Methods to Improve the Robustness of Right Whale Detection using CNNs in Changing Conditions

W. Vickers
School of Computing Sciences
University of East Anglia
Norwich, UK
w.vickers@uea.ac.uk

B. Milner
School of Computing Sciences
University of East Anglia
Norwich, UK
b.milner@uea.ac.uk

A. Gorpincenko
School of Computing Sciences
University of East Anglia
Norwich, UK
A.Gorpincenko@uea.ac.uk

R. Lee
Gardline Environmental
Gardline Geosurvey Limited
Great Yarmouth, UK
robert.lee@gardline.com

***Abstract*—This paper is concerned with developing a method of detecting right whales from autonomous surface vehicles (ASVs) that is robust to changing operating conditions. A baseline convolutional neural network (CNN) is first trained using data taken from a single operating condition. Its detection accuracy is then found to degrade when applied to different operating conditions. Two methods are then investigated to restore performance using just a single model. The first method is an augmented training approach where progressively more data from the new condition is mixed with the original data. The second method uses unsupervised adaptation to adapt the original model to the new conditions. Evaluation under changing environmental and noise conditions reveals the model produced from augmented training data to achieve higher detection accuracy across all conditions than the adapted model. However, the adapted model does not require label data from the new environment and in these situations is a more realistic solution.**

***Index Terms*—cetacean detection, autonomous surface vehicles, passive acoustic monitoring, CNN, augmentation, adaptation**

I. Introduction

This work is concerned with developing robust methods for detecting marine mammals from autonomous surface vehicles (ASVs). Detecting marine mammals is important for population monitoring and for mitigation as many species are endangered and protected by environmental laws. In particular, we consider detecting North Atlantic right whales (Eubalaena glacialis) in the vicinity of potentially harmful offshore activities. Detecting their presence before they enter a mitigation zone both protects the animal and avoids shutdown of costly sub-sea operations. Marine mammals have traditionally been detected by human observers on-board ships, but more recently ASVs have been used [1]. An ASV typically uses passive acoustic monitoring (PAM) to listen for whale sounds. This provides a cheaper and more accessible alternative that can also operate in low visibility conditions.

Many machine learning methods have been applied to cetacean detection in recent years. For example, methods such as vector quantisation and dynamic time warping have been effective in detecting blue and fin whales from their frequency contours extracted from spectrograms [2]. Hidden Markov models (HMMs) have also been used to recognise low frequency whale sounds using spectrogram features [3]. Comparisons have also been made between between artificial neural networks (ANNs) and spectrogram correlation for right whale detection [4]. A study of various time-series classification and deep learning approaches to right whale detection found convolutional neural networks (CNNs) to give highest accuracy [5]. Further studies have also found success in using CNNs for right whale detection when compared to other classification models such as recurrent neural networks [6], [7].

The aim of this work is to consider scenarios where deployment conditions for right whale detection are changing and not necessarily matched to the source data used to train the underlying model. We therefore propose to develop a single generalised model that can operate effectively under different operating conditions. We begin by optimising a detector on a specific set of environmental and noise conditions and analyse its performance as conditions change. We then explore two different methods to create a new model to restore performance in the mismatched conditions, whilst still retaining good performance in the original conditions. The first method is to augment the source training data with new, labelled, data taken from the new target operating conditions, with the aim of producing a single model that performs well in both the original (source) and new (target) conditions. The second method applies unsupervised adaptation to the source model to create a new model matched to the target data. In this situation we explore the unsupervised adaptation method of adversarial discriminative domain adaptation (ADDA) which has been effective in image classification applications [8].

The remainder of the paper is organised as follows. Section II gives a brief introduction to right whales and the sounds they produce. A baseline CNN right whale detector is introduced in Section III and is then analysed in terms of how its accuracy is affected by changes in environmental and noise conditions. The augmentation and unsupervised adaptation methods are presented in Section IV. A set of right whale detection results are presented in Section V that investigate the effectiveness of the two methods in changing operating conditions.

