Universal Incremental Learning for Few-Shot Bird Sound Classification

Manjunath Mulimani, Annamaria Mesaros

Signal Processing Research Centre, *Tampere University*, Tampere, Finland
{manjunath.mulimani, annamaria.mesaros}@tuni.fi

***Abstract*—Incremental learning aims to continually learn new input tasks while overcoming the forgetting of previously learned ones. Existing incremental learning methods for audio classification tasks assume that the incoming task either contains new classes from the same domain or the same classes from a new domain, referred to as class-incremental learning (CIL) and domain-incremental learning (DIL), respectively. In this work, we propose a universal incremental learning (UIL) method for few-shot bird sound classification, in which the incoming task contains new or a combination of new and previously seen bird classes from a new domain. Our method uses generalizable audio embeddings from a pre-trained model, which is trained on focal recordings, to develop an incremental learner that solves few-shot bird sound classification tasks from diverse soundscape datasets. These datasets are selected from BIRB (Benchmark for Information Retrieval in Bioacoustics), a large-scale bird sounds benchmark, and used to demonstrate the performance of the proposed method. Results show that our method adapts to the incoming tasks effectively with minimal forgetting of previously seen tasks.**

***Index Terms*—Incremental learning, class-incremental learning, domain-incremental learning, audio embeddings, few-shot bird sound classification**

1. INTRODUCTION

Incremental or continual learning for audio classification aims to acquire new knowledge from incoming audio data over time without significantly forgetting the previously acquired knowledge. In this work, our focus is on the incremental learning of bird sounds. Birds live in most environments, and their diverse species act as indicators of ecosystem health. Specifically, birds are sensitive to the environment, and investigation of bird sounds is highly useful for understanding the shifts in ecosystems and climate [1]. Deep learning models for bird sound classification have reached performance levels that allow their use as biodiversity monitoring systems [1]. In most cases, the annotated data for training these deep models is obtained from citizen science initiatives like Xeno-Canto (XC) [2] that includes over one million vocalizations from more than 10,000 bird species.

On the other hand, bird researchers usually collect their data from a specific soundscape using a noninvasive passive acoustic monitoring (PAM) method. The collected data is used locally to monitor species and perform different studies, such as determining the species' behavioral changes. However, XC and PAM follow different data collection procedures. XC recordings are typically focal, which concentrate on the bird's vocalization of interest. In contrast, PAM recordings are captured in natural soundscapes, which include a mixture of sounds of different species overlapped with background environmental noise. These differences in recording conditions create a domain shift between focal and soundscape recordings [3]. Recently, contrastive learning-based pre-trained models have been developed using focal recordings [4]–[6], and the robust audio embeddings from these pre-trained models were shown to be capable of generalizing to soundscape recordings [3].

In this work, we propose a method for few-shot incremental learning for bird sounds classification using soundscape datasets. From a learning perspective, the method needs to cope with domain shifts caused by a mismatch between acoustic conditions, background noises, and diverse recording devices. This domain shift can cause catastrophic forgetting of previously learned classes when the model learns to classify bird sounds from a new soundscape. Further, a soundscape dataset includes unique species, typical for the specific location, but also common species that are present in other datasets.

