UNSUPERVISED DOMAIN ADAPTATION FOR THE CROSS-DATASET DETECTION OF HUMPBACK WHALE CALLS

Andrea Napoli[*], *Paul R. White*

Institute of Sound and Vibration Research
University of Southampton, UK
{an1g18, P.R.White}@soton.ac.uk

ABSTRACT

Machine learning methods, and deep networks in particular, often underperform on data which lies outside the training distribution. Changes to the data distributions (known as domain shift) are particularly prevalent in bioacoustics, where many external factors can vary between datasets, although the effects of this are often not properly considered. This paper presents a benchmark for out of distribution (OOD) performance based on the detection of humpback whales in underwater acoustic data. Several humpback whale detectors from the literature are implemented as baselines, along with our own detector based on a convolutional neural network (CNN). Then, a set of unsupervised domain adaptation (UDA) algorithms are compared. Results show that UDA can significantly improve OOD performance when few distinct sources of training data are available. However, this is not a substitute for better data, as negative transfer (where the adapted models actually perform worse) is commonly observed. On the other hand, we find that training on a variety of distinct sources of data (at least 6) is sufficient to allow models to generalise OOD, without the need for advanced UDA algorithms. This allows our model to outperform all the baseline detectors we test, despite having 10,000 times fewer parameters and 100,000 times less training data than the next-best model.

Index Terms— Unsupervised domain adaptation, domain shift, passive acoustic monitoring, humpback whale detection

1. INTRODUCTION

Passive acoustic monitoring (PAM) forms a major part of marine mammal conservation. Acoustic surveys are an effective and non-invasive means to further our understanding of species-wise geographic distributions, migration patterns and feeding grounds, monitor ecosystem health, and help to mitigate the impacts of human activity. Automated analysis of survey data can improve our ability to achieve these goals, whilst substantially reducing the manual effort required [1].

An ideal solution to this end would be an off-the-shelf tool which can be easily deployed on any new data and identify all the vocalising species present (and indeed, any other relevant acoustic event). We argue a major obstacle exists to achieving this sort of generalisation ability that particularly affects PAM, but is seldom properly considered. This is the fact that dataset biases [2] in PAM are unusually large compared to other areas of machine learning research (consider, for example, ImageNet [3]: sourced by trawling images from the Internet, this may be a more representative sample of the "set of all possible images" for which it is a surrogate). This increases the likelihood of a mismatch between the data distributions of a model's training set and the data it then encounters when deployed (known as domain shift), violating the i.i.d. assumption and potentially causing significant reductions in performance on new data.

We support the view that shortcut solutions, in which the training distribution contains spurious correlations between classes which do not transfer to new data, are the primary cause of shift-induced performance drops in real-world problems [4]. If these patterns have lower descriptive complexity than the intended solution, models will preferentially use them to "cheat" on a task. This is a significant complication, as the learning bias for simpler solutions is a huge part (but not all) of what makes generalisation possible in the first place (in particular, it helps prevent overfitting). Although a form of data leakage, the introduction of shortcut solutions is oftentimes simply unavoidable when constructing datasets, so we believe these are better thought of as an integral part of the learning problem, rather than mere developer oversight.

Thus, our first aim is to design experiments that create more realistic testing scenarios for PAM algorithms. We can do this by ensuring the training and test sets never contain any domain overlap, to better mimic the distributional shifts which may occur "in the wild" (we call this the *OOD testing* setup).

What exactly constitutes a "domain" in this context we keep deliberately abstract; the primary aim is to confine any covariate which may cause shortcuts to a single domain. For example, in one data source we use [5], separate tapes are often digitised into a single master recording, so these are considered a single domain even if the original tapes were collected in different locations or years. As we are only testing on OOD samples, the fact that some domains have examples collected in different conditions, resulting in shortcuts within a single domain, is inconsequential (we also argue this happens unavoidably anyway).

Our second aim is to identify best practices for maximising OOD performance in these scenarios. Unsupervised domain adaptation (UDA) has previously been used to tackle domain shift across many areas of wildlife monitoring [6], including PAM [7], [8]. For marine mammal PAM, domain shifts have been shown to result in reduced performance [9], and basic supervised finetuning has been used to adapt models to new environments [10]. However, to our knowledge, UDA is unexplored in this context. Thus, in this paper, a range of UDA algorithms from the literature are applied to a test problem of humpback whale detection.