II. Characteristics of right whales

Right whales are one of the most endangered marine mammals and are at risk of extinction with as few as 350 individuals remaining [9]. Right whales make a range of vocalisations that have been well documented [10]. This work

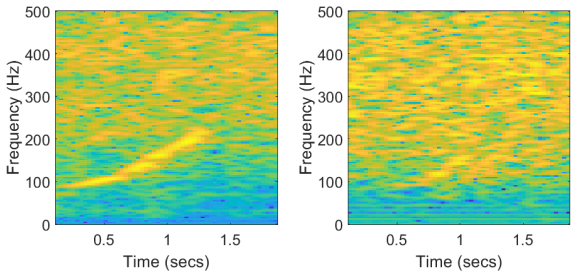

Fig. 1. *Example spectrograms of right whale up-sweep calls recorded from two different conditions.*

focuses on their most commonly documented sound, an up-sweep tone from approximately 60Hz to 250Hz typically lasting for one second. Two examples are shown in Figure 1 which illustrates calls taken in different conditions caused by marine noise and distance. Calls are not always consistent and vary in duration and frequency [11]. Calls can be difficult to hear and visualise in spectrograms, as these low frequency bands are often masked by sounds from ship noise, drilling, piling, seismic exploration or interference from other marine mammals [12]. These overlapping frequencies can cause large amounts of background noise in the signal making detection difficult. Current methods of collecting cetacean data involve towing a hydrophone array from a ship and using trained observers to listen and watch the water for mammal activity. We aim to produce a robust detection system that can be used on-board an ASV as a cheaper and more efficient solution.

III. Baseline detector and initial analysis

This section introduces the baseline right whale detector and investigates its accuracy when the operating environment and noise conditions change.

A. Right whale datasets

To investigate how changing operating conditions affect detection accuracy, two right whale datasets are used. The first is provided from the Marinexplore and Cornell University Whale Detection Challenge[1] of North Atlantic right whale up-calls, while the second was collected from the NRW Buoys in the Cape Cod region[2]. Although data from these is similar, they exhibit different mean spectral shapes, with the Cornell data being bandpass while that from Cape Cod has a flatter frequency response. As will be shown, these mismatches lead to different detection rates for the two datasets. Audio from both datasets is arranged as two-second segments where each either contains a right whale sound or does not, as indicated by an associated label. The audio is downsampled to 1 kHz, based on previous work that established that for right whale detection this introduces no loss in accuracy [5]. The Cape Cod dataset contains substantially more examples than the Cornell dataset and these are divided into non-overlapping training, validation and test sets. Specifically, Cornell and Cape Cod have 10,000 and 100,000 training examples, respectively. The validation and test data sets are set based on a data split of 70:15:15, which results in validation and test set sizes of 2,143 and 21,429, for the Cornell and Cape Cod datasets, respectively. All sets of data are balanced to contain equal proportions of segments with and without right whales present.

[1]https://www.kaggle.com/c/whale-detection-challenge/data

[2]https://portal.nrwbuoys.org/ab/dash/

B. Baseline CNN detector

The baseline right whale detector is taken from earlier work that compared a range of machine learning and deep learning techniques. This established that highest detection accuracy is achieved using a block of convolutional layers to encode input spectrogram features extracted from the audio signal, followed by a network of dense layers to perform classification [5].

The spectrogram features are created using a sliding window to convert short-duration frames of audio into a sequence of log power spectral vectors. Previous work established that best performance is obtained when extracting N=256-point frames of time-domain samples using a Hamming window and then applying a Fourier transform. The upper $N/2$ frequency points are discarded and the remaining amplitudes converted to a log power spectrum. Analysis windows are advanced by S=32 samples to compute each new log spectral vector which together produce a spectrogram feature. Normalisation is applied to the amplitudes to transform them into the range 0 to 1.

The CNN encoder, M_S, maps the input spectrogram into a new space and contains three convolutional layers that are each followed by max pooling layers. This outputs into two dense layers that form the classifier, C [5]. Each convolutional layer uses 3×3 filter kernels with 32, 64, 128 filters on each subsequent layer. The max pooling layers use a pool size of 2×2 and have ReLU non-linear activation function applied to their outputs. At the edges of the input, zero-padding is applied to convolutional layers to maintain the size of the output. After the last max pooling layer a dropout of 0.5 is applied and the latter dense layers use 200 and 50 nodes respectively also with a ReLU function. The final dense layer uses a sigmoid activation function and outputs a probability of a right whale being present. For training, an Adam optimiser is used with a learning rate of 0.001 and binary cross-entropy as the loss function [13]. Training took place over 200 epochs and was repeated 10 times for each test. The model that achieved highest validation accuracy was used for testing and reported accuracies were calculated as an average over all 10 tests (seen in Figures 2, 4, 5, and 6).

