Domain-Invariant Representation Learning of Bird Sounds

Ilyass Moummad, Romain Serizel, Emmanouil Benetos, Nicolas Farrugia

Domain-Invariant Representation Learning of Bird Sounds

Ilyass Moummad[1] Romain Serizel[2] Emmanouil Benetos[3] Nicolas Farrugia[1]

[1]IMT Atlantique, CNRS, Lab-STICC, Brest, France
[2]Université de Lorraine, CNRS, Inria, Loria, Nancy, France
[3]C4DM, Queen Mary University of London, London, UK

***Abstract*—Passive acoustic monitoring (PAM) is crucial for bioacoustic research, enabling non-invasive species tracking and biodiversity monitoring. Citizen science platforms like Xeno-Canto provide large annotated datasets from focal recordings, where the target species is intentionally recorded. However, PAM requires monitoring in passive soundscapes, creating a domain shift between focal and passive recordings, which challenges deep learning models trained on focal recordings. To address this, we leverage supervised contrastive learning to improve domain generalization in bird sound classification, enforcing domain invariance across same-class examples from different domains. We also propose ProtoCLR (Prototypical Contrastive Learning of Representations), which reduces the computational complexity of the SupCon loss by comparing examples to class prototypes instead of pairwise comparisons. Additionally, we present a new few-shot classification benchmark based on BirdSet, a large-scale bird sound dataset, and demonstrate the effectiveness of our approach in achieving strong transfer performance.**[1]

***Index Terms*—Supervised Contrastive Learning, Domain Generalization, Few-shot Learning, Bioacoustics.**

I. Introduction

Passive Acoustic Monitoring (PAM) is a non-invasive method for studying wildlife through sound. By using acoustic recorders, researchers can gather data on animal behavior, migration, and population trends without disturbance [1]. PAM is useful for monitoring endangered species, offering long-term insights for conservation.

In recent years, deep learning models have emerged as a powerful tool to process and analyze complex bioacoustic data [2]. A key source of training data for these models comes from citizen science platforms like Xeno-Canto [3], which contains over one million annotated vocalizations from more than 10,000 species, primarily birds. These citizen-led initiatives have significantly expanded the availability of labeled wildlife sound data, enabling the development of more robust and accurate deep learning models [4].

However, a challenge arises from the difference between data collected on platforms like Xeno-Canto and PAM recordings. In citizen science platforms, recordings are typically focal—where the recorder is aimed directly at the species of interest. In contrast, PAM systems passively capture sounds within natural soundscapes, leading to recordings that contain a mix of species vocalizations and background environmental noise. This difference in recording conditions creates a domain shift, complicating the ability of models trained on focal data to generalize to soundscape recordings in PAM. In practice, we require models that can perform well across diverse and potentially unseen environments.

Supervised contrastive learning (SupCon) [5], a supervised learning framework for training robust feature extractors, has demonstrated strong generalization capabilities for transfer learning in bioacoustics, particularly in few-shot classification [6] and detection [7]. However, these studies have been limited to settings where both training and testing rely on focal recordings, and therefore do not address the domain shift challenge associated with testing on PAM recordings when models are trained on focal recordings.

Domain Generalization (DG) [8] aims to develop models that learn robust features that are domain-invariant, i.e. capable of generalizing to new unseen domains without prior knowledge or access to target domain data during training.

SupCon offers a promising approach for DG. In SupCon, the objective is to learn an embedding space where same-class examples are pulled together and different-class examples are pushed apart. This clustering can promote domain-invariance when sufficient domain diversity is present in the dataset, allowing the model to focus on features that are domain-invariant. In contrast, its self-supervised counterpart SimCLR [9] lacks this explicit mechanism for domain-invariance, as it relies solely on augmentations to create positive pairs. Without label information, SimCLR requires carefully designed augmentations that account for domain shift [10].

Despite its effectiveness, SupCon is computationally expensive due to the need for pairwise similarity calculations between all examples. To address this, we introduce ProtoCLR (Prototypical Contrastive Learning of Representations), a more efficient variant. By analyzing SupCon's gradient and drawing inspiration from the generalization capabilities of prototypical networks [11] in few-shot learning, ProtoCLR replaces pairwise comparisons with a prototypical contrastive loss that compares examples to class prototypes, retaining the original objective while significantly reducing computational complexity.

We propose a new few-shot classification benchmark based on the BirdSet [12] dataset to evaluate the generalization ca-

This work was co-funded by Collège doctoral de Bretagne, Ecole doctorale SPIN, GdR IASIS, OSO-AI company, and AI@IMT program.