The UDA literature is dominated by the *distribution alignment* approach, which aims to minimise the distance between the

[*] Thanks to BAE Systems and the Engineering and Physical Sciences Research Council for funding.

feature distributions of the source and target domains. The crux of this approach is finding how to estimate this distance reliably using only samples from the distributions. Two main approaches exist: kernel methods, which embed the distributions in a reproducing kernel Hilbert space (RKHS) [11]–[15]; and adversarial training, pioneered by [16] and the current basis for practically all state-of-the-art methods. Various extensions to the original "domain adversarial neural network" (DANN) formulation have followed the better-known literature on generative adversarial networks, such as with the introduction of the cycle-consistency loss [17], conditional adversarial training [18] and the Wasserstein objective [19].

As a final note, we call attention to subsequent analyses of existing UDA (and, more broadly, domain generalisation) algorithms in new contexts, on additional, perhaps more realistic, datasets, or averaged across many tasks, which have failed to reproduce or report much-reduced benefits compared to their original publications [6], [8], [20], [21]. Thus, we also consider that testing existing algorithms on new data helps contribute to the bigger picture of how effective or useful these methods actually are.

In summary, in this paper, we compare 8 UDA algorithms on a novel benchmark of OOD humpback whale detection. We also analyse the effect of varying the number of domains used to train the base model.

2. DATA

Humpback whale (*Megaptera novaeangliae*) calls are perhaps the most studied of all marine mammal vocalisations, and also what non-biologists usually mean when they talk about "whale song". The complex nature of the song, its population-level variability, and the fact that humpbacks are found in a wide range of environments all over the world make for an attractive (i.e., challenging) OOD problem. Additionally, the large body of previous work means many acoustic datasets already exist online and there are several well-established baselines to compare our approach to.

We construct a dataset consisting of approximately 100 minutes of audio, labelled as either humpback whale (HW) or non-humpback whale (NHW), from 13 distinct sources. Most of these sources already contain both HW and NHW examples, although some have only a single class; these are paired together so that every domain has examples from both classes, for a total of 9 domains.

Most data was downloaded from freely available sources online: the Watkins Marine Mammal Sound Database (which includes locations in the Caribbean, North Atlantic and Antarctica) [5], the Pacific Islands Passive Acoustic Network [22], the Australian National Mooring Network [23], the Hawaiian Islands Cetacean and Ecosystem Assessment Survey [24] and moby-sound.org; the remaining data was recorded in Madagascar in an in-house collection project [25].

Samples were handpicked to create a diverse, representative, and challenging learning problem, covering a wide range of non-target underwater acoustic events, geographic locations, recording methods and environments. All audio was resampled to 8 kHz, although two domains have original sample rates of 4 and 6 kHz, so do not contain higher-frequency information – we just consider this an additional characteristic of the learning problem to be overcome. Some exemplar spectrograms are shown in Figure 1.

We use the same audio pre-processing pipeline as Allen et al. [26]: mel spectrograms are generated using 100 ms FFT windows with 50% overlap, normalised with per-channel energy normalisation [27], then split into 3.92 s analysis frames with 50% overlap. This results in 3,150 total spectrogram images, measuring 64 by 128 pixels. The number of images is broken down by class and domain in Figure 2.

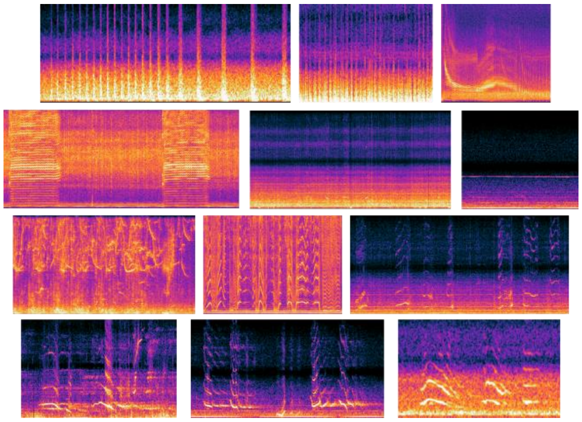

Figure 1: Some exemplar spectrograms of sounds in the dataset (4 kHz bandwidth, time axis scales variable). Top row: sperm whale clicks, pilot whale clicks, seal vocalisations. Second row: minke whale boings, right whale calls in strong vessel noise, electrical interference. Third row: dolphin whistles, dolphin creaks, right whale calls. Bottom row: three humpback whale calls.