C. Analysis

Two sets of analysis are now presented that explore the effect of changing operating conditions, first on the environment and second on additive noise.

1) Environment: To examine the effect of changing environment, the Cornell and Cape Cod datasets are used. Two right whale detectors are trained using the CNN architecture of Section III-B, one using Cornell training data and the other Cape Cod training data. Given the large difference in the number of training samples, the tests also examine how the

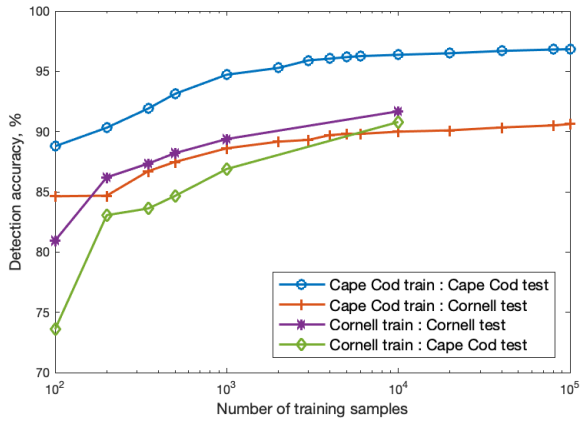

Fig. 2. *Detection accuracy as the number of training samples is increased for matched and unmatched Cape Cod and Cornell datasets.*

amount of training data affects detection accuracy. Specifically, the Cornell model is trained using 100 to 10,000 training samples while the Cape Cod model is trained using 100 to 100,000 training samples. Figure 2 shows detection accuracy on the two datasets as the number of training samples is increased. Using just 100 training examples from the Cape Cod dataset gives 88.8% detection accuracy which rises to 96.8% with 100,000 examples, with performance levelling at around 10,000 examples. Detection accuracy for the Cornell data follows a similar trend, although starts lower at 81.0% with 100 training examples and peaks at a lower level of 91.7% using 10,000 training examples.

In a practical scenario it is desirable to use a single model for detection, irrespective from where the test data is received. To evaluate this condition two further tests are presented in Figure 2 that show detection accuracy of the Cape Cod test data when applied to the Cornell model, and the accuracy of Cornell test data against the Cape Cod model. The Cornell test data attains peak detection accuracy of 90.0% when applied to the Cape Cod model which is a reduction of 1.7% compared to the matched model. Even using the 100,000 sample-trained Cape Cod model, accuracy increases to only 90.6%. Testing Cape Cod data on the Cornell model with 10,000 training samples attains 90.8% accuracy, which compares to 96.8% when tested in matched conditions. These tests show that sub-optimal performance results from the mismatched training/testing conditions.

2) Noise: The robustness of the baseline detection system to changing noise conditions is now examined. Many sources contribute to sub-sea noise and thereby reduce the received signal-to-noise ratio (SNR). Furthermore, sounds recorded from more distant whales will also be received with lower SNRs. To simulate noisy conditions we add white noise to the audio. Although this is not specific to any particular sub-sea noise, it gives a good indication of how robust detection is in noisy conditions. Specifically, white noise at an SNR of 0dB is added to the test samples of the Cornell dataset as this was found to have a significant impact on accuracy. These samples are tested against the baseline Cornell-trained CNN and also against a CNN trained on the Cornell dataset contaminated with white noise at an SNR of 0dB to give a matched condition model. As expected, noisy test data causes detection accuracy against the clean-trained model to reduce from 91.7% to 72.3%. However, testing using the matched noisy-trained model recovers detection accuracy to 83.1%.

Taking the clean-trained Cornell model as the baseline source model this analysis has shown that its performance deteriorates as both the environment changes and as background noise increases. Performance can be improved by training new models that match to the new conditions, but this is impractical in real situations. Instead, methods are now considered to improve the robustness of this single model, trained on source data, so that it performs robustly across a range of operating conditions.

