Towards Domain Shift in Location-Mismatch Scenarios for Bird Activity Detection

Amir Latifi Bidarouni
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
amir.latifi.bidarouni@idmt.fraunhofer.de

Jakob Abeßer
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

***Abstract*—Bioacoustic monitoring serves as a valuable tool for gaining insights into the well-being of wildlife. Sensor locations with diverse acoustic conditions pose a major challenge for deep learning-based audio classification systems. In this paper, we study unsupervised domain adaptation techniques for the task of bird activity detection in short audio segments using two bird recognition datasets with recordings from diverse locations. Furthermore, we explore various distance and divergence metrics to quantify the domain shift as a proxy to predict the expected drop in classification accuracy at different recording locations. Our results confirm the superior performance of the instance-wise feature projection-based domain adaptation (IFPDA) technique across multiple audio domains and demonstrate that useful domain shift metrics can be derived from the energy distribution across frequency bands.**

***Index Terms*—Domain shift, domain adaptation, bioacoustic monitoring, bird activity detection**

I. Introduction

Biodiversity monitoring methods yield invaluable insights into animal species' presence, distribution, behavior, and abundance in different environments. Bioacoustic monitoring, in particular, allows to evaluate the health of ecosystems in a non-invasive manner and to inform strategic decisions in habitat conservation and management. Although large amounts of data can be recorded by distributed acoustic sensors, their annotation is very complex and requires domain-specific expert knowledge.

One of the main challenges of bioacoustic monitoring in diverse environments is the domain shift, i. e., the differences in the acoustic data distribution caused, for example, by different acoustic conditions at the sensor locations. Domain shift can cause a diminished performance of deep learning models when faced with audio recordings from novel sensor locations, which further complicates the data analysis and interpretation.

Utilizing domain adaptation (DA) techniques allows to reduce domain shift between the source domain (SD) dataset used to train the model and the target domain (TD) datasets upon which the model will perform predictions. DA methods can be classified into model-based and data-based methods. In model-based DA, neural network architectures are equipped with various regularization techniques to improve generalization to unseen TD data [1]–[3]. However, these methods require relatively large SD and TD datasets to yield robust results. Data-based DA methods modify the data so that the model can be trained once on SD data and then utilized on arbitrary TD datasets and are discussed in previous studies [4]–[6]. Although such models do not fully achieve the performance of models re-trained for every TD, they have demonstrated reasonable performance without requiring a resource-intensive and time-consuming retraining step [7].

Although domain adaptation methods have been studied mainly to reduce the performance gap between source and target domains, quantifying the existing domain shift between different datasets is an important data analysis step that has received so far less attention in research. Several studies have examined the relationship between the decrease in model performance and domain shift. For example, in [8] a CNN-based method is introduced to measure domain shift, applying to a medical image. In another study [9], the hidden representation of neural network data preceding the classifier layer was utilized as a data representation to assess domain shift. In contrast to these studies that employed neural networks for predicting performance drop in regards to domain distortion using embedding data representations, our research focuses on analyzing domain shift in the feature space utilizing statistical approaches. In the context of bioacoustic monitoring, domain shift measurement offers a means to quickly assess the "difficulty" of making prediction on data recorded at novel sensor locations and estimating the expected performance drop of the classification model.

In this paper, we focus mainly on studying domain shift caused by mismatching sensor locations. We study the task of bird activity detection, i. e., a binary classification task where the presence or absence of bird calls is predicted for short audio segments. We run our experiments on two large multi-location bird recognition datasets and select data recorded from one sensor location as SD and the remaining locations as TD. Then, we study different unsupervised domain adaptation techniques for domain shift compensation, which solely modify the data in the feature space without requiring any modification to the neural network architectures. Second, we apply several distance metrics based on the overall dynamic range distributions as well as the energy distribution across different frequencies to quantify domain shift between audio

recordings from different sensor locations.

II. Methodology

A. Feature extraction

In this paper, audio recordings with a sample rate of 44.1 kHz are converted into Mel spectrograms with 128 Mel bands. Logarithmic magnitude scaling is applied to reduce the dynamic range between foreground and background sounds. The feature extraction parameters, including FFT size, window size, and hop size, are selected as 2048, 1024, and 512, respectively, similar to [7] for comparability.