[1]Models and code: https://github.com/ilyassmoummad/ProtoCLR

pabilities of models trained on Xeno-Canto's focal recordings and tested on diverse soundscape datasets. This benchmark is designed to assess how well models can generalize across domains in challenging few-shot scenarios. We validate our proposed loss ProtoCLR on this benchmark, demonstrating its effectiveness in improving DG in bird sound classification.

In this work, we make the following contributions:

- We establish a large-scale few-shot benchmark for bird sound classification using BirdSet datasets, evaluating model generalization from focal to soundscape recordings.
- We introduce ProtoCLR, a novel supervised contrastive loss that reduces computational complexity of SupCon by using class prototypes instead of pairwise comparisons.

II. Related Work

Nolasco et al. [13] reformulate bioacoustic sound event detection using a few-shot learning approach to recognize species from a few labeled examples, making it suitable for rare species but limited to single-species detection per task. Heggan et al. [14] introduce MetaAudio, a few-shot benchmark for audio classification, including BirdCLEF 2020 [15], which focuses on generalizing to new classes but only includes focal recordings.

To address the generalization challenge from focal to soundscape recordings, the BIRB [16] benchmark focuses on few-shot retrieval, retrieving labeled sounds from large, unlabeled datasets. BirdSet [12] emphasizes transfer learning, evaluating models across various downstream classification tasks.

DG [8] has emerged as a critical approach to tackle domain shift, where the test data distribution differs from the training data. It aims to learn robust, domain-invariant representations using only source domain data, without requiring access to target domain data during training [8]. DG methods typically focus on learning domain-invariant representations, using techniques like domain alignment [17], meta-learning [18], and data augmentation [19]. These approaches help the model learn features that remain consistent across varying domains. In bioacoustics, DG is especially important due to the difficulty in collecting annotated soundscape recordings compared to focal data [12], [16].

Invariant learning has gained attention, where models are trained to learn features that remain invariant across different variations in data, such as augmented versions in self-supervised learning [9], [20]–[22] or same-class examples in supervised learning [5]. This approach has proven effective for learning robust features in bird sound classification [6].

III. Method

A. Supervised Contrastive Loss (SupCon)

Given a batch with two views (transformations) of each example, let I denote the set of indices of all examples in the batch, $P(i)$ represent the set of indices of positive examples for examples i, and $A(i) = I \setminus \{i\}$ represent the set of all other indices excluding i. For an example i, let z_i be its l_2-normalized embedding and τ the temperature parameter.

The SupCon [5] loss is defined as:

$$\mathcal{L}^{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)}, \tag{1}$$

The gradient of SupCon loss for an example i with respect to its embedding z_i is (please refer to [5] for more details):

$$\nabla_{z_i} \ell_i^{\text{SupCon}} = \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{1}{\tau} z_p - \frac{1}{\tau} \frac{\sum_{a \in A(i)} S_{ia} z_a}{\sum_{a \in A(i)} S_{ia}}, \tag{2}$$

where $S_{ia} = \exp(z_i \cdot z_a / \tau)$ is the similarity between z_i and z_a.

This gradient consists of two terms: a positive term $\frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{1}{\tau} z_p$ pulling the embedding z_i towards its class centroid and a negative term $-\frac{1}{\tau} \frac{\sum_{a \in A(i)} S_{ia} z_a}{\sum_{a \in A(i)} S_{ia}}$ pushing it away from other examples.

B. Prototypical Contrastive Loss (ProtoCLR)

Motivated by the gradient of SupCon, we propose ProtoCLR (**Proto**typical **C**ontrastive **L**earning of **R**epresentations), which introduces class-level centroids into contrastive learning. The centroid for each class y in the batch is computed as $c_y = \frac{1}{|C(y)|} \sum_{i \in C(y)} z_i$, where $C(y)$ is the set of indices of examples with label y and $|C(y)|$ is its size. We define the ProtoCLR loss as:

$$\mathcal{L}^{\text{ProtoCLR}} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp\left(z_i \cdot c_{y_i} / \tau\right)}{\sum_{y \in Y} \exp\left(z_i \cdot c_y / \tau\right)}, \tag{3}$$

where c_{y_i} is the centroid of the class to which example i belongs, and Y is the set of all classes in the batch.

Similarly to SupCon, the gradient for ProtoCLR is:

$$\nabla_{z_i} \mathcal{L}_i^{\text{ProtoCLR}} = \frac{1}{\tau} c_{y_i} - \frac{1}{\tau} \frac{\sum_{y \in Y} S_{iy} c_y}{\sum_{y \in Y} S_{iy}}, \tag{4}$$

where $S_{iy} = \exp(z_i \cdot c_y / \tau)$ is the similarity between z_i and c_y.

