Overview of BirdCLEF 2024: Acoustic Identification of Under-studied Bird Species in the Western Ghats

Stefan Kahl[1,2,*], Tom Denton[3], Holger Klinck[1], Vijay Ramesh[1,4], Viral Joshi[5], Meghana Srivathsa[4], Akshay Anand[4], Chiti Arvind[5], Harikrishnan CP[5], Suyash Sawant[5], Robin V V[5], Hervé Glotin[6], Hervé Goëau[7], Willem-Pier Vellinga[8], Robert Planqué[8] and Alexis Joly[9]

[1]*K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, USA*

[2]*Chemnitz University of Technology, Chemnitz, Germany*

[3]*Google Deepmind, San Francisco, USA*

[4]*Project Dhvani, Bangalore, India*

[5]*Indian Institute of Science Education and Research (IISER) Tirupati, Tirupati, India*

[6]*University of Toulon, AMU, CNRS, LIS, Marseille, France*

[7]*CIRAD, UMR AMAP, Montpellier, France*

[8]*Xeno-canto Foundation, Groningen, Netherlands*

[9]*Inria, LIRMM, University of Montpellier, CNRS, Montpellier, France*

Abstract

The BirdCLEF 2024 challenge focused on the acoustic identification of understudied bird species in the Western Ghats, a biodiversity hotspot in India. This edition aimed to advance passive acoustic monitoring by tasking participants with developing reliable systems for detecting and identifying bird vocalizations from extensive soundscape recordings. Using training data provided by the Xeno-Canto community and new unlabeled soundscapes from the Western Ghats, participants addressed the challenges of domain adaptation and limited training data for many species. Participants employed techniques such as pseudo-labeling, test-time augmentation, and diverse ensembles, significantly improving model performance. Notable strategies also included the use of single-class cross-entropy and Contrastive Adversarial Domain (CAD) bottlenecks, which provided innovative solutions to acoustic data analysis challenges. The highest-scoring submission achieved an ROC-AUC score of 0.690 on the private leaderboard (0.738 on the public leaderboard), with the top 10 systems differing by only 1.5% in their scores.

Keywords

LifeCLEF, bird, song, call, species, retrieval, audio, collection, identification, fine-grained classification, evaluation, benchmark, bioacoustics, passive acoustic monitoring, PAM

1. Introduction

Passive acoustic monitoring (PAM), which uses autonomous recording units (ARUs) to study animals and their habitats at ecologically meaningful scales, has become an essential method in conservation [1]. The availability of affordable, off-the-shelf ARUs has enabled extensive data collection efforts in many regions worldwide. Typically, arrays of these recorders are deployed for long durations (weeks to months), producing large volumes of data that provide valuable insights into the abundance and distribution of vocalizing animals with high spatial and temporal resolution [2]. However, PAM faces several ongoing challenges. Data collection efforts can result in many terabytes of acoustic data that must be efficiently managed, stored, and analyzed [3]. In particular, the task of analyzing this data—reliably extracting relevant signals from often complex soundscapes—is still an active area of research. Additionally, while ample data for common species is usually available to train models, data for rare, listed, or endangered species is often scarce. This scarcity necessitates the development of innovative algorithmic approaches to monitor these species effectively.

The Western Ghats is a mountain range that runs along the southwestern coast of India [4]. This region is home to very high levels of biodiversity and supports the livelihoods of millions of people. Over 500 bird species have been reported in this region, of which several species are rare, endangered, and endemic (see figure 1). Automated identification of calls from different species is challenging in this region due to the high number of vocalizing bird species resulting in complex soundscapes with frequently overlapping calls.

The *Bird Recognition Challenge* (BirdCLEF) is an integral part of LifeCLEF 2024 [5], aimed at developing robust analytical frameworks for detecting and identifying bird vocalizations in continuous soundscape recordings. Initiated in 2014, BirdCLEF has grown into one of the largest bird sound recognition contests, featuring tens of thousands of recordings representing up to 1,500 species [6, 7]. The 2024 edition of BirdCLEF tasks participants with creating reliable systems for identifying bird calls within soundscapes from the Western Ghats, despite the challenge of having limited training data for many species.