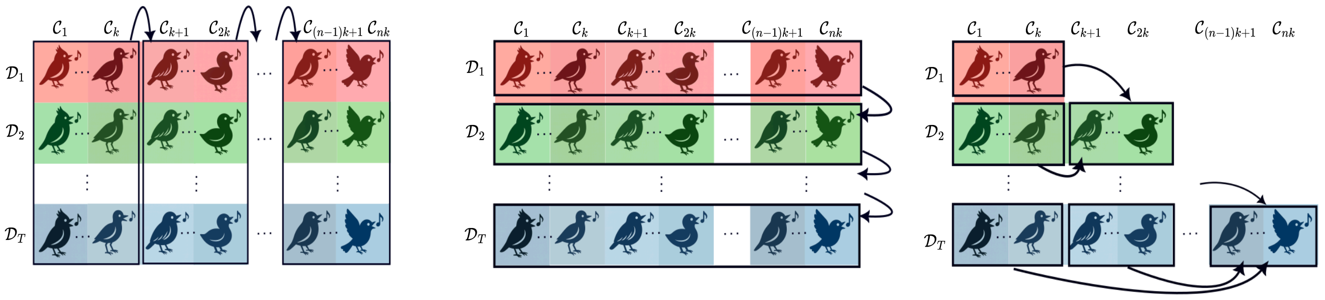
The existing incremental learning methods for audio classification typically fall into two scenarios. One is class-incremental learning (CIL), where incoming tasks contain new audio classes from the same domain [7] [8], as illustrated in Fig. 1a. Another is domain-incremental learning (DIL), where incoming tasks contain the same audio classes from the new domains [9], illustrated in Fig. 1b. These scenarios are based on a strong assumption that incoming sequential tasks either contain the same classes or domains. This assumption is not valid in the case of datasets containing bird sounds. An incoming soundscape dataset, in other words, a new domain, may include a combination of both new and already seen classes from previous domains. Specifically, a model learns a new domain in each incremental time step, and that domain includes new or previously seen classes, as illustrated in Fig. 1. For this, we propose a universal incremental learning method (UIL) for few-shot bird sound classification.

The existing few-shot incremental learning methods for audio classification [10]–[12] or bird sound detection [13] are based on CIL. Most of these methods include one base task or session, followed by multiple incremental tasks, with sufficient training samples available in the base task to train the model from scratch offline for several iterations, and only a few training samples available in incremental tasks to update the model to adapt to new classes.

In this work, we propose using a fixed pre-trained model trained on focal recordings as a generic feature extractor and use its embeddings to adapt to any number of incremental soundscape datasets. We use a cosine classifier that works based on the class prototypes to learn domains incrementally and accumulate the inexpensive class prototypes of previously learned domains to avoid forgetting. Additionally, we inject a random projection (RP) layer with an activation function between the audio embeddings and the classifier as in [14], which expands the dimension of the audio embeddings and enhances the linear separability before computing the class prototypes for the cosine classifier. The proposed method is free from a base session and is online. It adapts to a new domain by only a single forward pass through the training samples, with minimal forgetting of previously learned domains. The proposed method is domain-agnostic and does not require any domain ID during inference for prediction.

The contributions of this work are as follows. (1) We propose a universal incremental learning framework that is suitable for few-shot bird sound classification from diverse datasets; (2) We show that the proposed method incrementally adapts to any new soundscape datasets by passing through a few training samples only once; the

We acknowledge CSC Finland for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through Jane and Aatos Erkko Foundation, grant number 230048, ”Continual learning of sounds with deep neural networks”.



(a) Class-incremental learning (CIL) (b) Domain-incremental learning (DIL) (c) Universal incremental learning (UIL)

Fig. 1: Comparison of incremental learning protocols. Color shade denotes the domain groups, solid box denotes the class groups, and arrows denote incremental steps. (a) Class-incremental learning (CIL) learns new classes from the same domain; (b) domain-incremental learning (DIL) learns the same classes from new domains; (c) Proposed universal incremental learning (UIL) learns new and already seen classes from new domains in incremental steps.

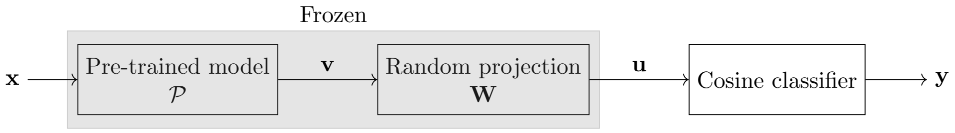


Fig. 2: Overview of the proposed universal incremental learning (UIL) method. Features $\mathbf{v}$ extracted from a frozen pre-trained model $\mathcal{P}$ for given input samples $\mathbf{x}$ are projected into a higher-dimensional space using frozen random weights $\mathbf{W}$, followed by a nonlinear activation function. Randomly projected features $\mathbf{u}$ are then used as inputs to the cosine classifier for prediction. The class prototypes of the cosine classifier are updated based on the new or old classes in the incoming new domain.

proposed method only requires audio embeddings from a fixed pre-trained model without any base session; (3) We analyze performance concerning each component of the proposed method in one-shot and five-shot incremental tasks.