IV. Improving Robustness

We consider two approaches for creating a single robust model that can operate effectively under different operating conditions. The first method is based on data augmentation and the second on unsupervised adaptation.

A. Augmentation-based training

This approach augments the existing source training data with examples matched to the target operating conditions. For all cases the baseline model is the Cornell-trained source CNN and classifier, introduced in Section III-B. To adapt to a new set of target conditions, for example a new environment or noise, varying amounts of target data from the new condition is used to augment the existing source training data. A new model is then trained that aims to deliver robust detection accuracy on both the original source condition and new target condition.

B. Unsupervised adaptation

Several methods have been proposed for unsupervised adaptation of CNN encoders to a new target domain, with some of the most effective being based on generative adversarial networks [14]. In particular, the adversarial discriminative domain adaptation (ADDA) has worked effectively in a range of image classification tasks and is consequently employed as the adaptation method for whale detection. For implementation it is assumed that target training data is available but without any labels which makes the method well suited to a new, unknown operating condition. Implementation of ADDA is a three-stage procedure which is shown in Figure 3. The first stage uses only the source data and associated class labels to train a CNN encoder, M_S, and classifier, C. This is the same procedure used to create the baseline whale detector in Section III-B and is shown in Figure 3a.

The second stage creates a target encoder, M_T, that aims to transform the target data into the same feature space as the source data and is illustrated in Figure 3b. In this way the same classifier, C, can be used for whale detection from both the source and target data. The target encoder, M_T, is initialised using the weights in the source encoder, M_S. A discriminator network, D, is employed to discriminate between source data and target data. Training involves optimising the discriminator

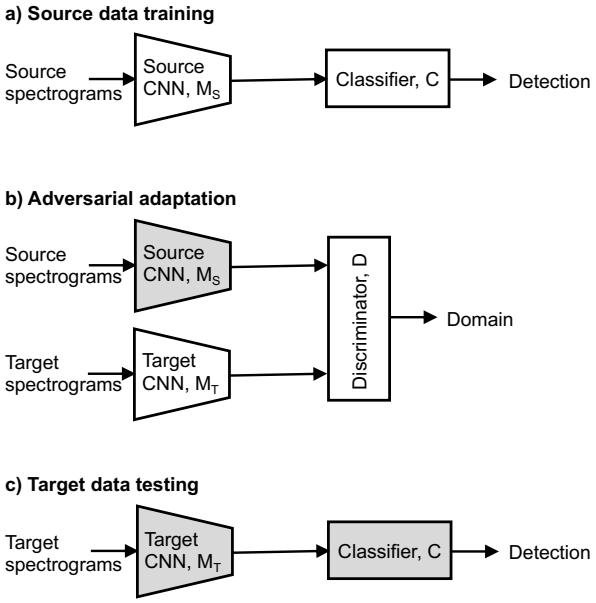

Fig. 3. *Method of adversarial discriminative domain adaptation (ADDA) applied to spectrogram-based right whale detection. Gray boxes indicate a network that is fixed during training.*

and finding a target encoder that transforms target spectrograms into a space by minimising the accuracy of the discriminator, which means that the source and target domains have become similar.

Testing on the target data is shown in Figure 3c where spectrograms extracted from the new environment or noise condition are transformed by the target encoder, M_T, into the source data space. The classifier, C, then determines whether or not a whale is present.

V. Experimental results

Experiments first examine how effective the methods in Section IV are at improving right whale detection in new operating environments. Second, their effect in new noise conditions is examined. Third, the effect of changing both the environment and noise is examined. The baseline whale detection method for all tests is the Cornell-trained CNN introduced in Section III-B.

