forum acusticum 2026
8-12 September · Graz, Austria

MITIGATING DOMAIN SHIFT IN BIRD CALL RECOGNITION

Giotto Frean[1], Stephen Marsland[1]

[1] *School of Mathematics and Statistics, Victoria University of Wellington, New Zealand*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Automated bird call recognition systems trained on one recording environment often produce worse results when deployed to others due to differences in background noise and recording equipment, a problem known as domain shift. To isolate this issue we process data from two different recording sources into comparable datasets with matched marginal label distributions and similar durations. We fine-tune a pre-trained bird recognition CNN on each dataset and find substantial domain shift in both directions, with 18–84% performance loss depending on preprocessing method. We test different methods for addressing this problem, including normalization of the spectrogram, adding noise, and domain adversarial training. Of the methods tested, adaptive background subtraction helps models generalize the most, reducing performance loss to as low as 18–32%. Moderate noise augmentation can help at the cost of reduced in-domain performance, while noise diversity shows mixed effects.

Keywords: *bird call recognition, domain shift, cross-domain evaluation, bioacoustic monitoring*

1. Introduction

The automated recognition of bird species from acoustic recordings has improved markedly in recent years [1], [2], [3]. Deep neural networks processing spectrograms of sound data are able to differentiate between the calls of many species, forming the basis for analysis tools such as AviaNZ [4], BirdNet [5], and Perch [6], amongst many others.

However, models often struggle in test environments due to differences in recording equipment and background noise [7]. This problem, called *domain shift*, remains a major challenge in bioacoustics [8], particularly for large-scale biodiversity monitoring where the background noise profile across the sites can differ markedly, and a variety of recorder types may be used [2].

In this paper we examine the extent of domain shift in bird call recognition, and evaluate strategies for mitigating it. To isolate the problem we construct comparable datasets with matched marginal label distributions and similar length, but different recording sources (locations and partially recorder types). We then systematically compare performance under preprocessing normalization, domain adversarial training, and noise augmentation.

2. Methods

2.1. Matched Dataset Construction

We started with two large datasets and created a matched pair from them. Individual recordings were capped at 10 seconds in length and may contain vocalizations of multiple species. The datasets were matched based on the distribution of species labels, rather than exact match.

The DOC dataset contains bird call recordings collected between 2010 and 2023 as part of New Zealand Department of Conservation (DOC) baseline biodiversity monitoring across public conservation land in Aotearoa New Zealand, using the DOC AR4[1] acoustic recorder sampling at 32 kHz.

The full initial DOC dataset is available at `https://www.kaggle.com/datasets/ollypowell/new-zealand-bird-sound` with Attribution 4.0 International (CC BY 4.0) license. Species presence within a minute was recorded based on manual identification by spectrogram visualisation and listening. Call were then extracted automatically to make short clips (3–10 seconds) that are assumed to match the bird labels of the minute. We subsequently processed the data to correct labels and ensure clear species presence.

Our second dataset comes from six sites in the Waitākere ranges, a national park approximately 25 km from Auckland in Aotearoa New Zealand. Data were collected as part of a MSc project [9] using Frontier Labs Bar-LT[2] and DOC AR4 recorders, both

***Corresponding author**: stephen.marsland@vuw.ac.nz.*

[1] electronics@doc.govt.nz

[2] `https://www.frontierlabs.com.au/bar-lt`

recording at 32 kHz. The data were manually annotated using the AviaNZ software by expert annotators, and clips for the focal birds extracted based on the annotations.
After controlling the species distribution the DOC dataset amounts to 766 files (64 minutes total duration) and the Waitākere dataset contains 684 files (73 minutes total duration).
The two datasets cover ten species (in nine classes) of birds commonly found in Aotearoa New Zealand, comprising:

1. two species of introduced European passerines (blackbird, *Turdus merula*; chaffinch *Fringilla coelebs*);

2. six species of native and self-introduced birds: four passerines (piwakawaka/fantail, *Rhipidura fuliginosa*; riroriro/grey warbler *Gerygone igata*; miromiro/tomtit, *Petroica macrocephala*; tauhou/silvereye, *Zosterops lateralis*), a parrot (kākā, *Nestor meridionalis*), and an owl (ruru/morepork, *Ninox novaeseelandiae*);

3. a final class combines the vocalisations of two oscine species with overlapping dialectic calls (tūī, *Prosthemadera novaeseelandiae*; korimako/bellbird, *Anthornis melanura*).