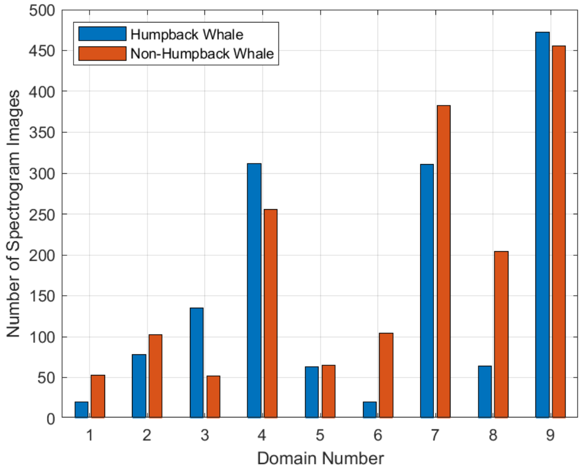

Figure 2: Total number of spectrogram images in the dataset, by class and domain.

Extracting a single value from the literature for what constitutes "acceptable" performance for this task is difficult. Helble et al. [28] state that any automated detector should perform at or above the level of a trained human analyst, although even this benchmark varies greatly based on the call's SNR, the nature of the background noise, as well as the human in question. However, based on values in [28], and without wishing to get too lost in the details, we consider a balanced accuracy of 87% to be the bare minimum required for this task, and anything above 90% to be good.

3. DETECTORS

A simple CNN is designed with 4 convolutional layers and one dense layer. The convolutional layers each have 3 by 3 kernels, (2, 2) stride, 16 filters and RELU activations, with 7,154 trainable parameters total. Batch normalisation was found in testing to deteriorate OOD performance, reproducing findings in the literature [20], so is not used. Training is performed using the Adam optimiser with an initial learning rate of 0.001 and a batch size of 32, for 500 iterations.

In addition to empirical risk minimization (ERM) (that is, the standard training paradigm with no adaptation), 8 UDA algorithms are compared:

- Principal component analysis (PCA)
- Correlation alignment (CORAL) [29]
- Geodesic flow kernel (GFK) [30]
- Transfer component analysis (TCA) [11]
- Joint distribution alignment (JDA) [12]
- Transfer joint matching (TJM) [13]
- Manifold embedded distribution alignment (MEDA) [14]
- Scatter component analysis (SCA) [15]

The CNN is first trained normally on the source domain data. The UDA algorithms are then applied to the activations of the final convolutional layer. Finally, a new dense layer is trained on the transformed source domain features. For the methods based on dimensionality reduction (all but CORAL), the output dimension is set to 8. The whole process is repeated 5 times to reduce the influence of parameter initialisation and provide a measure of the uncertainty for the results.

In addition to the shallow UDA algorithms listed above, various types of deep adversarial UDA [16], [18], [19] were also attempted, but failed to work, and are not included in these results. Other than the notorious difficulties that come with adversarial training (e.g., mode collapse), we also suspect that these methods require larger amounts of data than is available in our application, which may explain why they failed in this case.

3.1. Baselines

We also implement 3 baseline detectors for this task:

1) *Allen et al.* [26], a ResNet-50 [31] architecture (25.6 M parameters) trained on 187,000 hours of data from a single PAM program [22]. The decision threshold is set to the average of all the optimal thresholds stated in the paper (a different threshold is used per site), at 0.13. One domain of our dataset contains data overlap with the training set for this model, so we do not include it when calculating the average test accuracy for this baseline.

2) *YAMNet* [32], a MobileNet-V1 [33] architecture (3.7 M parameters) trained on AudioSet [34], a broad ontology of 527 classes of audio events drawn from YouTube. In particular, we are counting detections of the class “Whale vocalisation”. The training data for this class consisted of around 20 minutes of audio from 129 videos, most of which upon inspection are humpback whales.

3) *Template matching*, via cross-correlation of spectrograms [35]. For each test sample, a 2D correlation is performed with each humpback call training sample and the highest correlation value is taken as the recognition score. The nontarget training samples are unused. The decision threshold is chosen based on tests on a held-out subset of training data, and is set to 0.2.