A. Changing environment

The effectiveness of the augmentation method of Section IV-A is examined first within the changing environment scenario. Cornell training data is augmented with samples taken from Cape Cod and a new augmented detector is trained. The model is tested on both the Cornell and Cape Cod test sets as the amount of augmentation samples is increased from 10 to 10,000 (to match the total number of Cornell samples). Figure 4 shows that as more Cape Cod samples augment the Cornell training data, detection accuracy for Cape Cod improves rapidly and approaches the matched-condition performance. Importantly, when testing the same model with Cornell test data, the detection accuracy is almost unchanged from the

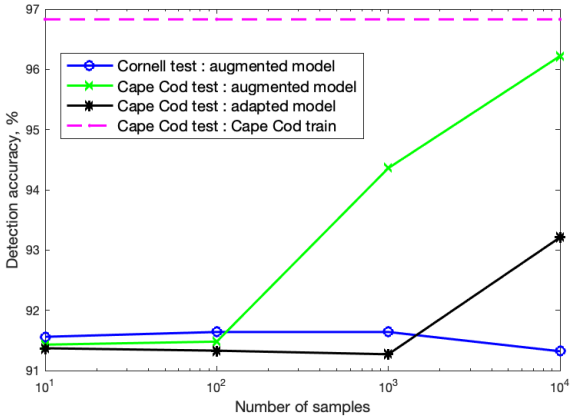

Fig. 4. *Detection accuracy of the augmented-trained model and unsupervised adapted model as the number of Cape Cod samples is increased. All models start with a baseline of 10,000 samples from the Cornell set.*

original Cornell-trained model that attained 91.7%. This shows that augmented training is able to generate a single model that has no degradation on Cornell and Cape Cod attains accuracy approaching the matched Cape Cod system - 96.8% compared to 96.2%.

Considering now the unsupervised adaptation of Section IV-B, in this scenario the Cornell-trained model forms the source model and increasing amounts of Cape Cod training data are used to create a new target model that is tested on Cape Cod. Figure 4 shows detection accuracy of this model when tested with Cape Cod data, using 10 to 10,000 Cape Cod training samples. Using up to 1,000 target samples has almost no effect but with 10,000 samples the accuracy increases to 93.2% which is approximately mid-way between the Cornell and Cape Cod trained models.

B. Changing noise

Tests in this section investigate changing noise conditions and begin by creating a copy of the Cornell training data that is mixed with white noise at an average SNR of 0dB. Varying amounts of this noisy data are used to augment the original Cornell training data and new models created. Figure 5 shows detection accuracy when tested against noisy Cornell data as the number of augmented noisy samples is increased from 10 to 10,000. As the amount of augmentation data increases, detection accuracy increases and when this equals the number of original training examples, performance equals the matched accuracy of 83.1%. Also shown in Figure 5 is the detection accuracy of the augmented model when testing against the original clean test data. This achieves almost the same performance as when tested against the clean trained model, 91.4% compared to 91.7%. Therefore, the single augmented-trained model attains close to optimal performance in both clean and noisy test conditions.

Considering the unsupervised adaptation, varying amounts of the noisy Cornell training data are used to adapt the source model. Figure 5 shows detection accuracy as the number of

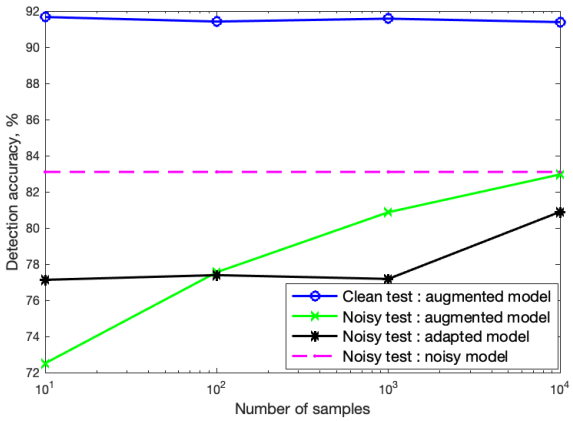

Fig. 5. *Detection accuracy of the augmented-trained model and unsupervised adapted model as the number of noisy samples is increased. These test use a baseline Cornell model before augmentation or adaptation is applied.*

