Domain Adaption for Birdcall Recognition: Progressive Knowledge Distillation with Semi-Supervised and Self-Supervised Soundscape Labeling⋆

Notebook for the BirdCLEF Lab at CLEF 2024

Lihang Hong[1,*,†]

[1]*Accenture Japan Ltd, Akasaka Intercity 1-11-44 Akasaka, Minato-ku, Tokyo, 107-8672, Japan*

Abstract

We present working notes for the BirdCLEF 2024 competition, focused on recogizing Indian bird species in soundscape recorded in Western Ghats. In this study, first, we utilize existing off-the-shelf models, BirdNET and Bird Vocalization Classifier, to address labeling challenges for training soundscapes from the same recording locations as the test soundscapes. Second, with the semi-supervised labeled soundscape, we execute a cycle of knowledge distillation training, self-supervised re-labeling and knowledge distillation training again. Our goal is to address the challenge of domain shift between train audio which focus on a certain species and test soundscape, and to maximize the performance of models. The solution based on the study achieves 7th rank among 974 teams at BirdCLEF 2024 challenge hosted in Kaggle.

Keywords

BirdCLEF2024, audio, bird species recognition, Semi-supervised, Self-supervised, Knowledge Distillation, Domain Adaption, CEUR-WS

1. Introduction

The rapid decline in global biodiversity has become a significant concern in recent years, putting numerous species at risk of extinction and threatening the stability of ecosystems. As birds serve as important indicators of biodiversity change, monitoring their populations is essential. Traditional bird surveys, which primarily rely on direct observation and human expertise, can be resource-intensive and face logistical challenges when applied at large scales and high temporal resolutions. This highlights the need for more efficient, scalable, and cost-effective methods to monitor bird populations. Advancements in passive acoustic monitoring (PAM) technology, combined with innovative machine learning algorithms, present a promising solution to these challenges.

Western Ghats are a biodiversity hotspot, home to diverse ecosystems and bird species, including those that are endemic and endangered. However, these ecosystems are threatened by landscape and climate changes. The aim of BirdCLEF 2024[1][2] is to develop conservation technologies to carry out automated detection and classification of bird species of the Western Ghats from soundscapes.

2. Domain Shift Challenge in Birdcall Recognition

The BirdCLEF 2024 competition focuses on recogizing Indian bird species in fully annotated 4-minute test soundscapes recorded in Western Ghats, which we call fully-annotated dataset. Two types of dataset are provided for training. One dataset, which we call weakly labeled dataset, comprises of short audios with ground truth label from Xeno-canto[3]. Another dataset, which we call unlabeled dataset, comprises of soundscapes without ground truth label recorded in the same locations as the fully-annotated dataset.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

rihanneko@gmail.com (L. Hong)
0009-0006-7840-7857 (L. Hong)

The model trained with weakly labeled dataset poses domain shift challenges when predicting fully-annotated dataset[4]. The challenges are:

1. Covariate shift. Short audios usually focus on one certain species and the bird call appears in the foreground. However, in soundscape, usually there are several species speaking over each other in the background. Making the classification model trained on short audios applicable to soundscape is very important because scientists need to identify birds recorded in a relatively noisy environment, while short audios are cost-effective as training data.
2. Label shift. Label shift can occur due to a variety of reasons such as seasonal variations in bird species and geographical disparities. For instance, the short audio may include a higher proportion of certain bird species that are not as prevalent in fully-annotated dataset. The implications of label shift are significant, as it can lead to biased predictions and poor model performance. If the model is trained on a dataset with a high proportion of certain bird species, it might over-predict these species in fully-annotated dataset. Conversely, it might under-predict species that were less prevalent in the training audio but more common in fully-annotated dataset.

Under the hypothesis that unlabeled dataset share a similar distribution with fully-annotated dataset, we focus our efforts on addressing domain shift challenge by labeling unlabeled dataset with semi-supervised and self-supervised approach. After labeling unlabeled dataset, we train the model with the union of weakly labeled dataset and unlabeled dataset.