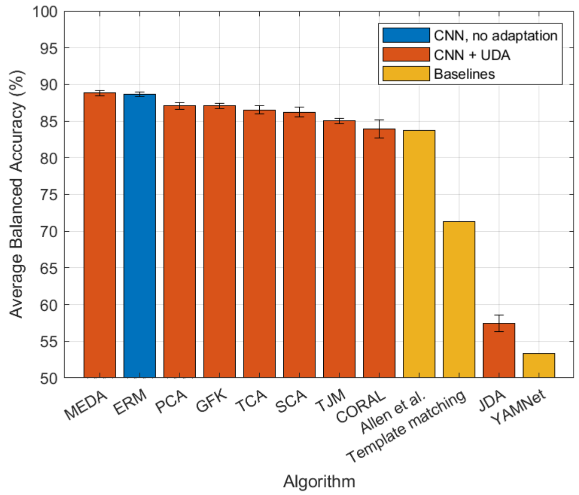

Figure 3: Average balanced accuracy across all domains for each algorithm. Error bars denote standard error in the mean.

4. RESULTS

First, the algorithms and baselines are compared using leave-one-domain-out cross-validation – that is, the model is trained using data from all but one domain at a time. The performance measure used is balanced accuracy, equal to the mean of the true positive rate and true negative rate, and averaged across all domains. The results are shown in Figure 3, where the error bars denote standard error in the mean across the repeats (note, the baselines do not have error bars).

Our tests show that no UDA algorithm exceeds ERM by a significant margin – at most 0.2 percentage points for MEDA. This reproduces recent findings on OOD generalisation from the literature [6], [8], [20], [21] – where the ERM baseline has been described as “frustratingly strong” [20]. It is clear that, in this case, the diversity of the training data makes a far larger difference than the learning algorithm, with our best models significantly outperforming the Allen et al. [26] baseline, despite having 10,000 times fewer parameters, 100,000 times less training data and no pretrained backbone. A total of 4 algorithms, including ERM, exceed the 87% accuracy criterion. Template matching also performs surprisingly well, although this is rather dependent on the domain being tested.

What is perhaps most striking is how often UDA actually reduces performance when it is applied – a phenomenon known as *negative transfer* [36]. Some algorithms completely destroy the model’s predictive power (e.g., JDA) and *every* algorithm underperforms ERM in at least one domain. This behaviour has been observed consistently throughout our work on UDA – not least for the adversarial methods. We suspect that a bias exists in commonly used UDA benchmarks which may, in particular, explain why our application of UDA fails to reproduce the massive improvements on ERM often seen elsewhere. This is that the distribution alignment is often performed on features from a pretrained backbone (usually ResNet-50) which has already “seen” target domain samples. The biased feature distributions then make the alignment task far easier than if no such pretraining is available. Otherwise, the phenomenon of *modal misalignment* (also called false alignment [36], essentially analogous to overfitting) is far