The rest of the paper is organized as follows: Section 2 presents the incremental tasks setups, notations and the proposed method for UIL. Section 3 introduces the datasets, training setup, baselines, implementation details, evaluation metrics, and results. Finally, conclusions are given in Section 4.

2. UNIVERSAL INCREMENTAL LEARNING

2.1. Tasks setup and notations

In our universal incremental learning framework, a sequence of T domains or tasks, here presented as different datasets, $\mathcal{D}_1, \cdots, \mathcal{D}_T$ is introduced to the model for few-shot classification of bird sounds. A domain $\mathcal{D}_t$ includes the audio samples of different bird species, and the model does not have access to previous domains at any learning stage. The proposed UIL combines the functionalities of both CIL and DIL protocols to learn the classes present in the incoming domain. CIL learns new classes from the same domain. In contrast, UIL learns the new classes from the new incoming domain, and it works like DIL whenever the same classes are present in the new domain. We refer to $\mathcal{D}_t$ as the soundscape dataset, domain, task, or stage interchangeably throughout this paper.

We denote the total number of training samples in each soundscape dataset as M_t, and the total number of classes learned so far as C. $\mathbf{x}_{t,m}$ and $\mathbf{y}_{t,m}$ are the m-th training sample and corresponding one-hot encoded label of length C. $\mathbf{v}_{t,m} \in \mathbb{R}^K$ and $\mathbf{v}_{test} \in \mathbb{R}^K$ are embeddings of dimension K, extracted from a frozen pre-trained model $\mathcal{P}$ for input training sample $\mathbf{x}_{t,m}$ and test sample $\mathbf{x}_{test}$, respectively. An overview of the proposed UIL method is given in Fig. 2. It includes two major components: a random projection layer and a cosine classifier, which are explained as follows.

2.2. Random projection layer

RP layer is injected between the pre-trained model and the classifier. The feature embeddings of an input training sample, obtained via the pre-trained model, are projected into a dimension G ($G > K$) using a random weight matrix $\mathbf{W} \in \mathbb{R}^{K \times G}$, followed by an element-wise nonlinear activation function ψ (e.g., ReLU) in each domain as:

$$\mathbf{u}_{t,m} = \psi(\mathbf{v}_{t,m}^\top \mathbf{W}), \quad \mathbf{u}_{test} = \psi(\mathbf{v}_{test}^\top \mathbf{W}), \tag{1}$$

where $\mathbf{u}_{t,m}$ and $\mathbf{u}_{test}$ are new features of length G in training and inference phases. $\mathbf{W}$ contains training-free random weights, which are generated once and kept frozen in all incremental learning stages. The projected features are used to compute class prototypes for the cosine classifier.

2.3. Cosine classifier

Inspired by [14], [15], we propose using a cosine classifier. The weights of the cosine classifier are the class prototypes, computed by averaging the randomly projected embeddings of the training samples within classes. We denote class prototypes of a class y as $\mathbf{c}_y$.