B. Data partitioning

Following the notation previously introduced in [7], we denote a batch as feature tensor $X \in \mathbb{R}^{B \times F \times T \times C}$ where B denotes the batch size (*batch dimension*), F denotes the number of frequency bins (*frequency dimension*) and T denotes the number of frames (*time dimension*). Furthermore, C denotes the number of channels (*channel dimension*), while in this study, we only use one channel ($C = 1$) to store the magnitude spectrogram.

In this study, we explore various data normalization with different partitioning strategies w.r.t. the frequency and batch dimensions. Regarding the batch dimension, we differentiate between normalizing "per instance" $X_{i,:,:,:}$ with $i \in [1 : B]$ and normalizing "per batch" across all instances of the batch. In contrast to 2D images where pixels convey spatial information, "pixels" in Mel spectrograms represent energy at specific frequencies and time positions. As a result, the statistical properties within the frequency bands provide additional insight, as demonstrated in [10]. Therefore, another partitioning approach would be that, in the frequency dimension, we also distinguish between normalizing "per frequency" $X_{:,j,:,:}$ with $j \in [1 : F]$ and normalizing "globally" across the entire frequency range. Furthermore, normalization can be performed "within domain" based on the statistics computed and applied within each domain separately or "between domains", where the statistics derived from the source domain are used for normalizing all domains.

C. Model architecture and training parameters

As in [7], we use the "CNN420" [5] network architecture with 799,050 trainable parameters. The model comprises an initial convolutional layer, four residual blocks, each incorporating a dropout layer with a rate of 0.1, a global average pooling layer, and a final dense layer with a softmax activation function. All batch normalization layers have been intentionally removed to systematically study different feature space domain adaptation techniques. We use a batch size of 32 and the Adam optimizer with a learning rate of 0.001 and train for 250 epochs. In addition, early stopping is used with patience of 75 epochs.

D. Distance and divergence measurement methods

Statistical methods such as divergences or distances are often used to quantify the difference/similarity between two datasets. For a metric to be classified as a distance, it must satisfy the conditions outlined in definition II.1.

Definition II.1 (Distance Metric [11], [12])**.** Given a set X, a real-valued function $d(x, y)$ on the Cartesian product $X \times X$ is a distance metric if for any $x, y, z \in X$, it satisfies the following conditions [11]:

1) $d(x, y) \geq 0$ (non-negativity),
2) $d(x, y) = d(y, x)$ (symmetry),
3) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality),
4) $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles).

Since divergence metrics are only required to fulfill the criterion of non-negativity and the identity of indiscernibles, they cannot be regarded as distance metrics from a mathematical perspective, but as a distortion measure that is exclusively applicable to probability distributions [13]. This distinguishes them from distances, which apply to various other data types as well. In the following, the methods used in this research are provided.

1) Kullback-Leibler divergence (KLD): The discrete form of the relative entropy or Kullback-Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same space is [14]:

$$d_{KLD}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{1}$$

Kullback-Leibler divergence is the entropy difference of two probability distributions and can be interpreted as a measure of data distortion or loss of information between one probability distribution and the reference distribution. It can also be considered as the number of bits, necessary to transform one distribution to another one.

2) Wasserstein distance (WSD): The Wasserstein distance is a distance metric used to compare two distributions and quantify the minimum amount of work required to transform one probability distribution into another.

Definition II.2 (P_{th} Wasserstein distance [15])**.** If $(X, d(x, y))$ be a Polish metric space, for any two probability distributions P and Q on X, the Wasserstein distance of order $p \geq 1$ between P and Q is:

$$d_{\mathrm{WS}}(P, Q) = \Big(\inf_{\pi \in \prod(P,Q)} \int_X d(x, y)^p \, d\pi(x, y) \Big)^{\frac{1}{p}} \tag{2}$$

Where inf refers to infimum, $\prod(P, Q)$ represents all joint probability measures on $X \times X$ with marginals P and Q and $d(x, y)$ is the distance between the x and y in X.

The first-order Wasserstein distance with $p = 1$, often referred to as the Earth Mover's Distance (EMD), is the specific form of the Wasserstein metric utilized in this investigation.

TABLE I: Experiment configurations with the corresponding normalization methods and dataset partition approach.