3. Method

3.1. Dataset

3.1.1. Short Audio from Xeno-canto

As in previous BirdCLEF challenges, training data is provided by the Xeno-canto community. 24459 short audios covering 182 species are provided by the competition host. To further expand the dataset size, we collect additional 25710 short audios from Xeno-canto community. For pretraining, audios from previous BirdCLEF challenges were included [5][6][7][8]. The total dataset size was 234104 covering 992 species. We call short audios from Xeno-canto weakly labeled dataset.

3.1.2. Semi-supervised Labeled Soundscape

In addition to weakly labeled dataset, 8444 unlabeled soundscapes recorded in the same locations as the fully-annotated dataset are provided by the competition host, which we call unlabeled dataset. we utilize existing off-the-shelf models, BirdNET[9] and Bird Vocalization Classifier[10], to extract audio clip with high probability of birdcall presence. We call audio clips extracted from soundscapes with BirdNet and Bird Vocalization Classifier semi-supervised unlabeled dataset.

BirdNET is able to predict presence of all competition species except Nilgiri Wood Pigeon, while Bird Vocalization Classifier is able to predict presence of Nilgiri Wood-Pigeon. We process every soundscape using BirdNet to extract a prediction logit vector in 181 dimensions for every 3-second interval and using Bird Vocalization Classifier to to extract a prediction logit vector in 1 dimension for every 5-second interval. With the prediction logit, we extract 15-second audio clip with birdcall presence probability larger than 30 percent.

34829 audio clips are extracted from 5162 soundscapes. Comparation of species distribution between weakly labeled dataset and semi-supervised unlabeled dataset is shown in Figure 1.

As we can see in Figure 1, species distribution of weakly labeled dataset differs from that of semi-supervised unlabeled dataset, indicating the existance of label shift between weakly labeled dataset and fully-annotated dataset, under the hypothesis that unlabeled dataset share a similar distribution with fully-annotated dataset.

