CROSS-ADAPT: CROSS-DATASET GENERATION AND DOMAIN ADAPTATION TECHNIQUE FOR FEW-SHOT LEARNING

Technical Report

Amir Latifi Bidarouni, Jakob Abeßer

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

ABSTRACT

Bioacoustic monitoring is an invaluable tool for understanding wildlife well-being. However, the scarcity of annotated data for effective model training coupled with domain shifts resulting from data recorded at various sensor locations with diverse acoustic environments poses significant challenges for deep learning-based audio classification systems. In this paper, we propose a novel cross-dataset data augmentation technique designed to effectively use the limited annotated data available, exemplified by the few-shot learning task 5 of the DCASE challenge. Furthermore, we employ Instance-wise Feature Projection-based Domain Adaptation (IFPDA) to mitigate the domain shifts caused by variations in recording locations or devices. We use a modified ResNet model architecture for a multitask learning setting, which combines multi-class species classification on a patch level and binary classification for frame-level sound event detection.

Index Terms**—** Domain adaptation, data augmentation, few-shot learning, IFPDA

1. INTRODUCTION

Bioacoustic monitoring plays a crucial role in assessing wildlife well-being. Despite its importance, the limited availability of annotated data for training models and the domain shifts imposed by recordings from different sensor locations with varying acoustic conditions, present substantial challenges for deep learning-based audio classification systems. Few-shot learning (FSL) offers an effective tool for bioacoustic researchers, presenting a promising approach to address the scarcity of annotated data. In FSL, the model must learn to classify N classes, each with K annotated samples, and then make predictions on the remaining data (N-way-K-shot classification).

Task 5 of the DCASE Challenge includes training and validation sets for development, alongside an evaluation set. The objective is to develop a FSL system for bioacoustic sound event detection (SED). This system is supposed to be trained using the provided training set and the first five annotated sound events (support set, $K = 5$) of a class in a given audio recording from the evaluation set. The developed model should then predict all sound events of the same class in the remainder of that specific file (query set).

As shown in Table 1, the training set, comprises five datasets covering different animal species with a total of 47 classes. Each recording is accompanied by an annotation file that indicates whether a species is vocalizing (positive/Pos.) or not (negative/Neg.) within a specified segment defined, using its start and end times. In addition, if the species cannot be identified or classified, the annotation is marked as unknown. Fig. 1 illustrates an example of a recording from the `ME` dataset with Pos. segments (green horizontal bars in the bottom row) and Neg. segments (red bars).

Table 1: Overview of the development set of the DCASE Challenge 2024 task 5 [1].

	# Files	Duration (h)	# Classes	Sample rate (Hz)
Training Sets				
BV	5	10	11	24000
HT	5	5	3	6000
JD	1	0.16	1	22050
MT	2	1.16	4	8000
WMW	161	4.66	26	-
Overall	174	21	47	-
Validation Sets				
HB	5	2.63	1	44100
PB	6	3	2	44100
ME	2	0.33	2	48000
RD	6	18	1	48000
PB24	4	2	2	44100
PW	15	24	1	96000
Overall	43	49.95	7	-

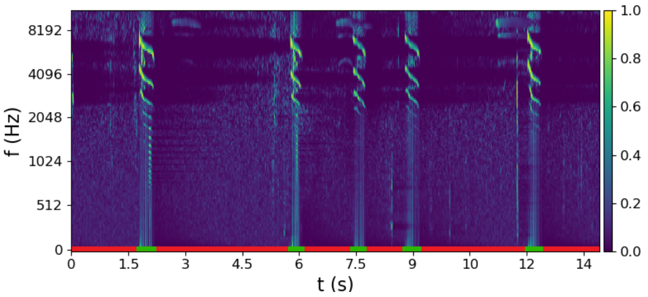


Figure 1: Example of a recording from `ME` dataset. positive and negative segments are shown with green and red area at the bottom row respectively.