Configuration	Normalization Method	Partitioning Dimension		
		Batch (B)	**Frequency** (F)	**Domain**
Base	-	-	-	
Zs-IGW	Z-score	instance	global	within
Zs-IFW	Z-score	instance	frequency	within
Zs-BGW	Z-score	batch	global	within
Zs-BGB	Z-score	batch	global	between
Zs-BFW	Z-score	batch	frequency	within
Zs-BFB	Z-score	batch	frequency	between
Rf-IFW	RFN	instance	frequency	within
Rf-BFW	RFN	batch	frequency	within
Fp-IFW	FPDA (**IFPDA**)	instance	frequency	within

3) Cramér-Von Mises (CVM): The Cramér-Von Mises distance serves as a measure to evaluate whether a set of N empirical samples $x_1, x_2, ..., x_N$ has been drawn from a continuous distribution $F(x)$ [16]. Mathematically, it is expressed as:

$$d^2_{CVM} = \int_{-\infty}^{\infty} [F_N(x) - F(x)]^2 dF(x) \tag{3}$$

Here $F_N(x)$ denotes the empirical distribution function. In essence, this method allows to evaluate how closely $F_N(x)$ aligns with $F(x)$. Consequently, it can also be interpreted as a measure of how similar or dissimilar two distributions are to each other.

E. Domain adaptation methods

1) Z-score: Z-score normalization (standardization) is a prevalent method to rescale data to zero mean and unit variance.

2) Relaxed Instance Frequency-Wise Normalization (RFN): In computer vision, normalization is commonly performed globally, either on individual (instance) or multiple (batch) images. In audio processing, however, the frequency axis of spectrograms holds vital information, rendering frequency-wise normalization highly effective. RFN [10] is a frequency-wise normalization technique that incorporates instance-wise information into normalization outcomes using a relaxation factor $\lambda \in [0, 1]$ to preserve crucial global information from being lost.

$$RFN(x) = \lambda \cdot IN(x) + (1 - \lambda) \cdot FN(x) \tag{4}$$

where $IN(x)$ and $FN(x)$ are instance-wise and frequency-wise normalization, respectively.

3) Instance-Wise Feature Projection-Based Domain Adaptation (IFPDA): The IFPDA method [7] performs normalization in frequency dimension for each instance individually. The covariance matrix of each instance assesses the relationships among the frequency bands. The eigenvectors associated with the L highest eigenvalues of this matrix serve as a mapping. When the normalized instance is multiplied by this mapping, its magnitude is projected along the direction of the highest variation within the covariance matrix. The complete algorithm of IFPDA is presented in algorithm 1.

Algorithm 1 IFPDA [7]

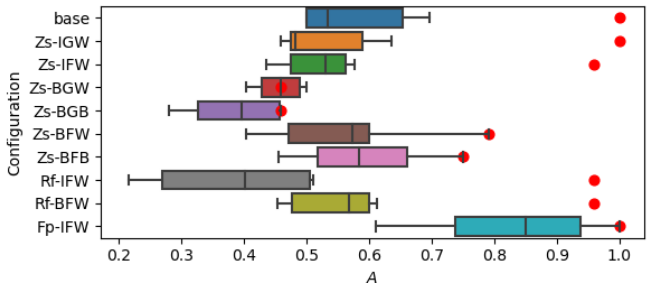
for $x_i^{T,F} = X_{i,1:T,1:F,:}$ `with` $i \in [1:B]$ **do**
  $x_i^{F,T} \leftarrow transpose(x_i^{T,F})$
  $\mu_i^{F,1} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(x_i^{F,T})$
  $\sigma_i^{F,1} \leftarrow \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^{F,T} - \mu_i^{F,1})^2}$ ▷ standard deviation
  $\bar{X}_i^{F,T} \leftarrow (x_i^{F,T} - \mu_i^{F,1}) \,/\, (\sigma_i^{F,1})$
  $Cov_i^{F,F} \leftarrow covariance(\bar{X}_i^{F,T})$ ▷ covariance of $\bar{X}_i^{F,T}$
  $V_i^{F,F} \leftarrow eigenvector(Cov_i^{F,F})$ ▷ eigenvectors of L highest $Cov_i^{F,F}$
  $\bar{X}_{\text{sub}}^{T,F} \leftarrow transpose(\bar{X}_i^{F,T}) \cdot V_i^{F,F}$
  $\hat{X}_i^{T,F} \leftarrow \bar{X}_{\text{sub}}^{T,F} \cdot transpose(V_i^{F,F})$
end for