In the proposed UIL setup, the incoming domain either contains only new classes or a combination of both new and previously seen (old) classes. For new classes in the incoming domain, we expand the classifier to accommodate new classes and compute their class prototypes. For old classes in the incoming domain, we compute class prototypes and add these to the class prototypes of the same classes in previously seen domains. This update to the class prototypes of the old classes helps to acquire incremental domain-specific knowledge. During inference, the class of a test sample y_{test} is predicted by finding the maximum cosine similarity s_y between its randomly projected feature embeddings and the set of class prototypes as:

$$y_{test} = \underset{y' \in \{1,\cdots,C\}}{\text{argmax}} \; s_{y'}, \qquad s_y = \frac{\mathbf{u}_{test}^\top \mathbf{c}_y}{||\mathbf{u}_{test}|| \cdot ||\mathbf{c}_y||}. \tag{2}$$

Table 1: Final average accuracy (AA_T) and forgetting (FR_T) of the methods over all the tasks $T = 6$ for one-shot and five-shot bird sounds classification. Higher accuracy and lower forgetting are better.

Method	RP	ReLU	One-shot AA_T	One-shot FR_T	Five-shot AA_T	Five-shot FR_T
Online						
Linear probe			0.5±0.3	0.6±0.3	0.5±0.2	0.9±0.3
Joint linear probe			0.8±0.4		4.5±0.8	
Offline						
Linear probe			6.1±1.7	13.1±1.4	8.8±0.4	27.3±1.5
Joint linear probe			14.9±1.3		36.7±1.2	
Proposed						
Universal incremental learning	✓	✓	16.5±1.4	1.8±0.2	33.5±0.8	5.6±1.0
Ablations						
Universal incremental learning			16.3±1.6	2.4±1.1	32.1±1.1	6.0±1.0
Universal incremental learning	✓		16.6±1.8	2.3±0.7	32.5±0.9	6.5±1.0

During training, we only update the class prototypes in the cosine classifier by a single forward pass through the few-shot training samples.

3. EVALUATION AND RESULTS

3.1. Datasets, training setup and baselines

We use 6 soundscape datasets from BIRB (benchmark for information retrieval in bioacoustics) [16]; these datasets were further preprocessed and provided in [3], [17]. A summary of these datasets is provided in the Table 2, including the number of audio recordings and the number of classes/species. Each soundscape dataset includes unique bird species and common bird species which are present in other soundscapes.

Table 2: A summary of the soundscape datasets from BIRB.

Dataset	Abbreviation	Recordings	Classes
Peru [18]	PER	14,798	132
Colombia, Costa Rica [19]	NES	6,952	89
Island of Hawai'i, USA [20]	UHH	59,583	27
High Sierra Nevada, USA [21]	HSN	10,296	19
New York State, USA [22]	SSW	50,760	96
Sierra Nevada [23]	SNE	20,147	56

We adopt an episodic training and testing strategy. For training, we randomly select one and five training samples per class from a dataset for one-shot and five-shot bird sound classification tasks, incrementally. The remaining test samples of all the datasets seen so far are considered for evaluation. Samples in the training and testing sets are different in all experiments. In this work, we learn the datasets in the following order: PER → NES → UHH → HSN → SSW → SNE.

We compare the performance of the proposed UIL method with two different methods used to solve the same problem: (1) a linear probe classifier trained on each incremental soundscape dataset using embeddings extracted from the pre-trained model. A linear probe is widely used to demonstrate the generalization ability of a pre-trained model's embeddings in classification tasks [24], [25]. (2) a joint linear probe classifier trained using embeddings of all the soundscape datasets seen so far. This joint training violates the incremental learning setup, but it is given for completeness.

3.2. Implementation details and evaluation metrics

We use the CvT-13 pre-trained transformer model [26], trained on focal recordings of XC by [3]. The 384-dimensional feature vectors are extracted for audio recordings in each soundscape dataset using CvT-13 that was trained using Prototypical Contrastive Learning of Representations (ProtoCLR). For complete details about the training procedure and parameters of CvT-13, we refer the reader to [3]. We compute log-mel spectrograms from audio recordings of soundscape datasets and apply the augmentation techniques using the configuration provided in [3]. ReLU is used as a nonlinear activation function in Eq. (1). The dimension G is set to 2048 based on preliminary experiments performed with different values of G.