2. BirdCLEF 2024 Competition Overview

Recent progress in machine listening techniques for identifying animal vocalizations has significantly improved our ability to analyze long-term acoustic datasets comprehensively [8, 9]. Nevertheless, achieving high precision and recall remains challenging, especially when dealing with numerous species simultaneously. A key difficulty in acoustic event detection and classification lies in bridging the gap between high-quality training samples (focal recordings) and noisy test samples (soundscape recordings). The 2024 BirdCLEF competition, hosted on Kaggle[1], tackled this complex issue by tasking participants with identifying bird calls in soundscape recordings from the Western Ghats in India. The competition followed the "code competition" format, encouraging participants to share their code for the benefit of the community, particularly scientists and practitioners monitoring bird populations for conservation in India. Additionally, submissions were required to complete inference within two hours to ensure the models could run efficiently on the modest computing resources available to conservationists.

(a) Black-and-orange Flycatcher

(b) Chestnut-headed Bee-eater

(c) Gray-headed Canary-Flycatcher

(d) Velvet-fronted Nuthatch

Figure 1: More than 500 bird species have been reported in the Western Ghats, of which several species are rare, endangered, and endemic. 108 were featured in this year's competition. Photos: Chandrasekar Das

2.1. Goal and Evaluation Protocol

This year's competition featured two major changes compared to the previous few years: A new metric was used for evaluation (macro-averaged ROC-AUC that skips classes that have no true positive labels), and inference was limited to two CPU hours.

2.2. Metric

This year, we used class-averaged ROC-AUC as the competition metric. ROC-AUC is best considered a rank-based metric: it is the probability that a positive example scores higher than a negative example when the positive and negative examples are independently chosen uniformly at random. We compute the ROC-AUC independently for each class present in the test data and then average over classes to obtain the model score.

As a threshold-free metric, ROC-AUC allows comparing overall model quality, without requiring participants to engage in difficult (and opaque) threshold-selection processes. It is also, by construction, indifferent to the positive/negative label balance within the dataset, though values can be noisy for extremely rare classes [10].

2.3. Time Limits

Competitors were limited to two hours of inference time on a CPU. This ensures that models are cost-effective for real-world usage. A side effect is reducing the impact of *ensembling*, a common Kaggle tactic obscuring underlying model quality.

2.4. Dataset

2.4.1. Training Data

As in previous editions, the training data for the competition was sourced from the Xeno-Canto community, comprising over 25,000 recordings spanning 182 species. Participants were permitted to use metadata to enhance their systems and to download/utilize additional Xeno-Canto recordings. Additionally, we offered detailed information on the locations and times of both focal and soundscape recordings, enabling participants to consider the spatio-temporal occurrence patterns of bird species in their analyses.

In addition, we supplied 8,444 unlabeled soundscape recordings from the same sites as the test data, though recorded on different dates to ensure no overlap. Participants were allowed to use these recordings to fine-tune their models or apply them for unsupervised learning during model training.

2.4.2. Test Data

As in previous years on Kaggle, the test data was completely hidden from participants. Hidden test data consisted of 1,073 soundscape recordings of 4-minute duration and were recorded at multiple locations within the Western Ghats. Most of the audio data was collected across the Anamalai and the Palani hills. These hill ranges largely consist of mid-elevation tropical wet evergreen rainforests and span an elevational gradient of ∼700 meters to 2,300 meters above sea level.