III. Datasets

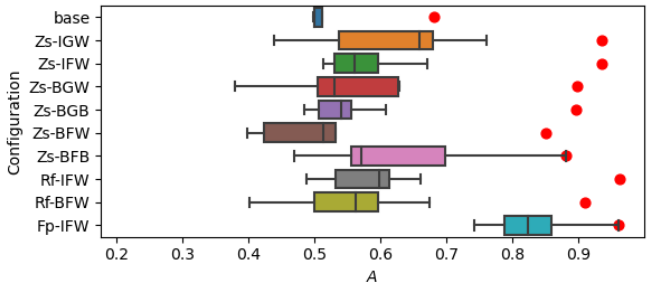
In this paper, we study two public bird monitoring datasets to measure domain shift between audio recordings from different sensor locations and to study the effectiveness of different domain adaptation methods. The BirdVox-296h (`BV`) dataset [17] contains 148 two-hour audio recordings, which have been captured with identical audio sensors at 9 recording sensors located near Ithaca, NY, USA, in the fall of 2015. The audio files are accompanied by sound event annotations that specify start times and a taxonomy code to classify the associated sound source.

The Southwestern Amazon Basin (`AMZ`) dataset [18] includes 21 fully-annotated one-hour soundscape recordings from 7 distinct locations within the Inkaterra Reserva Amazonica in early 2019. The dataset includes bounding box annotations to localize bird calls in both time and frequency for a taxonomy of 132 bird species, with a minimum distance of 5 seconds between consecutive bounding boxes.

In this paper, we study the task of predicting bird activity within 10 s segments of audio. To convert the event-level annotations of both datasets into segment-level bird activity annotations, we proceed as follows. All audio recordings are initially divided into a fixed set of 10 s long segments. Segments preceding the first annotated bird sound event in the original recording are classified as negative (no bird). In the `BV` dataset, which only provides starting points of bird calls, segments following the first sound event are marked as positive (has bird), if at least one bird call is detected within the segment. Segments without a starting point for bird sounds are excluded from the training pipeline due to uncertainty about the duration of each bird sound event. Conversely, in the `AMZ` dataset, where annotations include the start and end of each bird sound event, segments lacking any overlap with bird sound events are designated as negative, while those with overlap are classified as positive.

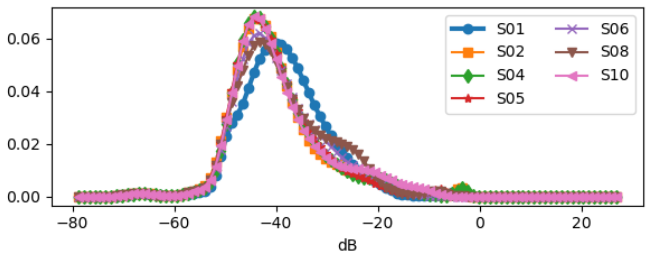


(a) AMZ dataset.

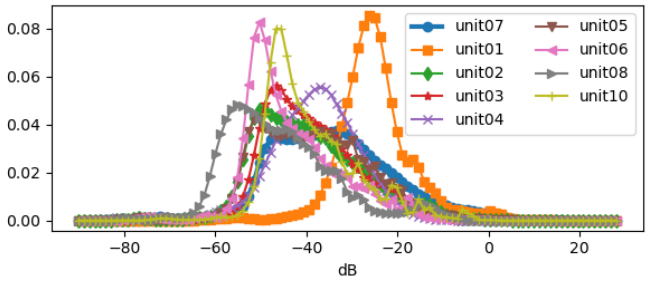


(b) BV dataset.

Fig. 1: Figure summarizing the accuracy (A) values obtained on the source domain data (red dot) and across multiple target domains (boxplot) for the AMZ dataset (a) and the BV dataset (b).



(a) AMZ dataset.



(b) BV dataset.

Fig. 2: Raw data probability distribution of different recording locations for the AMZ dataset (a) and the BV dataset (b).

IV. Evaluation

In the realm of domain adaptation and domain shift measurement, it is imperative to designate one dataset as the source domain while others serve as the target domains. In our case, we designate the unit07 subset of the BV dataset as the source domain due to its extensive file count with nearly balanced class labels and 47.62 % of all segments comprising bird sounds. In contrast, in the AMZ dataset, none of the locations exhibit balanced bird activity labels. Consequently, we designate the S01 subset as the source domain dataset and randomly select a subset of positive files in addition to the negative files to achieve a dataset with balanced classes.