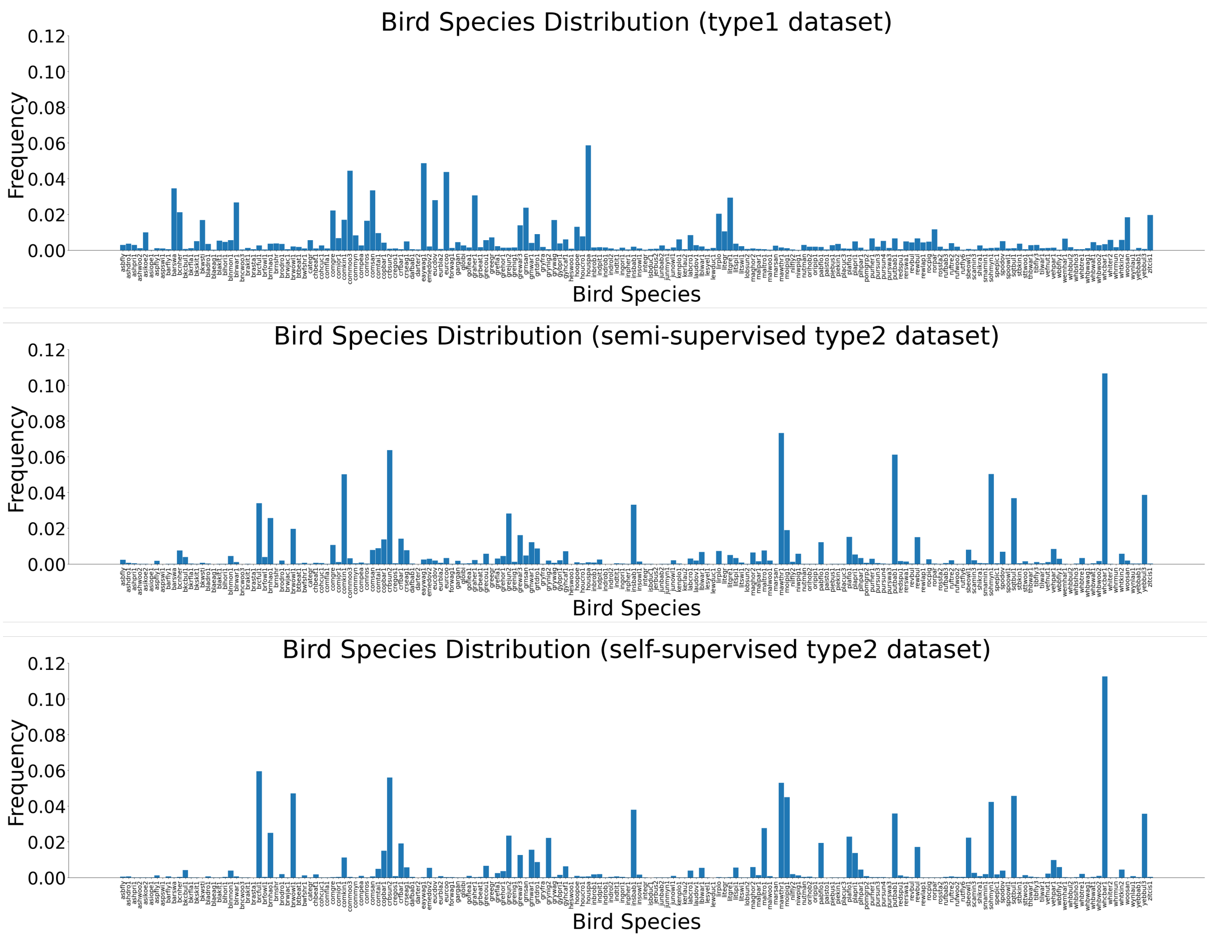

Figure 1: Comparation of species distribution between weakly labeled dataset, semi-supervised unlabeled dataset and self-supervised unlabeled dataset.

3.1.3. Self-supervised Labeled Soundscape

After training models with weakly labeled dataset and semi-supervised unlabeled dataset, we utilize trained models to further extract audio clip with high probability of birdcall presence. We call audio clips extracted from soundscapes with trained models self-supervised unlabeled dataset.

67260 audio clips are extracted from 6654 soundscapes. As we can see in Figure 1, self-supervised unlabeled dataset share a similar species distribution with semi-supervised unlabeled dataset.

3.2. Training Details

3.2.1. Model Architecture

Table 1
Mel-spectrogram parameters for each model

Model type	CNN encoder	Mel bins	Frequency	Window	Hop length
SED	EfficientNetV2-s	128	(0 Hz, 16000 Hz)	2048	417
SED	SeResnext26t-32x4d	128	(0 Hz, 16000 Hz)	2048	627
Custom CNN	ResNet34d	128	(0 Hz, 16000 Hz)	2048	627

We use two types of model architecture from our work in BirdCLEF 2023[11]. One is Sound Event Detection model[12], which we call SED model. Another is CNNs with simple pooling layer, which we call Custom CNN[13][14]. Details of Mel-spectrogram parameters for each model are shown in Table 1.

3.2.2. Knowledge Distillation and Temperature

Knowledge distillation is a technique used in deep learning where a smaller, simpler model, or the student model, is trained to mimic the behavior of a larger, more complex model, or the teacher model [15]. The goal is to transfer the knowledge from the teacher, which may be impractical to use in real-world applications that require fast predictions due to its complexity, to the student. The key idea behind knowledge distillation is to use the output probabilities of the teacher model, known as soft targets, to train the student model. These soft targets provide more information than just the correct class labels (hard targets). This additional information helps the student model learn more effectively.