Acoustic data were collected as part of an ongoing project to assess the impacts of ecological restoration work on bird diversity. Across a gradient of forest regeneration (consisting of actively restored, naturally regenerating, and undisturbed benchmark forest sites, see Figure 2), AudioMoth ARUs were deployed to collect acoustic data [11]. These passive monitoring devices were placed on trees, approximately 2 meters above the ground at each site. Using a sampling rate of 48 kHz and a gain of 40 dB, each recorder was deployed to record data in 4-minute segments every 5 minutes for seven consecutive days at each site between March 2020 to January 2021. (Data could not be collected in April 2020 due to the COVID-19 pandemic). For more details, please see [12].
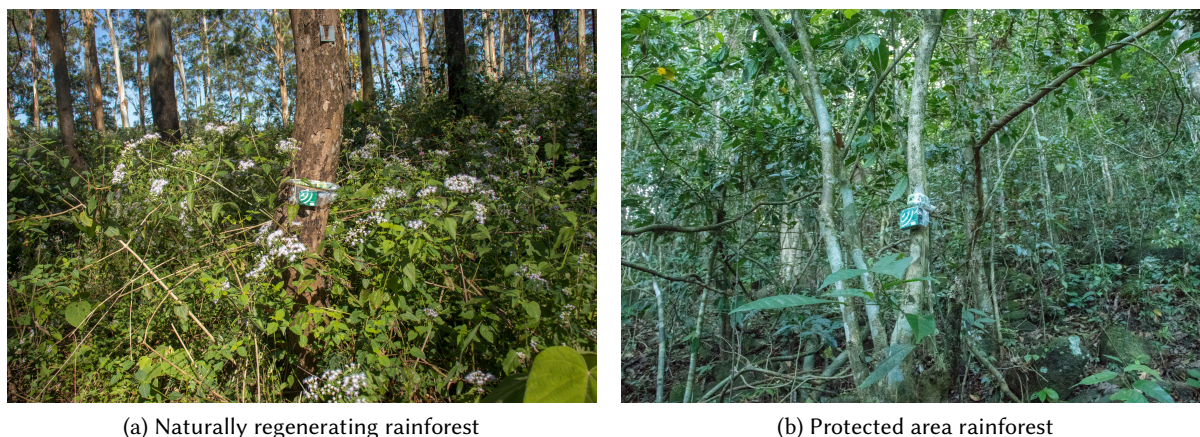

(a) Naturally regenerating rainforest   (b) Protected area rainforest

Figure 2: Most of the audio data for this competition were collected using AudioMoth ARUs deployed across a gradient of forest regeneration in the Anamalai and Palani hills. Photos: Vijay Ramesh

We identified all vocalizing bird species at a given site on a subset of the data recorded across each site. Each audio segment was broken down into 10-second audio segments for bird species identification. This was the shortest time period necessary to identify vocalizing bird species accurately. The annotation process resulted in 13,701 labels for 108 species.

3. Results

A total of 974 teams with nearly 1,200 competitors participated in the BidCLEF 2024 competition, submitting a total of 30,118 runs. As in recent years, two-thirds of the test data was allocated to the private leaderboard and one-third to the public leaderboard. Based on the ROC-AUC metric, the baseline score was 0.5, with random confidence scores for all birds across all segments. The highest-scoring submission achieved 0.690 (0.738 on the public leaderboard), with the top 10 systems differing by only 1.5% in their scores. There was a notable shake-up in the ranking between the public and the private leaderboard. While the top teams largely maintained their positions, many lower-ranked teams experienced significant drops due to the influence of a highly effective public code notebook[2], which led to many ranks being assigned based on execution date.

3.1. Online write-ups

A few common themes from online write-ups[3] emerged in the top solutions: the use of pseudo-labeling for unlabeled data, the implementation of test-time augmentation, and the deployment of diverse ensembles.

The public unlabeled data was a new addition to this year's competition, and perhaps unsurprisingly, many of the top competitors found ways to take advantage of it. Pseudo-labeling in this context provides aspects of both domain adaptation and knowledge distillation. Domain adaptation helps models cope with distributional differences between the train and test data: in the bioacoustic context, this includes changes in class frequency, geographic variation in vocalizations (dialects), and differences in recording characteristics (signal-to-noise ratio, device characteristics, and/or compression artifacts). When only unsupervised data is available for adaptation, as in this competition, the problem is known as *source-free domain adaptation* (SFDA). The SFDA task is particularly challenging in the multi-class, multi-label context [13]. Pseudo-labeling can also be interpreted as a form of knowledge distillation, as the pseudo-labels can be produced by large, pre-trained models (or ensembles); many of the top teams used models too slow for submission (such as the Google Perch classifier) or larger ensembles to produce pseudo-labels on the unlabeled data and the weakly-labeled Xeno-canto data.

Most of the top competitors also used a specific form of test-time augmentation: producing predictions for time-shifted audio windows and averaging with the predictions for the target window. This provides diverse views of the target data for the ensemble.

Finally, two competitors (in 4th and 5th place) produced a raw-waveform model, which ran in an ensemble with the standard spectrogram models. While these models underperformed spectrogram-based models individually, they improved the overall ensemble, presumably by obtaining diverse features from the audio. These competitors were the highest-ranking competitors who did not use pseudo-labeling, which suggests that this is a strong technique, orthogonal to pseudo-labeling.