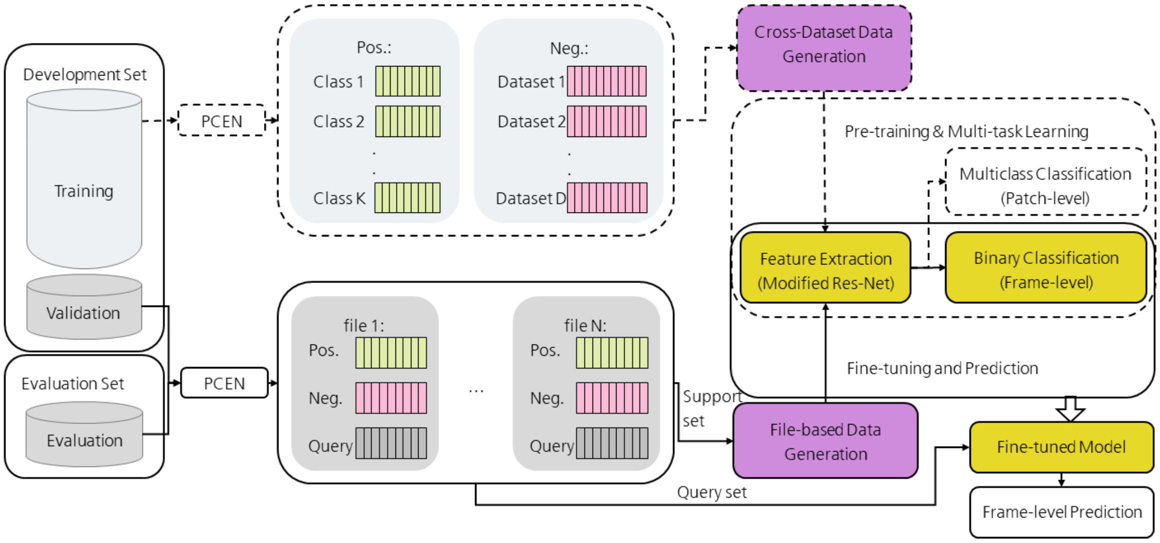

Development Set
Training
Validation
PCEN
Pos.:
Class 1
Class 2
Class K
Neg.:
Dataset 1
Dataset 2
Dataset D
Cross-Dataset Data Generation
Pre-training & Multi-task Learning
Multiclass Classification (Patch-level)
Feature Extraction (Modified Res-Net)
Binary Classification (Frame-level)
Fine-tuning and Prediction
Evaluation Set
Evaluation
PCEN
file 1:
Pos.
Neg.
Query
...
file N:
Pos.
Neg.
Query
Support set
File-based Data Generation
Fine-tuned Model
Query set
Frame-level Prediction

Figure 2: The flowchart illustrates the proposed workflow for FSL, which encompasses data preprocessing (upper and lower blocks, middle column), data generation (purple boxes), and model pre-training and fine-tuning. Dash-lined lines denote the pre-training process, while solid arrows indicate the fine-tuning phase.

From Table 1, it becomes evident that this task impose several significant challenges. First, the sampling rates vary widely, ranging from 6 kHz in the training set to 96 kHz in the validation set. This disparity poses risks such as the introduction of artifacts and reduced frequency resolution during upsampling, or potential loss of information during downsampling. Secondly, the highest sampling rate in the training datasets is only slightly more than half of the lowest sampling rate in the validation set, which could complicate transfer learning between the training and validation sets. Furthermore, the varying durations of each dataset result in highly unbalanced sound classes. Finally, since all datasets have been recorded at different locations with different acoustic recording equipment, we expect that, domain shift imposes a significant challenge in developing a robust FSL system [2].

In this paper, we present the CROSS-ADAPT technique to simultaneously address the challenges posed by imbalanced datasets and domain shifts arising from discrepancies in recording locations. This approach integrates a cross-dataset data generation strategy with a domain adaptation component, namely Instance Feature Projection Domain Adaptation (IFPDA), seamlessly incorporated into a modified ResNet model for feature extraction.

2. PROPOSED METHOD

The proposed workflow for SFL in this work consists of pre-training and fine-tuning phases each consists of two main parts: Data-Generation and training the model. The flowchart of the entire pipeline is illustrated in Fig. 2. The upper part of the figure, depicted with dash-lined lines related to pretraining the multi-task learning model and bottom part marked with solid lines representing the fine-tuning and prediction in frame-level.

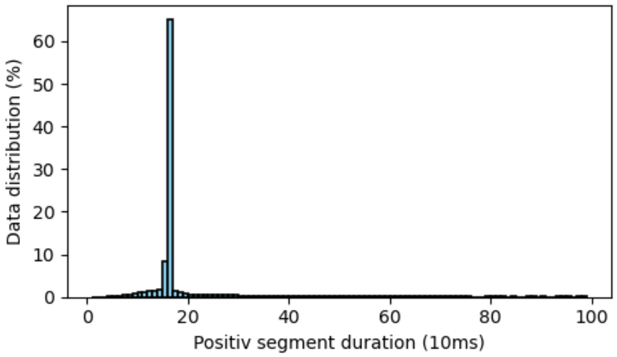


Figure 3: Probability distribution of positive segment with lengths shorter than 1 s, and frame duration of 10 ms, in the training dataset.