For the baseline methods, linear and joint linear probes are trained using cross-entropy loss, Adam optimizer [27] with a learning rate of 5×10^{-4} and the CosineAnnealingLR [27] scheduler. The proposed method updates only the cosine classifier and does not require Adam-based weight updates. The number of epochs is set to 1 for online and 25 for offline training, to train both linear and joint linear probes.

Following the standard practice in incremental learning [9], [14], we evaluate the performance of the method after learning the current domain $\mathcal{D}_t$ using average accuracy and forgetting of all the domains seen so far. Average accuracy is defined as:

$$AA_t = \frac{1}{t}\sum_{j=1}^{t} ACC_{t,j}, \tag{3}$$

where $ACC_{t,j}$ is the accuracy of j-th domain after learning the t-th domain. Average forgetting is defined as:

$$FR_t = \frac{1}{t-1}\sum_{j=1}^{t-1} \max_{\hat{t}\in\{1,2,\cdots,t-1\}} (ACC_{\hat{t},j} - ACC_{t,j}), \tag{4}$$

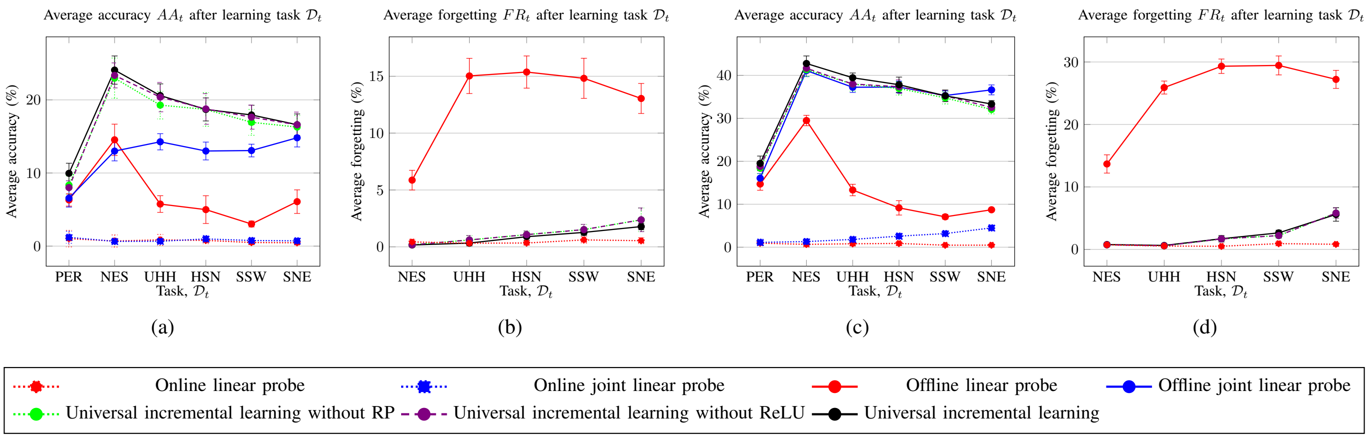
where $ACC_{\hat{t},j}$ is the maximum accuracy of the previously seen j-th domain after learning the $\hat{t}$-th domain.

We compute AA_t and FR_t after learning $\mathcal{D}_t$ in both one-shot and five-shot classification setups. We run each incremental few-shot experiment 10 times with different random seeds and report the results using mean and standard deviation over average accuracy values and forgetting values across the runs.

3.3. Results

We report final average accuracy (AA_T) and forgetting (FR_T) of different methods after learning all the tasks ($T = 6$) in Table 1. For a detailed task-wise analysis, we show average accuracy (AA_t) and forgetting (FR_t) of different methods after learning each soundscape dataset, in Fig. 3.

As expected, the offline linear probe performs better than the online linear probe, which is trained using a single epoch. The linear probe



(a) (b) (c) (d)

Fig. 3: Average accuracy and forgetting of the methods after learning the task $\mathcal{D}_t$. Average accuracy (a) and forgetting (b) of the current $\mathcal{D}_t$ and previously seen tasks for one-shot classification; Average accuracy (c) and forgetting (d) of the current $\mathcal{D}_t$ and previously seen tasks for five-shot classification.