Overall, the message from the top competitors is clear: robust pseudo-labeling strategies and diverse ensembles (whether from test-time augmentation or raw-waveform members) consistently made a significant impact. Two unique strategies were also notable among the top ten submissions. The first-place submission employed single-class cross-entropy for training, noting that multi-label samples were relatively rare in the unlabeled data. This approach provided strong regularization during model training but also necessitated additional efforts to generate meaningful per-class predictions at test time. The ninth-place submission utilized a Contrastive Adversarial Domain (CAD) bottleneck to obtain domain-invariant features [14], ensuring that model embeddings for the training data were indistinguishable from those of the unlabeled in-domain data, effectively minimizing domain-shift issues.

---

[2]https://www.kaggle.com/code/zulqarnainalipk/birdclef-2024-species-identification-from-audio
[3]Individual write-ups can be accessed via the "Solution" icon on the leaderboard: https://www.kaggle.com/competitions/birdclef-2024/leaderboard
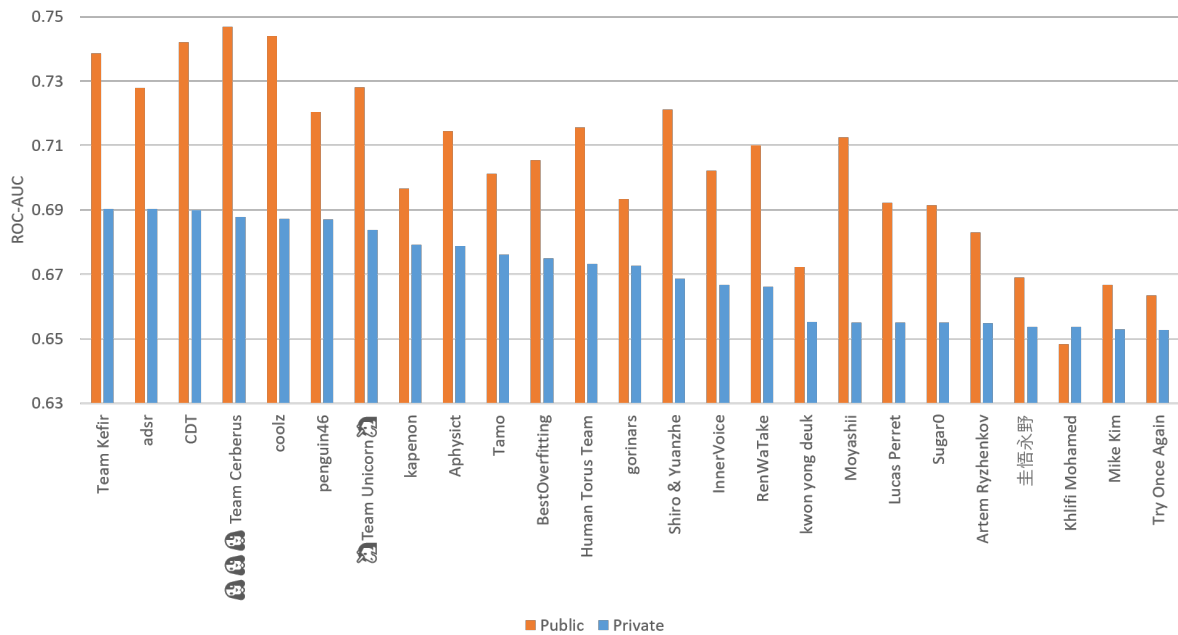
Figure 3: Top 25 private leaderboard scores achieved by the best systems evaluated within the bird identification task of LifeCLEF 2024. Public and private test data were split randomly. The private scores remained hidden until the submission deadline. Participants were able to optimize the recognition performance of their systems based on public scores, which likely explains some differences in scores.