For both datasets, spectrograms were extracted in AviaNZ using 25 ms Hann window with a 10 ms hop between windows, and processed into 224 mel-scale frequency bins. While originally intended to match human hearing, the mel scale is a standard representation used for birdsong, and is used by the BirdClef pre-trained model we fine-tune. We used 224 bins instead of the standard 128 as this yielded consistently better performance in prior experimentation.
Data Splitting: To prevent data leakage from recording conditions, we employed a two-stage splitting strategy. For the Waitākere dataset, which contains multiple annotated segments from the same source recordings, we performed file-level splitting: all segments from a given source file were assigned to either training or test set, never both. Source files were randomly assigned to achieve approximately 75/25 train/test ratio, though actual segment-level ratio varies slightly due to differing annotations per file. For the DOC dataset we performed sample-level splitting, downsampling to match the species distribution from the Waitākere file-level split, ensuring both datasets have identical total size and comparable class balance. This introduces asymmetry between datasets and may partially contribute to directional performance differences. The DOC data spans geographically diverse locations, so we cannot guarantee full independence between samples due to unknown recording correlations.

2.2. Model Architecture and Training

We use a BirdClef 2025-pretrained RegNetY-008 CNN [10] fine-tuned to our nine classes. Training uses the AdamW optimizer (learning rate 10^{-4} for backbone, 10^{-3} for classifier), batch size 16, cosine annealing, early stopping with patience 15, and mixup with $\alpha = 0.25$.
Data Augmentation: During training, we apply standard spectrogram augmentations to improve generalization: time stretching (±10%, 50% probability), frequency shifting (±10 mel bins, 50% probability), temporal rolling (random circular shift along time axis), and SpecAugment-style time/frequency masking (2 rounds per sample, masking up to 20% of time and 48 frequency bins). These augmentations are disabled during evaluation.

2.3. Normalization Strategies

Using the spectrogram – a plot of the histogram of power in time-frequency bins – as an image as input to a deep neural network has generated effective results once appropriate normalization methods have been applied to remove background and smooth the peaks of the power distribution.
We compared six normalization approaches to assess their cross-domain robustness (an example of each is given in Figure 1). All normalization methods were applied to the linear mel spectrograms. They were:

Log baseline is the standard log transform;

Log+median applies a log transform followed by 5-frame median filter along the time axis;

Log+normalize begins with the log transform, then estimates background noise using the quietest 10% of frames per frequency band. Values exceeding 4 standard deviations from this estimate are treated as outliers and replaced by samples drawn from the non-outlier distribution within the same band. We then subtract this background estimate from the original;

Log+median+normalize applies the 5-frame median filter before background estimation and subtraction;

PCEN is per-channel energy normalization [11] with parameters gain = 0.8, bias = 10, power = 0.25, time constant $t = 60$ ms, and $\varepsilon = 10^{-6}$;

Box-Cox is a standard power transformation for variance stabilization [12], with λ estimated from the data via maximum likelihood.

2.4. Noise Augmentation

To investigate whether synthetic acoustic variation can improve generalization, we augment training spectrograms with background noise sampled from the Freesound dataset[3]. During training, noise is applied in the spectrogram domain by linear mixing of linear-amplitude (pre-log) spectrograms: $S_{\text{aug}} = (1-\alpha) \cdot S_{\text{orig}} + \alpha \cdot N_{\text{scaled}}$, where S_{orig} is the linear-amplitude spectrogram, N_{scaled} is a noise spectrogram