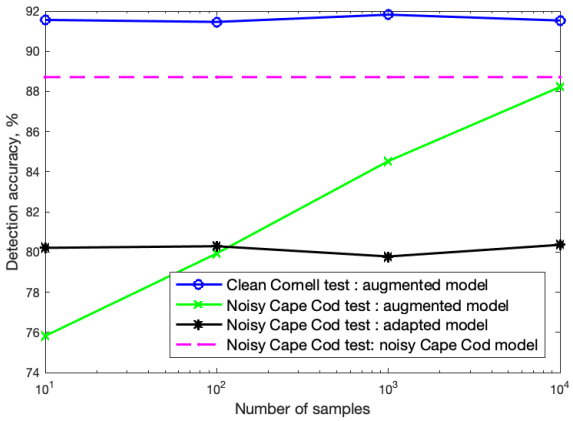

Fig. 6. *Detection accuracy of the augmented-trained model and unsupervised adapted model as the number of noisy Cape Cod samples is increased.*

target samples is increased from 10 to 10,000. With a relatively small number of target samples, accuracy is increased from 72.3% with no adaptation to 77%. Using the maximum number of 10,000 target samples, accuracy increases to 80.9%.

C. Changing environment and noise

As a final test, changes to both the environment and noise conditions are considered together. Target data is created by combining white noise at 0dB with the Cape Cod data. Figure 6 shows the detection accuracy of noisy Cape Cod test data when tested against the augmented-trained model and the adapted model. For the augmented model, a increasing improvement in accuracy is observed as more augmentation data is used. With the maximum amount, performance approaches that of matched condition training, 88.2% compared to 88.7%. For unsupervised adaptation, an initial increase from 73.6%, with no adaptation data, to 80.4% with as little as 10 target samples is seen. However, further increases in adaptation data give no increase in accuracy. This we attribute to difficulties in creating suitably stable models.

VI. Conclusion

Experiments have shown that both the augmented training and unsupervised adaptation are able to improve baseline performance when the operating conditions change. Using augmented training data produces a new model which has been shown to operate effectively in both original and new conditions. However, to train the model, labels for the target data are necessary. The ADDA unsupervised adaptation method also improves detection accuracy but not to the level of the augmented training. However, this method does not require labelled target samples which is a clear advantage when moving to a potentially unknown new operating condition.

Acknowledgement

We acknowledge the support of the Next Generation Unmanned Systems Science (NEXUSS) Centre for Doctoral Training, Gardline Geosurvey Limited and NVIDIA.

References

[1] U. K. Verfuss, A. S. Aniceto, D. V. Harris, D. Gillespie, S. Fielding, G. Jiménez, P. Johnston, R. R. Sinclair, A. Sivertsen, S. A. Solbø *et al.*, "A review of unmanned vehicles for the detection and monitoring of marine fauna," *Marine Pollution Bulletin*, vol. 140, pp. 17–29, 2019.

[2] X. Mouy, M. Bahoura, and Y. Simard, "Automatic recognition of fin and blue whale calls for real-time monitoring in the st. lawrence," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2918–28, 12 2009.

[3] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, May 2000.

[4] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Canadian Acoustics*, vol. 32, no. 2, pp. 55–65, Jun. 2004.

[5] W. Vickers, B. Milner, J. Lines, and R. Lee, "A comparison of machine learning methods for detecting right whales from autonomous surface vehicles," in *EUSIPCO*, 2019.

[6] E. Smirnov, "North atlantic right whale call detection with convolutional neural networks," in *Proc. Int. Conf. on Machine Learning, Atlanta, USA*. Citeseer, 2013, pp. 78–79.

[7] W. Vickers, B. Milner, J. Lines, and R. Lee, "Detecting right whales from autonomous surface vehicles using RNNs and CNNs," in *EUSIPCO - Satellite Workshop: Signal Processing, Computer Vision and Deep Learning for Autonomous Systems*, 2019.

[8] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017.

[9] S. D. Kraus, M. W. Brown, H. Caswell, C. W. Clark, M. Fujiwara, P. K. Hamilton, R. D. Kenney, A. R. Knowlton, S. Landry, C. A. Mayo, W. A. McLellan, M. J. Moore, D. P. Nowacek, D. A. Pabst, A. J. Read, and R. M. Rolland, "North Atlantic Right Whales in Crisis," *Science*, vol. 309, no. 5734, pp. 561–562, Jul. 2005.

[10] C. W. Clark, "Acoustic communication and behavior of the southern right whale (eubalaena australis)," *Communication and behavior of whales*, pp. 163–198, 1983.

[11] K. Pylypenko, "Right whale detection using artificial neural network and principal component analysis," Apr. 2015, pp. 370–373.

[12] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, Jun. 2004.

[13] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.