2.1. Feature Extraction Parameters

In this paper, we resample all audio recordings at 48 kHz sample rate and compute Mel spectrograms with 128 Mel bands combined with Per-Channel Energy Normalization (PCEN) [3]. We use an FFT size of 2048, a window size of 1024, and a hopsize of 480 samples resulting to 10 ms time resolution (frame size). The spectrograms are normalized to a 0-1 range per audio recording.

Given the various duration of the analyzed audio files, the proposed FSL model for SED, processes spectrogram segments of uniform length (patches). The choice of a suitable patch size is critical for the potential of the system to generalize well to new animal vocalizations of unknown duration. As illustrated in Fig. 3, we analyzed the entire set of annotation files from the training set and found that most positive segments are shorter than 200 milliseconds with only a small fraction of longer segments up to 1 s. Based on this observation, we decided to use 1 s long patches.

2.2. Cross-Dataset Data Generation

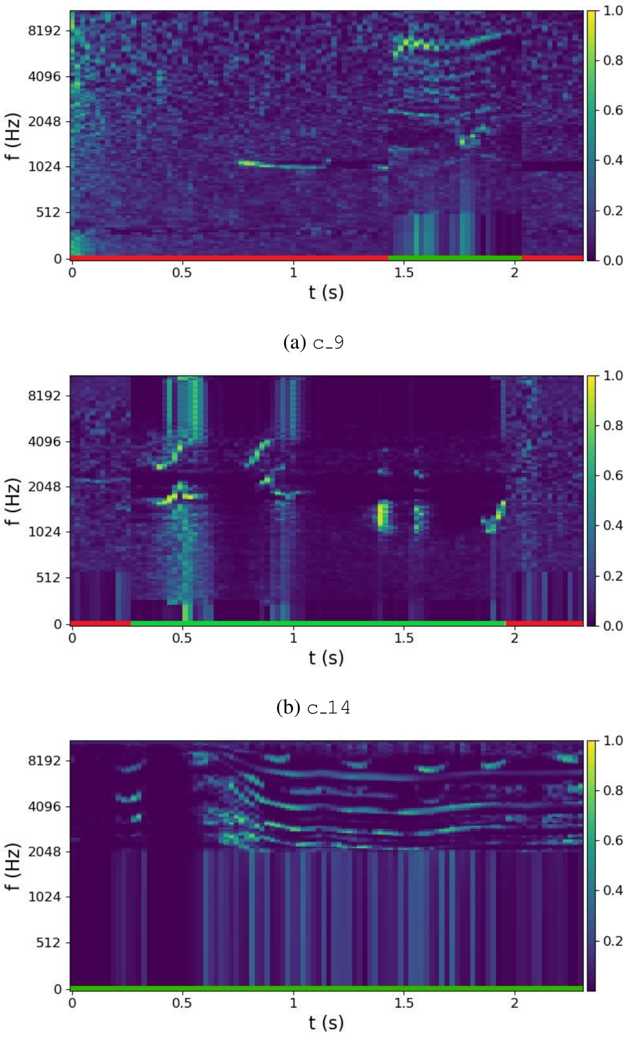
Processing the datasets provided as part of this DCASE task comes with some major challenges: On the one hand, not all classes of a dataset are necessarily present in each file, and each species displays diverse behaviors regarding how frequently they call and for how long they call leads to intra-dataset imbalance. On the other hand, the total duration of the datasets varies significantly, ranging from 10 minutes for the JD dataset to 10 hours for the BV dataset resulting to inter-dataset imbalance.

To overcome these challenges and efficiently utilize the annotated data, we introduced a novel cross-dataset data generation technique (CROSS-ADAPT), which includes the following steps: (1) For each file of a given dataset, we first collect all positive samples for each class. (2) We selected 28 distinct classes from the training set after excluding classes with only a few positive samples or a very short overall duration. (3) Subsequently, we created one pool of negative segments for each dataset. Each pool includes not only segments that are labeled as negative for all classes but also segments without any label.

As an underlying principle of the proposed data generation process, we create spectral patches by randomly combining positive segments (animal calls) and negative segments (background noises) across datasets. We extract negative segments using an overlap of up to 80 % from smaller datasets and without overlap from larger datasets to ensure sufficient variability during patch creation. All negative segments are shuffled and evenly distributed into 28 parts corresponding to the 28 classes to ensure diverse background noises across all animal call classes. Finally, patches are created by combining positive segments from each of the five training sets with 20 % of negative segments coming from the same dataset and 80 % from other datasets.