[3] http://freesound.org

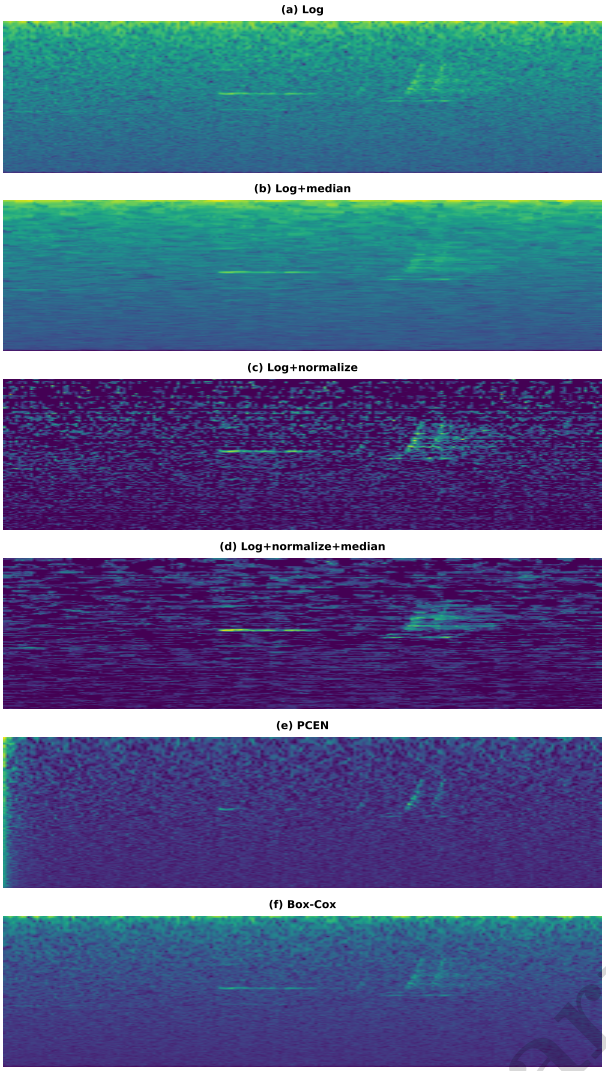

Figure 1: Spectrogram comparison under different normalizations

RMS-scaled to match the signal energy, and α is a fixed mixing ratio. The normalization is then applied to S_{aug}.

We systematically vary two augmentation parameters: **intensity** ($\in \{0.0, 0.2, 0.4, 0.6, 0.8\}$) controls the noise mixing ratio, and **variety** (1, 10, 100, or 1000 unique noise samples) controls acoustic diversity. Noise samples are preprocessed identically to bird calls. This design allows us to separate the effects of noise strength from noise diversity [13].

2.5. Domain Adversarial Training

We evaluate domain adversarial neural networks (DANN) [14] as a learning approach to domain adaptation. DANN extends standard classification networks with an auxiliary domain classifier trained adversarially via gradient reversal, learning representations that are discriminative for species classification yet invariant to domain identity.

Our implementation adds a domain classification head to the model backbone, taking features from the final layer. The domain classifier consists of three fully connected layers (running from 512 units to 256 to 1) with batch normalization, ReLU activations, and dropout (0.3). This head predicts binary domain labels (DOC vs Waitākere). A gradient reversal layer with $\lambda = 0.3$ inverts gradients flowing back to the feature extractor, encouraging domain-invariant representations. The domain classifier is trained jointly with the species classifier using binary cross-entropy with logits loss.

2.6. Evaluation Metrics

We evaluate models using exact-match accuracy: a prediction is correct only if all species labels match the ground truth. This is a harsh metric for multi-label evaluation, particularly when most clips contain a single species. However, for presence/absence detection of species it is the most appropriate metric.

Let D_s and D_t denote source (training) and target (test) domains, and let $A(D_s \to D_t)$ denote exact accuracy (all classes) when training on D_s and testing on D_t.

In-domain accuracy is $A(D_s \to D_s)$, the performance when training and testing on the same domain;

Cross-domain accuracy is $A(D_s \to D_t)$ where $D_s \neq D_t$;

Reduction is the proportional performance loss relative to the target domain's in-domain baseline:

$$\text{Reduction}(D_s \to D_t) = \frac{A(D_t \to D_t) - A(D_s \to D_t)}{A(D_t \to D_t)}. \quad (1)$$

This measures how much worse a model trained on D_s performs on D_t compared to a model trained directly on D_t, controlling for the intrinsic difficulty of D_t.

We report both absolute accuracy differences and proportional reduction relative to target-domain baseline; the latter controls for differences in target-domain difficulty, but can become unstable when baseline accuracy is low. Negative reduction values indicate cross-domain performance exceeding in-domain baseline, typically due to degraded in-domain performance rather than genuine improvement.