A. Experiment 1 - Domain Adaptation

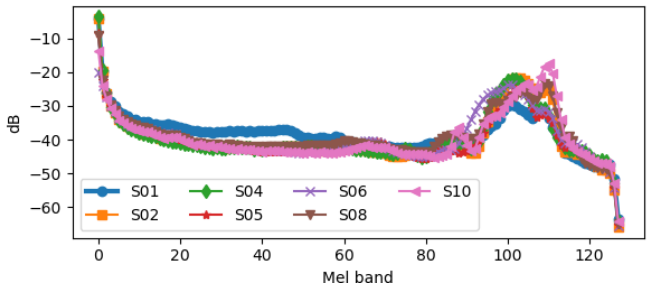
In conducting this experiment, we employ various domain adaptation techniques analogous to [7] to identify the most effective DA method for bioacoustic monitoring scenarios. Table I lists all combinations of DA methods and partitioning approaches ("Configurations"), which were examined by training the models on the SD dataset and evaluating them on the individual TD datasets. The average classification accuracy A for each model over all TD datasets is illustrated in Figure 1.

The results indicate that Zs-BGW and Zs-BGB exhibit the lowest performance in the source domain (SD) for the AMZ dataset, while Zs-BFW exhibits the lowest performance for the BV dataset, compared to the other methods examined. In addition, methods such as Zs-BGB and Rf-IFW on AMZ or Zs-BFW on BV demonstrate even lower accuracy than base models in target domains (TD), indicating poor generalization of the model to domains using these methods. As our main finding, IFPDA not only demonstrates high performance in the source domain but also in the target domains of both datasets, surpassing the other studied DA methods by a considerable margin. This finding is consistent with our previous study that focused on industrial and urban audio datasets [7].

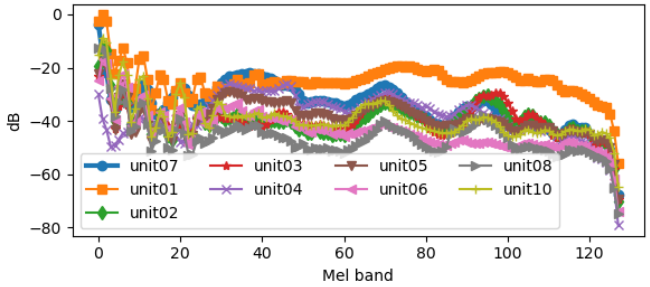
B. Experiment 2 - Quantifying Domain shift

To quantify the domain shift, we utilize the distance and divergence methods introduced in Section II-D as well as mean square error (MSE) to compare pairs of SD and TD datasets based on either a histogram over all Mel spectrogram magnitude values ("Magnitude") as shown in Figure 2 or a time-averaged Mel spectrogram ("Frequency") as shown in Figure 3. While the former one characterizes the overall dynamic range, the latter measures the energy distribution across frequencies for different recordings. For each configuration, we fit a Huber regression model (with $\epsilon = 1.35$, $\alpha = 0$), which is less sensitive to outliers compared to linear regression, to predict the accuracy drop between the source domain and the target domain by the computed domain shift between both. Naturally, we constrain the intercept to be zero as we do not expect an accuracy drop in case no domain shift can be measured. In this experiment, we use the data normalized with the best-performing method from Experiment 1 (IFPDA).

Table II summarizes the estimated regression coefficient and the coefficient of determination R^2 for each configuration. Although several models show poor performance ($R^2 < 0$), the energy distribution across frequencies ("Frequency") appears



(a) AMZ dataset.



(b) BV dataset.

Fig. 3: Mean frequency of raw data per mel band for different recording locations for the AMZ dataset (a) and the BV dataset (b).

TABLE II: Huber regression coefficients (HRC) and coefficient of determination R^2 for different domain shift measurements on both datasets.