To transfer the knowledge from off-the-shelf models, we use prediction logit vector extracted by BirdNET and Bird Vocalization Classifier as soft target for model training. Using soft target is also an effective way to address the challenge of weak labels of weakly labeled dataset. In weakly labeled dataset, we have no information about where the birdcall appears and there is a chance that the audio clip does not contain birdcall when we clip the audio. In that case, the presence probability of hard target is still set to 1 for the species, which introduce noise to the training process. On the contrary, presence probability of soft target generated by the teacher model is expected to be a value near 0, which suppress the noise in training process.

In the context of knowledge distillation, the concept of temperature comes into play when generating soft targets. Temperature is a parameter that smooths out the probability distribution produced by the teacher model. When the temperature is high, the differences between the probabilities of the different classes are smaller, making the distribution softer and more informative. When the temperature is low, the distribution becomes sharper, with one class having a much higher probability than the others. By using a higher temperature, the student model can learn more nuanced information from the teacher's predictions.

For our experiments, we found that using a temperature value of 20 provided a good balance, making the soft targets informative enough to significantly improve the student model's performance.

Models are trained with the following loss function:

$$\text{loss function} = 0.1 \cdot \text{hard target loss} + 0.9 \cdot \text{soft target loss} \tag{1}$$

$$\text{hard target loss} = \text{BCELoss}(\text{model prediction}, \text{hard target}) \tag{2}$$

$$\text{soft target loss} = \text{KLDivLoss}\left(\frac{\text{model prediction}}{\text{T}}, \frac{\text{soft target}}{\text{T}}\right) \cdot \text{T}^2 \tag{3}$$

$$\text{T} = 20 \tag{4}$$

3.2.3. Sampling Strategy

To address to the challenge of domain shift, we compare sampling strategies in Table 2 to find the best sample strategy for the training.

Table 2
sampling strategies for domain shift

sampling strategy	description
uniform distribution sampling[14]	Sample the audios of each species according to a uniform distribution. The purpose of this sampling strategy is to address to imbalanced distribution of the species and to prevent model overfitting to certain species.
uniform distribution sampling with geometric weight	In addition to uniform distribution sampling, the sampling probability of audio from unlabeled dataset and subset of weakly labeled dataset recorded in Western Ghats is set to ten times of the probability of other audios. The purpose of this sampling strategy is to address to covariate shift.

4. Results

Macro-average ROC-AUC is calculated as the metrics in BirdCLEF 2024 challenge's Leaderboard, denoted as LB which consists of two variants of public and private. Table 3 presents the experimental results ofmodel trained with knowledge distillation method and unlabeled dataset. In our experiment, adding both unlabeled dataset and knowledge distillation to training significantly improves both Public LB and Private LB of single model. In addition, utilizing self-supervised unlabeled dataset extracted with model trained with semi-supervised unlabeled dataset further improves both Public LB and Private LB. Models with different model type and different cnn encoder share similar LB score.

Table 3
model performance with knowledge distillation training and unlabeled dataset

No	Model	Distillation	Dataset	Public LB	Private LB
1	SED (EfficientNetV2-s)	False	weakly labeled	0.651293	0.637508
2	SED (EfficientNetV2-s)	False	weakly labeled + unlabeled(semi-supervised)	0.680923	0.653730
3	SED (EfficientNetV2-s)	True	weakly labeled + unlabeled(semi-supervised)	0.696951	0.667098
3	SED (EfficientNetV2-s)	True	weakly labeled + unlabeled(self-supervised)	0.701729	0.675409
4	SED (SeResnext26t-32x4d)	True	weakly labeled + unlabeled(self-supervised)	0.694686	0.681450
5	Custom CNN (ResNet34d)	True	weakly labeled + unlabeled(self-supervised)	0.694892	0.671009

From Table 3, we can see that adding type 2 dataset significantly improves the model performance, which means that model trained with unlabeled dataset is more adaptive to fully-annotated dataset, implying that unlabeled dataset share similar distribution with fully-annotated dataset. Applying knowledge distillation also improves the model performance, implying that soft target is an effective way to decrease the label noise in train audio. Further training the model with self-supervised unlabeled dataset improve the model performance. Self-supervised unlabeled dataset contains more birdcall sample than semi-supervised unlabeled dataset, enabling further domain adaption for the model.