2.7. Experimental Design

For every test we train a model on the training data from one dataset and then test it on the test data of both datasets, comparing within-domain to cross-domain accuracy. We test the six normalization methods (Log, Log+median, Log+normalize, Log+median+normalize, PCEN, Box-Cox), DANN, and augmentation with Freesound noise at varying intensities (0.0, 0.2, 0.4, 0.6, 0.8) and diversity (1, 10, 100, 1000 noise samples). All experiments use five random seeds and we report mean±std over those five tests.

Table 1: Domain shift analysis across preprocessing methods. For each transfer direction (e.g., Waitākere→DOC), **In-Dom** reports source in-domain accuracy, **Cross-Dom** reports cross-domain accuracy, and **Red%** is the percentage reduction from the target domain's in-domain baseline (e.g., for Waitākere→DOC, reduction uses DOC→DOC as baseline). Mean±std over 5 seeds. **Bold**: best cross-domain performance.

Method	Waitākere→DOC			DOC→Waitākere		
	In-Dom	Cross-Dom	Red%	In-Dom	Cross-Dom	Red%
Log	35.1±2.6	30.7±3.6	46.9±6.7	57.8±2.4	21.7±4.5	38.2±13.6
Log+median	11.8±3.7	7.0±2.3	83.8±5.7	43.0±6.0	28.2±2.9	-137.7±78.4
Log+normalize	17.1±8.1	20.3±8.5	52.2±21.7	42.5±7.3	23.4±6.2	-36.9±74.5
Log+median+normalize	**37.5±3.2**	**37.2±2.6**	**32.2±6.2**	**54.9±3.2**	**30.6±5.0**	**18.4±15.0**
PCEN	29.1±3.2	32.5±3.2	42.6±6.6	56.5±3.2	20.3±3.1	30.2±13.1
Box-Cox	34.6±3.8	34.2±1.9	40.3±5.0	57.2±3.7	18.9±5.2	45.3±16.1

Table 2: Domain adversarial training (DANN) comparison. **In-Dom**: In-domain accuracy. **Cross-Dom**: Cross-domain accuracy. **Red%**: Reduction from target in-domain baseline. Mean±std over 5 seeds.

Method	Transform	Waitākere→DOC			DOC→Waitākere		
		In-Dom	Cross-Dom	Red%	In-Dom	Cross-Dom	Red%
Baseline	Log	35.1±2.6	30.7±3.6	46.9±6.7	57.8±2.4	21.7±4.5	38.2±13.6
DANN	Log	41.4±3.0	30.1±3.8	49.8±7.4	60.0±4.7	21.2±4.3	48.7±11.0
Baseline	Log+median+normalize	37.5±3.2	37.2±2.6	32.2±6.2	54.9±3.2	30.6±5.0	18.4±15.0
DANN	Log+median+normalize	36.9±3.6	37.5±2.4	39.1±5.4	61.6±3.7	30.3±2.9	17.9±11.3

3. Results

3.1. Normalization Methods

Log+median+normalize achieves the best cross-domain accuracy in both directions (Table 1), retaining 68% of the DOC baseline and 82% of the Waitākere baseline. All methods show substantial domain shift, with reduction percentages ranging from 18–84% depending on the method and transfer direction.

Some preprocessing methods (e.g., Log+median) catastrophically degrade in-domain performance, producing misleadingly large negative reduction values that indicate method failure rather than genuine cross-domain improvement. PCEN occasionally improves cross-domain performance beyond in-domain baselines (Waitākere→DOC: 32.5% cross vs 29.1% in-domain), though this likely reflects inconsistent in-domain optimization rather than true generalization gains.

3.2. Domain Adversarial Training

DANN provides minimal improvements over the Log baseline and no meaningful gains when combined with strong preprocessing (Table 2). Following Log+median+normalize normalization, DANN produces nearly identical cross-domain accuracy in both directions.

3.3. Noise Augmentation

Noise augmentation shows limited and inconsistent benefits (Figure 2). Moderate noise intensity (around 0.2) maintains cross-domain performance while slightly improving in-domain accuracy on Waitākere, though higher intensities (0.6–0.8) degrade it. Noise variety also shows little clear effect.

3.4. Per-Species Analysis

Domain shift effects vary substantially across species (Table 3). Note that the per-species accuracy is computed as independent binary classification per class and is not directly comparable to exact-match accuracy reported elsewhere.