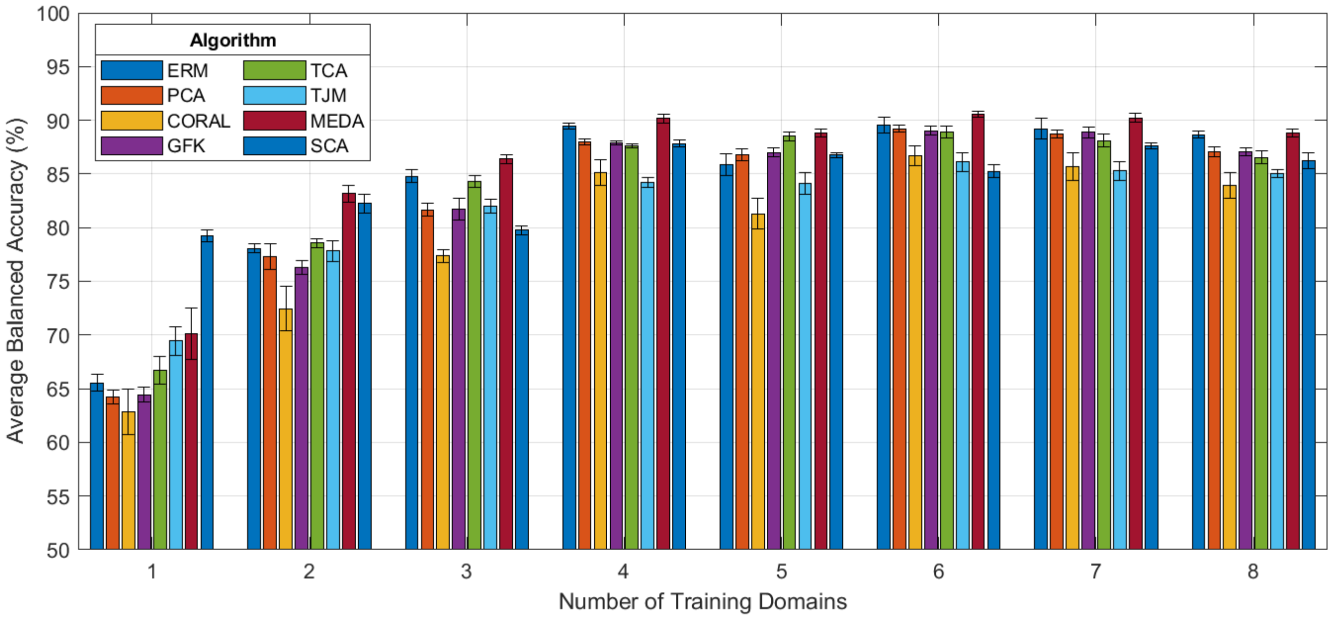

Figure 4: Balanced accuracy by number of domains in the training set, for each algorithm.

more prevalent. Extending the current UDA setup (where all training domains are pooled together and considered a single "source" domain) to multi-source UDA has previously been proposed to alleviate this difficulty [36], and will be investigated in future work.

4.1. How many training domains are needed?

The previous section showed that, given an abundance of training domains, no UDA algorithm significantly outperforms ERM. However, the question arises: when training domains are limited (as can easily happen in PAM, particularly for rare species), can UDA compensate for the lack of diversity in the training data?

In this section, the number of training domains is varied from 1 to 8. The domains that are not used for training are used for validation. This is done across at least 3 cross-validation folds, subject to the training set being large enough (we use a cut-off of at least 500 instances). The average balanced accuracy across all validation folds and domains, along with standard errors, is shown in Figure 4.

It can be seen that UDA is increasingly beneficial as fewer training domains become available. With a single training domain, SCA provides 14 percentage points improvement over ERM, although it is not a complete substitute for better data. Having at least 6 training domains appears to be a necessary and sufficient condition for achieving maximal performance on this dataset: it is the point where the performance of most algorithms no longer increases, as well as the point where UDA no longer significantly improves on ERM.

The fact that OOD accuracy does not clearly increase monotonically with the number of training domains (for example, there is a definite drop for most algorithms in going from 4 to 5 domains) suggests that, as found in [37], the design of the dataset, including the relative abundance of each domain, is an important factor, and naively combining as much data as possible may not be the best strategy. This will be investigated further in future work.

5. CONCLUSION

This paper presented a novel benchmark for OOD generalisation, namely the cross-dataset detection of humpback whales in PAM data. A total of 8 UDA algorithms, applied to a simple CNN detector, were tested on this benchmark, as well as 3 existing baseline detectors. It was shown that large domain shifts exist between data from different PAM projects, resulting in significant underperformance OOD if training data from only one domain is used. However, training on a variety of distinct sources of data (at least 6) is sufficient to allow models to generalise OOD, without the need for advanced algorithms. In cases where limited training domains are available, UDA can be used to recover a large part of the shift-induced performance drop.

Although some algorithms may exceed ERM on average, no algorithm consistently outperforms ERM every time, highlighting the challenges still faced in achieving reliable, trustworthy OOD generalisation. Being able to predict which algorithms will work in a particular domain would be a significant step towards achieving this goal – for example, the best model could then be dispatched automatically using a specialty-aware ensemble [38]. As of yet, no pattern appears to exist, although this will be investigated further in future work.

6. REFERENCES

[1] P. Nguyen, "Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context," Doctoral Thesis, Sorbonne University, 2020.

[2] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Mar. 2010.

[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant Risk Minimization,” *arXiv*, 2019.

[5] L. Sayigh *et al.*, “The Watkins Marine Mammal Sound Database: An online, freely accessible resource,” in *Proceedings of Meetings on Acoustics*, 2016.

[6] S. Sagawa *et al.*, “Extending the WILDS Benchmark for Unsupervised Adaptation,” *ICLR*, 2021.

[7] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, “Robust sound event detection in bioacoustic sensor networks,” *PLoS One*, 2019.

[8] M. Boudiaf, T. Denton, B. van Merriënboer, V. Dumoulin, and E. Triantafillou, “In Search for a Generalizable Method for Source Free Domain Adaptation,” *ICML*, 2023.