Table 3: Accuracy of the methods on the current task $\mathcal{D}_t$ for one-shot and five-shot bird sounds classification.

Method	PER	NES	UHH	HSN	SSW	SNE
One-shot classification						
ProtoCLR [3]	9.23±1.6	38.6±5.1	18.4±2.3	21.2±7.3	15.5±2.3	25.8±5.2
Proposed universal incremental learning	9.9±1.3	39.4±4.1	14.2±3.2	11.6±3.9	11.7±1.2	21.7±3.3
Five-shot classification						
ProtoCLR [3]	19.2±1.1	67.9±2.8	36.1±4.3	48.0±4.3	34.6±2.3	48.6±2.8
Proposed universal incremental learning	19.6±1.2	68.4±2.8	33.3±3.1	36.9±6.1	30.4±1.6	41.2±1.9

does not have access to the previous soundscape datasets. Training an offline linear probe on the current soundscape dataset for multiple iterations improves the performance on the current soundscape dataset, but overwrites the parameters of the previously learned classes, leading to increased average forgetting, as observed in Fig. 3b and 3d, and reduced average accuracy in Fig. 3a and 3c after learning each dataset. The performance of the online joint linear probe is worse, but the offline joint linear probe achieves competitive results by taking advantage of full access to the training samples of all the datasets seen so far. A joint linear probe requires multiple iterations and more training samples from all the datasets to perform better.

The proposed UIL outperforms the joint linear probe in a one-shot incremental learning setup, as can be seen in Fig. 3a and Table 1. We can see from Fig. 3c that the proposed UIL also shows competitive performance as compared to the joint linear probe in a five-shot incremental learning setup. UIL suffers from minimal forgetting despite potential abrupt changes in the domain distributions and increasing numbers of classes in incremental stages.

Without the random projection or only with the cosine similarity classifier, UIL gives comparable results. However, class prototypes from the randomly projected feature embeddings perform better, and a nonlinear activation function, ReLU, helps extract the important nonlinear interactions in feature embeddings, reducing the average forgetting.

We compare the accuracy of the proposed UIL on the current domain $\mathcal{D}_t$ with the existing (non-incremental) few-shot bird sound classification method in [3]; the numbers for each domain/dataset are presented in Table 3. The work in [3] uses the feature embeddings from the same ProtoCLR-based CvT-13 pre-trained model, but uses a separate classifier to learn each soundscape dataset, with each classifier handling only the classes present in the current dataset. In the incremental learning setup, the number of classes in the cosine classifier keeps increasing as the model learns a new dataset; therefore, a single classifier handles all the classes seen so far. The proposed UIL approach outperforms ProtoCLR [3] in the first datasets (PER and NES in this experimental setup) in both one-shot and few-shot learning. As it learns more datasets in sequence, the number of classes increases and known classes are updated to the new domain, affecting the overall accuracy.

4. CONCLUSION

In this paper, we presented a universal incremental learning method for few-shot bird sound classification. Our proposed approach is specifically designed to learn bird sound classes from multiple datasets incrementally, dealing with the presence of both new, previously unseen classes and classes that were already seen, but are now recorded in different acoustic conditions. It outperforms online and offline linear probes in both one-shot and five-shot incremental learning setups. The proposed approach relies on the audio embeddings of the CvT-13 pre-trained model, that uses focal recordings for training, but needs to update only the class prototypes in the cosine classifier to account for the domain mismatch. Our method is suitable for realistic deployments, being able to adapt to any new domains and new classes in a few-shot scenario. Future research includes the investigation of the effectiveness of the proposed approach on other animal datasets or combinations or different kinds of audio data.

REFERENCES

[1] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, “Birdnet: A deep learning solution for avian diversity monitoring,” *Ecological Informatics*, vol. 61, p. 101236, 2021.