4. Discussion

4.1. Substantial Domain Shift

Domain shift in bird call recognition remains substantial even with matched species distributions. Using the log baseline, cross-domain models achieve only 30.7% (Waitākere→DOC) and 21.7% (DOC→Waitākere), representing 46.9% and 38.2% reductions from target in-domain baselines.

4.2. Explicit Normalization Outperforms Learned Adaptation

Adaptive background subtraction (Log+median+normalize) achieves better cross-domain performance than domain adversarial training. Based on the approximately 700 training samples available for the data in this study, DANN provides little to no benefit. We hypothesize this occurs because explicit normalization already removes domain-specific structure, leaving insufficient signal for adversarial alignment. In low-data regimes, preprocessing targeting known variation sources (background noise, recording characteristics) may be preferable to learned adaptation methods, which may require larger samples to disentangle domain from task-relevant features.

4.3. Dataset Asymmetry

The DOC dataset consistently achieves higher in-domain accuracy than Waitākere. At the same time, the relative reduction when transferring to DOC is larger than when transferring to Waitākere. This would seem to indicate our model is learning better features from DOC data. This could be explained by the range of locations across the dataset. However, another possibility is that the DOC splitting strategy

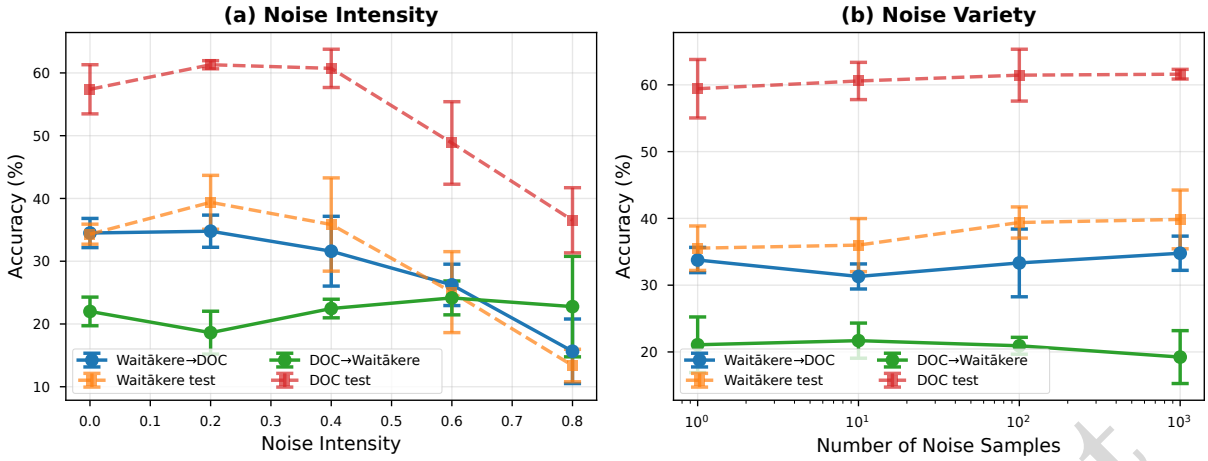

Figure 2: Noise augmentation effects on cross-domain performance under Log transform regime. (a) Accuracy by noise intensity. (b) Accuracy by noise variation (intensity set to 0.2). Error bars show standard deviation over 5 seeds.

Table 3: Per-species cross-domain performance using Log+median+normalize preprocessing. **%**: Percentage of dataset samples containing this species. **In**: In-domain accuracy (%). **Cross**: Cross-domain accuracy (%). **Red%**: Reduction from target domain's in-domain baseline (negative values indicate cross-domain performance exceeding target baseline). Mean±std over 5 seeds.