3.2. Working notes

We accepted seven working notes for the proceedings, which document the approaches and methodologies used by individual teams:

Dmitriev, Konstantin V. [15]: The author used semi-supervised and self-supervised labeling to create pseudo-labels for unlabeled datasets, applied data augmentation techniques like MixUp and CutMix, and employed advanced post-processing such as sliding window averaging. Data preprocessing methods standardized recording lengths, and additional noise sources such as traffic, human voices, and weather sounds were incorporated to improve model generalization. Location data was utilized to address geographical variations in bird calls, and inference time optimization was achieved using techniques like weight rounding and conversion to efficient frameworks such as ONNX and OpenVino. The highest score achieved by the participant was a public leaderboard score of 0.684 and a private leaderboard score of 0.637[4].

Hong, Lihang [16]: This participant employs semi-supervised and self-supervised labeling of soundscapes, knowledge distillation, and data augmentation. Off-the-shelf models BirdNET [8] and the Google Bird Vocalization Classifier[5] were used to label large unlabeled datasets, which were then employed in training. Data augmentation techniques such as MixUp and CutMix were used. The combined approach of using labeled soundscapes and knowledge distillation significantly improved performance, achieving a maximum private leaderboard score of 0.681 (public leaderboard score 0.695).

Witting et al. [17]: The authors implemented a combination of data augmentations and pre- and post-processing techniques to improve model robustness. Specifically, they used noise reduction methods, location-specific data augmentation, and temporal context adjustments. The best-performing models incorporated spectrogram-based architectures enhanced with pseudo-labeling and test-time

---

[4]The highest scores in the working notes don't always match the official leaderboard scores because participants choose two runs for official scoring based only on public leaderboard performance.
[5]https://www.kaggle.com/models/google/bird-vocalization-classifier

augmentation, achieving a maximum private leaderboard score of 0.651 and a public leaderboard score of 0.738.

Lasseck, Mario [18]: The approach of this participant involves creating pseudo-labels for a large number of unlabeled recordings from the target location and using them in training. The best-performing models utilized the EfficientNetB0 architecture with MixUp and CutMix augmentations. The method includes pre- and post-processing techniques such as noise reduction, location-specific data augmentation, and temporal context adjustments. Extensive experiments showed that these strategies significantly improved performance, achieving a maximum ROC-AUC of 0.728 on the public leaderboard and 0.690 on the private leaderboard.

Kumar et al. [19]: This team employed methods like using pseudo-labels for large unlabeled datasets, data augmentations like MixUp and CutMix, and noise reduction techniques to overcome the shift in acoustic domains. The best-performing models utilized ViT (Vision Transformer) and DeiT (Data-efficient image Transformers) architectures with positional encoding to improve spatial context. The training process involved cosine annealing and weighted sampling, and the use of the transformer model presented some challenges, such as increased computational requirements and the need for extensive pre-training. Despite these constraints, the team achieved a maximum private leaderboard score of 0.629 (public leaderboard score 0.638).

Miyaguchi et al. [20]: This team investigated the distributional shift caused by the addition of unlabeled soundscapes, representative of the hidden test set, by using transfer learning for birdcall classification with embeddings from pre-trained models like Google's Bird Vocalization Classification Model, BirdNET, and EnCodec[21]. They experimented with different training losses, including Binary Cross-Entropy, Asymmetric Loss, and sigmoidF1, and proposed a pseudo multi-label classification strategy to utilize the unlabeled data. Efficient framework conversions and targeted optimizations addressed computational challenges posed by restricted inference runtime. The best-performing models achieved a maximum private score of 0.586 (public 0.556).

Porwal, Aaditya [22]: In this working note, the participant details an approach using an ensemble of EfficientNet-B0 and EfficientNet-B1 models. EfficientNet-B0 was exclusively trained on this year's data with heavy augmentations, while EfficientNet-B1 was pre-trained on previous datasets. Mel spectrograms were used for audio preprocessing, enhanced by augmentations like mixup and masking. The ensemble method, combining predictions from both models, achieved a maximum private score of 0.653 and a public score of 0.663

4. Conclusions and Lessons Learned

Many top-performing solutions leveraged pseudo-labeling techniques to effectively use the unlabeled soundscape data, demonstrating the importance of domain adaptation in improving model accuracy. Using diverse ensemble models, combining predictions from various architectures and configurations proved critical for enhancing performance and robustness in acoustic bird identification. Addressing the domain shift between high-quality training samples and noisy, real-world test soundscapes remains a major challenge. Successful strategies included using domain adaptation techniques and robust data augmentation methods like MixUp and CutMix. Balancing model complexity and inference time within the two-hour CPU limit posed a significant challenge, leading to the development of more efficient algorithms and optimization strategies. This greatly improves the real-world applicability of the developed approaches and models. Submitted solutions included some innovative approaches: The first-place submission utilized single-class cross-entropy for training, which provided strong regularization and improved performance despite the rarity of multi-label samples. CAD was used to obtain domain-invariant features, effectively minimizing the domain-shift issues and enhancing model robustness. Additionally, Integrating raw-waveform models with traditional spectrogram-based models in ensembles provided diverse feature sets and improved overall performance.