We allocate 70 % of all positive segments for model training and divide the remaining 30 % equally for model validation and testing. If positive segments exceed the patch size, we slice them into smaller segments with varying lengths and strides. In cases of insufficient positive segments, we replicate them as needed, provided their length is less than 60 % of the patch size. This repetition of positive segments, coupled with the uniqueness of the negative segments, mitigates the risk of overfitting and aids the model in learning the characteristics of the positive segments, even when they are divided into smaller sections. In general, we place the positive segments in a random position within each patch. Some Example of generated patches is shown in Fig. 4

We use the same data generation procedure as explained above for the validation and evaluation datasets with the key difference being that each file is processed individually and negative segments are not shared across files. Specifically, to utilize approximately 70 % of the total duration of the first five shots, 3 to 4 shots are allocated for few-shot learning. The remaining 1 to 2 shots are used to generate validation and test data. Additionally, negative segments preceding the 5th shot are used for patch creation. If the negative segment is too short for patch creation, other augmentation techniques such as horizontal flipping, time stretching, or compression are applied to extend its duration.



(a) c_9

(b) c_14

(c) Q

Figure 4: Example of generated patches for the classes c_9 (a) and c_14 (b) from WMW dataset and (c) class GIG from HT dataset

2.3. Domain Adaptation

The IFPDA technique [4] implements a separate frequency-wise normalization of each patches. The covariance matrix of each patch captures the interdependencies among its frequency bands. The eigenvectors corresponding to the L largest eigenvalues of this matrix are utilized as a transformation matrix. By multiplying the normalized patch with this transformation matrix, its magnitude is projected along the direction that exhibits the greatest variability within the covariance matrix. We integrate this technique into our processing pipeline to mitigate domain shift as it has demonstrated strong performance for domain adaptation in various audio domains [4], [5].

2.4. Model Pre-Training using Multitask Learning

As shown in Fig. 2, our model has a dual output branch architecture. It consists of a residual network (ResNet) as the front-end and two separate classification heads. Conceptually, the first head uses multiclass classification to classify the corresponding animal class for a given patch, while the second head implements a binary classification on a frame-level to locate positive segments. As detailed in

Table 2: Summary of multitask classification network architectures (37,234,841 parameters)

Network Block	**Layers/Parameters**
Feature Extraction Model (Modified ResNet)	
Norm. Layer	IFPDA
Conv. Block	$k = 64 * (3 * 3), s = (2 * 1)$ BatchNorm2D ReLU MaxPooling2D: $k = (3*3), s = (2*1)$
Res. Layer 1	3*Conv. Blocks $k_{out} = 64 * (3 * 3)$
Res. Layer 2	4 Conv. Blocks $k_{out} = 64 * (3 * 3)$
Res. Layer 3	8 Conv. Blocks $k_{out} = 128 * (3 * 3)$
Res. Layer 4	3 Conv. Blocks $k_{out} = 256 * (3 * 3)$
Multiclass classification branch (patch-level)	
Pooling	AdaptiveAveragePooling2d((1,1) Flatten
Output	Dense (number of classes)
Binary classification branch (frame-level)	
Conv. Block 1	$K_{out} = 128 * (3 * 3)$ ReLU BatchNorm2D(128)
Conv. Block 2	$K_{out} = 32 * (3 * 3)$ ReLU BatchNorm2D(32) Flatten
TimeDistributed	Dense (number of classes)

Table 2, the front-end includes a normalization layer using IFPDA and a sequence of four residual layers with different numbers of convolutional blocks each. The multiclass classification head combines average pooling and flattening to predict the overall class of species. The binary classification head includes two additional convolutional blocks followed by a time-distributed dense layer to obtain frame-level predictions.

2.5. Model Finetuning and Prediction

After the model is pre-trained on the generated dataset from training set, the multiclass classification head is discarded, and the ResNet front-end layers are frozen. Given any new audio recording from validation or evaluation dataset, the binary classification head is fine-tuned using the support set of that file and the fine-tuned model is then used to make predictions on the query set.

Our method demonstrates an overall performance of 47.05 % on the validation set, achieving the highest F-score of 56.48 % on the ME dataset and the lowest F-score of 40.85 % on the PW dataset.

Acknowledgment

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

3. REFERENCES

[1] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos, and D. Stowell, "Mind the Domain Gap: a Systematic Analysis on Bioacoustic Sound Event Detection," 2024. [Online]. Available: https://arxiv.org/abs/2403.18638

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

[3] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-Channel Energy Normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[4] 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.

[5] ——, "Towards domain shift in location-mismatch scenarios for bird activity detection," in *32nd European Signal Processing Conference, EUSIPCO 2024*, Lyon, France, 2024, pp. 1–5.