Species	%	Waitākere→DOC			DOC→Waitākere		
		In	Cross	Red%	In	Cross	Red%
Blackbird	9.3	92.2±2.0	90.3±0.4	0.3	90.6±1.8	86.0±3.5	6.7
Chaffinch	7.4	91.4±2.1	93.8±0.6	−0.5	94.3±1.4	86.3±6.2	5.6
Fantail	3.6	93.8±0.0	94.8±0.3	−0.3	94.5±0.6	90.2±5.3	3.8
Grey warbler	13.1	86.3±2.1	85.7±1.9	7.2	92.3±1.1	79.7±0.4	7.7
Kākā	4.0	92.0±1.4	91.2±0.3	2.1	93.2±1.1	95.2±4.2	−3.5
Morepork	25.1	89.4±3.6	98.1±1.8	1.2	99.3±0.5	86.9±1.6	2.8
Silvereye	13.4	86.2±1.7	88.1±0.4	6.8	94.5±1.5	80.0±3.2	7.2
Tomtit	9.2	91.7±1.7	69.4±3.1	24.1	91.4±2.6	93.1±0.0	−1.5
Tui/bellbird	36.5	80.0±2.1	76.2±3.5	8.7	83.5±4.0	83.7±5.2	−4.6

is artificially boosting its in-domain performance, and therefore also increasing the reduction when training on Waitākere instead. The DOC data-points are treated independently, since we have no information about the location or time of recording. In comparison, with the Waitākere data we perform train/test separation on the basis of recording sessions. This means a model could potentially learn features of a particular recording session in the DOC training set, where that same session also appears in its test set. To distinguish between these two possibilities we would need to assess the quality of learned features under both models, which would require a third dataset.

4.4. Limitations and Future Work

Our results have limited replication (five repetitions) and cover only nine classes (ten species) present in New Zealand with two recording sources and approximately 700 samples total. The datasets differ in multiple dimensions (equipment, geography, recording practices), preventing causal attribution to specific factors. We evaluate only one CNN-based architecture, although testing on an Audio Spectrogram Transformer [15] shows similar results. Our noise augmentation uses synthetic background noise; real environmental noise from target domains may provide more effective regularization. Aggregate accuracy obscures per-class patterns and potential annotation inconsistencies, and we infer mechanisms from accuracy patterns without direct feature analysis.

Acknowledgments

This research was funded by a Ministry of Business, Innovation and Employment Smart Ideas Grant (RSCHTRUSTVIC2443).
We thank the New Zealand Department of Conservation and Olly Powell for making the biodiversity monitoring recordings available, and Joe Turner-Steele and Monika Nowicki for annotation of the Waitākere dataset.

References

[1] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, e13152, 2022.

[2] N. Priyadarshani, S. Marsland, and I. Castro, "Automated birdsong recognition in complex acoustic environments: A review," *Journal of Avian Biology*, vol. 49, no. 5, jav–01 447, 2018.

[3] D. Stowell, T. Petrusková, M. Šálek, and P. Linhart, "Automatic acoustic identification of individual animals: Improving generalisation across species and recording conditions," *arXiv e-prints*, 2018. arXiv: 1810.09273.

[4] S. Marsland, N. Priyadarshani, J. Juodakis, and I. Castro, "AviaNZ: A future-proofed program for annotation and recognition of animal sounds in long-time field recordings," *Methods in Ecology and Evolution*, vol. 10, no. 8, pp. 1189–1195, 2019.

[5] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021.

[6] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, "Perch 2.0: The bittern lesson for bioacoustics," *arXiv preprint arXiv:2508.04665*, 2025.

[7] K. Darras, P. Pütz, K. Rembold, and T. Tscharntke, "Measuring sound detection spaces for acoustic animal sampling and monitoring," *Biological Conservation*, vol. 201, pp. 29–37, 2016.

[8] H. Venkateswara and S. Panchanathan, "Introduction to domain adaptation," in *Domain Adaptation in Computer Vision with Deep Learning*, H. Venkateswara and S. Panchanathan, Eds. Cham: Springer International Publishing, 2020, pp. 3–21.

[9] J. Turner-Steele, "Avian community composition and its variation across kauri forests in the Waitākere ranges infected with Phytophthora agathidicida," MSc thesis, Victoria University of Wellington, 2024.

[10] shionao7, *Regnrty008 (pytorch model)*, `https://www.kaggle.com/models/shionao7/regnrty008/PyTorch/default/1`, Accessed: 2026-03-29, 2024.

[11] V. Lostanlen et al., "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[12] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[13] A. S. Kumar, T. Schlosser, S. Kahl, and D. Kowerko, "Improving learning-based birdsong classification by utilizing combined audio augmentation strategies," *Ecological Informatics*, vol. 82, p. 102 699, 2024.

[14] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[15] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proceedings of Interspeech*, 2021.