[9] O. S. Kirsebom, F. Frazao, Y. Simard, N. Roy, S. Matwin, and S. Giard, “Performance of a deep neural network at detecting North Atlantic right whale upcalls,” *Journal of the Acoustical Society of America*, 2020.

[10] B. Padovese *et al.*, “Adapting deep learning models to new acoustic environments - A case study on the North Atlantic right whale upcall,” *Ecol Inform*, 2023.

[11] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans Neural Netw*, 2011.

[12] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” *ICCV*, 2013.

[13] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer joint matching for unsupervised domain adaptation,” *ICCV*, 2014.

[14] J. Wang, W. Feng, Y. Chen, M. Huang, H. Yu, and P. S. Yu, “Visual Domain Adaptation with Manifold Embedded Distribution Alignment,” *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 402–410, Jul. 2018.

[15] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, “Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1414–1430, Oct. 2015.

[16] Y. Ganin *et al.*, “Domain-Adversarial Training of Neural Networks,” *JMLR*, 2015.

[17] J. Hoffman *et al.*, “CyCADA: Cycle-Consistent Adversarial Domain Adaptation”, *ICML*. 2017.

[18] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional Adversarial Domain Adaptation,” *Advances in Neural Information Processing Systems*, May 2017.

[19] J. Shen, Y. Qu, W. Zhang, and Y. Yu, “Wasserstein Distance Guided Representation Learning for Domain Adaptation,” *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 4058–4065, Jul. 2017, doi: 10.1609/aaai.v32i1.11784.

[20] I. Gulrajani and D. Lopez-Paz, “In Search of Lost Domain Generalization,” *ICLR*, 2021.

[21] A. Dubey, V. Ramanathan, A. Pentland, and D. Mahajan, “Adaptive Methods for Real-World Domain Generalization,” *CVPR*, 2021, doi: 10.1109/CVPR46437.2021.01411.

[22] NOAA Pacific Islands Fisheries Science Center, “Pacific Islands Passive Acoustic Network (PIPAN) 10kHz Data.” NOAA National Centers for Environmental Information, 2021.

[23] Integrated Marine Observing System, “Australian National Mooring Network,” 2017. https://imos.org.au/facilities/nationalmooringnetwork (accessed Apr. 09, 2023).

[24] NOAA Pacific Islands Fisheries Science Center, “Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) towed array data. Edited and annotated for DCLDE 2022,” *NOAA National Centers for Environmental Information*. 2022.

[25] F. Pace, P. White, and O. Adam, “Hidden Markov modeling for humpback whale (Megaptera Novaeanglie) call classification,” *Proceedings of Meetings on Acoustics*, 2012.

[26] A. N. Allen *et al.*, “A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset,” *Front Mar Sci*, 2021.

[27] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, “Trainable Frontend For Robust and Far-Field Keyword Spotting,” *ICASSP*, 2016.

[28] T. A. Helble, G. R. Ierley, G. L. D’Spain, M. A. Roch, and J. A. Hildebrand, “A generalized power-law detection algorithm for humpback whale vocalizations,” *The Journal of the Acoustical Society of America*, Apr. 2012.

[29] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 2058–2065, 2016.

[30] B. Gong, Y. Shi, F. Sha, and K. Grauman, “Geodesic flow kernel for unsupervised domain adaptation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, 2012.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *CVPR*, 2015.

[32] “YAMNet.” https://github.com/tensorflow/models/tree/master/research/audioset/yamnet (accessed Jun. 23, 2022).

[33] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *CoRR*, Apr. 2017, doi: 10.48550/arxiv.1704.04861.

[34] J. F. Gemmeke *et al.*, “Audio Set: An ontology and human-labeled dataset for audio events,” *ICASSP*, 2017.

[35] E. T. Vu *et al.*, “Humpback whale song occurs extensively on feeding grounds in the western North Atlantic Ocean,” *Aquatic Biology*, 2012.

[36] Z. Pei, Z. Cao, M. Long, and J. Wang, “Multi-Adversarial Domain Adaptation,” *AAAI*, Sep. 2018.

[37] T. Nguyen, G. Ilharco, M. Wortsman, S. Oh, and L. Schmidt, “Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP,” *NeurIPS*, Aug. 2022.

[38] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li, “Domain Generalization using Pretrained Models without Fine-tuning,” *CoRR*, Mar. 2022.