Acknowledgments

All results, code notebooks, and forum posts are publicly available at:
https://www.kaggle.com/c/birdclef-2024

References

[1] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, D. Llusia, Terrestrial passive acoustic monitoring: review and perspectives, BioScience 69 (2019) 15–25.

[2] L. S. M. Sugai, C. Desjonqueres, T. S. F. Silva, D. Llusia, A roadmap for survey designs in terrestrial acoustic monitoring, Remote Sensing in Ecology and Conservation 6 (2020) 220–235.

[3] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, et al., Perspectives in machine learning for wildlife conservation, Nature communications 13 (2022) 1–15.

[4] N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. Da Fonseca, J. Kent, Biodiversity hotspots for conservation priorities, Nature 403 (2000) 853–858.

[5] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, et al., Overview of lifeclef 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[6] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, R. Ruiz De Castañeda, I. Bolon, H. Glotin, R. Planqué, W.-P. Vellinga, A. Dorso, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of LifeCLEF 2021: a System-oriented Evaluation of Automated Species Identification and Species Distribution Prediction, in: Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), 2021.

[7] S. Kahl, M. Clapp, W. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, A. Joly, Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments, in: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.

[8] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236.

[9] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, H. Klinck, Deep neural networks for automated detection of marine mammal species, Scientific reports 10 (2020) 1–12.

[10] B. van Merriënboer, J. Hamer, V. Dumoulin, E. Triantafillou, T. Denton, Birds, bats and beyond: Evaluating generalization in bioacoustics models, Frontiers in Bird Science 3 (2024) 1369756.

[11] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, A. Rogers, Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment, HardwareX 6 (2019) e00073.

[12] V. Ramesh, P. Hariharan, V. Akshay, P. Choksi, S. Khanwilkar, R. DeFries, V. Robin, Using passive acoustic monitoring to examine the impacts of ecological restoration on faunal biodiversity in the western ghats, Biological Conservation 282 (2023) 110071.

[13] M. Boudiaf, T. Denton, B. Van Merrienboer, V. Dumoulin, E. Triantafillou, In search for a generalizable method for source free domain adaptation, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 2914–2931. URL: https://proceedings.mlr.press/v202/boudiaf23a.html.

[14] Y. Ruan, Y. Dubois, C. J. Maddison, Optimal representations for covariate shift, 2022. arXiv:2201.00057.

[15] K. V. Dmitriev, Methods for training convolutional neural networks to identify bird species in complex soundscape recordings, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[16] L. Hong, Domain Adaption for Birdcall Recognition: Progressive Knowledge Distillation with Semi-Supervised and Self-Supervised Soundscape Labeling, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[17] E. Witting, J. Lim, H. de Heer, C. T. Kopar, K. Sándor, Addressing the Challenges of Domain Shift in Bird Call Classification for BirdCLEF 2024, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[18] M. Lasseck, Improving Bird Recognition using Pseudo-Labeled Recordings from the Target Location, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[19] A. S. Kumar, T. Schlosser, D. Kowerko, TUC Media Computing at BirdCLEF 2024: Improving Birdsong Classification Through Single Learning Models, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[20] A. Miyaguchi, A. Cheung, M. Gustineli, A. Kim, Transfer Learning with Pseudo Multi-Label Birdcall Classification for DS@GT BirdCLEF 2024, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.

[21] A. Défossez, J. Copet, G. Synnaeve, Y. Adi, High fidelity neural audio compression, arXiv preprint arXiv:2210.13438 (2022).

[22] A. Porwal, Bird-Species Audio Identification, Ensembling of EfficientNet-B0 and Pre-trained EfficientNet-B1 model, in: CLEF Working Notes 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, 2024.