Table 4
model performance with different sampling strategies

No	Model	Sample Strategy	Public LB	Private LB
1	SED (EfficientNetV2-s)	uniform distribution sampling	0.696951	0.653730
2	SED (EfficientNetV2-s)	uniform distribution sampling with geometric weight	0.679231	0.658835

The experiment for different sampling strategies is shown in Table 4. Compared to uniform distribution sampling, uniform distribution sampling with geometric weight decreases public LB but slightly improves private LB.

5. Conclusion and future work

In this study, we have presented a novel approach to address the challenge of domain shift in birdcall recognition by leveraging semi-supervised and self-supervised soundscape labeling. Our method utilizes existing off-the-shelf models, BirdNET and Bird Vocalization Classifier, to extract audio clips with high probability of birdcall presence from the unlabeled unlabeled dataset. These semi-supervised labels are then used to train our models, which are subsequently used to extract more audio clips in a self-supervised manner.

Our experimental results demonstrate that this approach significantly improves the performance of our models, indicating that the unlabeled dataset shares a similar distribution with the fully-annotated dataset. Furthermore, we find that applying knowledge distillation further enhances the performance, suggesting that soft target is an effective way to decrease the label noise in training audio. Our solution achieve a remarkable 7th rank among 974 teams at the BirdCLEF 2024 challenge hosted on Kaggle, demonstrating its effectiveness. However, the study also revealed some areas for potential improvements. We find that while adding unlabeled dataset significantly improved the model performance, the model performance varied slightly with different sampling strategies.

In future work, we plan to conduct further experiments to refine our approach. Specifically, we plan to further explore and refine our sampling strategies to improve the model's adaptability to the domain shift in birdcall recognition. Furthermore, we aim to train our models with the semi-supervised unlabeled dataset and then extract and train multiple times with the self-supervised unlabeled dataset. This iterative process is expected to progressively improve the performance of our models by continuously adapting them to the domain of the fully-annotated dataset.

Through these efforts, we aim to further advance the field of birdcall recognition and contribute to the development of more efficient, scalable, and cost-effective methods for monitoring bird populations.

References

[1] 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.

[2] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. CP, S. Sawant, V. V. Robin, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the western ghats, Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024).

[3] Xeno-canto: Sharing bird sounds from around the world, 2022. URL: https://xeno-canto.org.

[4] M. V. Conde, U. Choi, Few-shot long-tailed bird audio recognition, in: Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, 2022.

[5] 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 (2020).

[6] S. Kahl, T. Denton, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of birdclef 2021: Bird call identification in soundscape recordings, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, 2021.

[7] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[8] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W. Vellinga, R. Planqué, A. Joly, Overview of birdclef 2023: Automated bird species identification in eastern africa, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

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

[10] Google, bird-vocalization-classifier, Kaggle, 2023. URL: https://www.kaggle.com/models/google/bird-vocalization-classifier.

[11] L. Hong, Acoustic bird species recognition at birdclef 2023: Training strategies for convolutional neural network and inference acceleration using openvino, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[12] S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, IEEE Journal of Selected Topics

in Signal Processing 13 (2018) 34–48. URL: https://ieeexplore.ieee.org/abstract/document/8567942. doi:10.1109/JSTSP.2018.2885636.
[13] C. Henkel, P. Pfeiffer, P. Singer, Recognizing bird species in diverse soundscapes under weak supervision, 2021. URL: https://arxiv.org/abs/2107.07728. doi:10.48550/ARXIV.2107.07728.
[14] E. Martynov, Y. Uematsu, Dealing with class imbalance in bird sound classification, in: Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, 2022.
[15] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (2015).