Metric	Data representation	BV dataset HRC	BV dataset R^2	AMZ dataset HRC	AMZ dataset R^2
MSE	Magnitude	260580	-0.41	446414	-0.68
MSE	Frequency	42.59	-0.11	94.37	**0.10**
KLD	Magnitude	147.16	**0.08**	192.83	**0.01**
WSD	Magnitude	5365.72	-0.31	8533	-0.73
WSD	Frequency	82.36	**0.21**	61.75	-0.25
CVM	Magnitude	22.48	-2.71	15.72	-1.66
CVM	Frequency	65.77	-1.86	3390	**0.36**

to be better suited for domain shift measurement compared to the global dynamic range ("Magnitude") as the best models are WSD+Frequency ($R^2 = 0.21$) for the BV dataset and CMV+Frequency ($R^2 = 0.36$) for the AMZ dataset.

V. Conclusions

In this investigation, we explored various domain adaptation techniques for the detection of bird activity in bioacoustic datasets characterized by location discrepancies. Our findings confirm that the IFPDA method is not only the most efficient of the investigated DA methods for urban sound and industrial audio data sets, as shown in our previous work [7], but also in the field of bioacoustics. Furthermore, we systematically tested combinations of four distance and divergence metrics with two different audio representations to predict the drop in accuracy when applying a pre-trained model from the SD to the TD. We found that the energy distribution across frequency bands within a dataset provides a better representation for domain shift measurement compared to global statistics of the dynamic range. In general, positive regression coefficients confirm our intuition that the accuracy drop from the SD to the TD increases with increasing domain shift.

Acknowledgment

This study was supported by the German Research Foundation (Grant No. 350953655), funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101081964, as well as internal funding by Fraunhofer Society.

References

[1] X. Wang, P. Guo, and Y. Zhang, "Domain adaptation via bidirectional cross-attention transformer," 2022.

[2] Y. Shao, X. Ma, Y. Ma, and W.-Q. Zhang, "Thuee submission for DCASE 2020 challenge task1a," DCASE2020 Challenge, Tech. Rep., June 2020.

[3] L. Jie, "Acoustic scene classification with residual networks and attention mechanism," DCASE2020 Challenge, Tech. Rep., June 2020.

[4] B. Sun, J. Feng, and K. Saenko, "Return of Frustratingly Easy Domain Adaptation," *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 2058–2065, 2016.

[5] D. Johnson and S. Grollmisch, "Techniques improving the robustness of deep learning models for Industrial Sound Analysis," in *Proceedings of the 2020 European Signal Processing Conference (EUSIPCO)*, Online, 2021, pp. 81–85.

[6] A. I. Mezza, E. A. P. Habets, M. Müller, and A. Sarti, "Feature Projection-Based Unsupervised Domain Adaptation for Acoustic Scene Classification," in *Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, 2020, pp. 1–6.

[7] A. L. Bidarouni and J. Abeßer, "Unsupervised feature-space domain adaptation applied for audio classification," in *2023 4th International Symposium on the Internet of Sounds*, Pisa, Italy, 2023, pp. 1–7.

[8] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "Measuring domain shift for deep learning in histopathology," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 325–336, 2021.

[9] H. ElSahar and M. Gallé, "To annotate or not? predicting performance drop under domain shift," in *Conference on Empirical Methods in Natural Language Processing*, 2019.

[10] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in *Proceedings of the INTERSPEECH conference*, Incheon, Korea, 2022, pp. 2393–2397.

[11] S. Chen, B. Ma, and K. Zhang, "On the similarity metric and the distance metric," *Theoretical Computer Science*, vol. 410, no. 24, pp. 2365–2376, 2009, formal Languages and Applications: A Collection of Papers in Honor of Sheng Yu.

[12] J. Muscat, *Distance*. Cham: Springer International Publishing, 2014, pp. 13–26.

[13] R. M. Gray, *Relative Entropy*. Boston, MA: Springer US, 2011, pp. 173–218.

[14] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. USA: Copyright Cambridge University Press, 2003.

[15] C. Villani, *The Wasserstein distances*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 93–111.

[16] T. W. Anderson, "On the Distribution of the Two-Sample Cramer-von Mises Criterion," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1148 – 1159, 1962.

[17] A. Farnsworth, S. Kelling, V. Lostanlen, J. Salamon, A. Cramer, and J. P. Bello, "BirdVox-296h: a large-scale dataset for detection and classification of flight calls," Jan. 2022.

[18] W. A. Hopping, S. Kahl, and H. Klinck, "A collection of fully-annotated soundscape recordings from the Southwestern Amazon Basin," Sep. 2022.