[2] W.-P. Vellinga and R. Planqué, “The xeno-canto collection and its relation to sound recognition and classification.” in *CLEF (Working Notes)*, 2015.

[3] I. Moummad, R. Serizel, E. Benetos, and N. Farrugia, “Domain-invariant representation learning of bird sounds,” *arXiv preprint arXiv:2409.08589*, 2024.

[4] I. Moummad, R. Serizel, and N. Farrugia, “Pretraining representations for bioacoustic few-shot detection using supervised contrastive learning,” in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 126–130.

[5] I. Moummad, N. Farrugia, and R. Serizel, “Self-supervised learning for few-shot bird sound classification,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 600–604.

[6] ——, “Regularized contrastive pre-training for few-shot bioacoustic sound detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1436–1440.

[7] M. Mulimani and A. Mesaros, “A closer look at class-incremental learning for multi-label audio classification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1293–1306, 2025.

[8] ——, “Incremental learning of acoustic scenes and sound events,” in *Proceedings of the 8th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023, pp. 141–145.

[9] ——, “Domain-incremental learning for audio classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[10] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, “Few-shot continual learning for audio classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 321–325.

[11] Y. Li, W. Cao, W. Xie, J. Li, and E. Benetos, “Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes,” *IEEE Transactions on Multimedia*, vol. 26, pp. 1346–1360, 2023.

[12] Y. Li, J. Li, Y. Si, J. Tan, and Q. He, “Few-shot class-incremental audio classification with adaptive mitigation of forgetting and overfitting,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2297–2311, 2024.

[13] X. Wu, D. Xu, H. Wei, and Y. Long, “Few-shot continual learning with weight alignment and positive enhancement for bioacoustic event detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] M. D. McDonnell, D. Gong, A. Parvaneh, E. Abbasnejad, and A. Van den Hengel, “Ranpac: Random projections and pre-trained models for continual learning,” in *NeurIPS*, 2023.

[15] D.-W. Zhou, Z.-W. Cai, H.-J. Ye, D.-C. Zhan, and Z. Liu, “Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need,” *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1012–1032, 2025.

[16] J. Hamer, E. Triantafillou, B. Van Merriënboer, S. Kahl, H. Klinck, T. Denton, and V. Dumoulin, “Birb: A generalization benchmark for information retrieval in bioacoustics,” *arXiv preprint arXiv:2312.07439*, 2023.

[17] I. Moummad, “Soundscape datasets for few-shot bird sound classification,” Oct. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13994373

[18] W. A. Hopping, S. Kahl, and H. Klinck, “A collection of fully-annotated soundscape recordings from the southwestern amazon basin,” Sep. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7079124

[19] Álvaro Vega-Hidalgo, S. Kahl, L. B. Symes, V. Ruiz-Gutiérrez, I. Molina-Mora, F. Cediel, L. Sandoval, and H. Klinck, “A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica,” Jan. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7525349

[20] A. Navine, S. Kahl, A. Tanimoto-Johnson, H. Klinck, and P. Hart, “A collection of fully-annotated soundscape recordings from the island of hawai’i,” Sep. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7078499

[21] M. Clapp, S. Kahl, E. Meyer, M. McKenna, H. Klinck, and G. Patricelli, “A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range,” Jan. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7525805

[22] S. Kahl, C. M. Wood, P. Chaon, M. Z. Peery, and H. Klinck, “A collection of fully-annotated soundscape recordings from the western united states,” Sep. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7050014

[23] S. Kahl, R. Charif, and H. Klinck, “A collection of fully-annotated soundscape recordings from the northeastern united states,” Aug. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7018484

[24] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” *arXiv preprint arXiv:1610.01644*, 2016.

[25] Y. Liang, L. Zhu, X. Wang, and Y. Yang, “A simple episodic linear probe improves visual recognition in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9559–9569.

[26] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “Cvt: Introducing convolutions to vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.

[27] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *International Conference on Learning Representations (ICLR)*, 2017.