The positive term remains the same as in the gradient of SupCon, pulling the embeddings $\mathbf{z}_i$ towards the centroids of their respective classes. The difference is the negative term $-\frac{1}{\tau} \frac{\sum_{y \in Y} S_{iy} c_y}{\sum_{y \in Y} S_{iy}}$: in ProtoCLR, the embeddings are pushed away from the weighted average of the centroids as opposed to the individual embeddings in SupCon.

C. ProtoCLR vs SupCon

In order to comprehensively assess the efficacy of ProtoCLR, we conduct a comparative analysis with SupCon.

1) Complexity: SupCon has a computational cost of $\mathcal{O}(N^2)$ due to computing dot products between all pairs of examples in a batch of size N, independent of the number of classes C. In contrast, ProtoCLR reduces this to $\mathcal{O}(N \times C)$ by computing dot products with class prototypes. Since C is usually smaller than N, ProtoCLR is more efficient, particularly for large batches.

2) Variance: SupCon relies on pairwise comparisons within the same class, which can lead to higher variance due to intra-class variability. In contrast, ProtoCLR compares embeddings z_i with class prototypes c_{y_i}, reducing variance as the prototype variance $\text{Var}(c_{y_i})$ decreases by $N_{y_i}^2$, where N_{y_i} is the number of examples in class y_i. This leads to lower noise and more stable gradients in ProtoCLR:

$$\text{Var}(c_{y_i}) = \frac{\text{Var}(\sum_{j \in y_i} z_j)}{N_{y_i}^2}.$$

3) Near Convergence Equivalence: Near convergence, both SupCon and ProtoCLR promote intra-class compactness with embeddings clustering tightly around class centroids: $z_i \approx c_{y_i}$ for all $i \in I$. In SupCon, the negative can be rewritten as:

$$\frac{S_{ia} z_a}{S_{ia}} \approx \frac{\exp(z_i \cdot z_a/\tau) z_a}{\exp(z_i \cdot z_a/\tau)} \approx \frac{\exp(z_i \cdot c_{y_a}/\tau) c_{y_a}}{\exp(z_i \cdot c_{y_a}/\tau)} \approx \frac{S_{iy_a} c_{y_a}}{S_{iy_a}}. \tag{5}$$

Thus, SupCon and ProtoCLR converge to similar strategies, ensuring intra-class compactness and inter-class separability in the final learned representations.

IV. Experiments

A. Few-Shot Classification Benchmark

The BirdSet benchmark [12] comprises two tasks: multi-label classification, where each audio recording is segmented into 5-second intervals to detect the presence of one or more species (or none), and multi-class classification, where individual bird events are detected using peak detection and bambird [23], a tool for identifying bird events in audio recordings, with each event containing a single species. In line with the MetaAudio framework, we focus on the multi-class classification task to define a few-shot evaluation for assessing the generalization capabilities of pre-trained models [24]. BirdSet offers two training sets: XCL (Xeno-Canto Large) with focal recordings across nearly 10,000 species, and XCM (Xeno-Canto Medium), a specialized subset of XCL with recordings from 411 species represented in the test datasets. The validation and test datasets contain soundscape recordings. The benchmark is detailed in Table II.

Following BioCLIP [25], we sample *k*-shot learning tasks by randomly selecting *k* examples for each class and obtain the audio embeddings from the audio encoder of the pre-trained models. We then compute the average feature vector of the *k* embeddings as the training prototype for each class. All the examples left in the dataset are used for testing.

To make predictions, we employ SimpleShot [26] by applying mean subtraction and L2-normalization to both centroids and test feature vectors. We then select the class whose centroid is closest to the test vector as the prediction. We repeat each few-shot experiment 10 times with different random seeds and report the mean and standard deviation accuracy in Table I.

B. Reference Systems

To compare our domain-invariant pre-training approach (SupCon and ProtoCLR), we train reference systems using cross-entropy (CE) loss as a supervised baseline and SimCLR as a self-supervised contrastive baseline. Additionally, we evaluate large-scale, state-of-the-art models in bioacoustics: the encoder of BioLingual [27], an HT-SAT transformer pre-trained on AudioSet and fine-tuned using contrastive language-audio training to align animal sounds with text captions describing the class across a large collection of data including Xeno-Canto, iNaturalist, Animal Sound Archive, ... etc; and the encoder of Perch [16], an EfficientNet-B1 [28] trained on Xeno-Canto for species classification, as well as taxonomic ranks genus, family, and order.

C. Pre-training Details

We train all models with CvT-13 [29], an efficient transformer architecture, on XCM and XCL datasets for 100 epochs using the AdamW optimizer with a batch size of 256, with a weight decay of 1×10^{-6}. Following Moummad et al. [6], we apply the augmentations found to be effective for bird sound representations: circular time shift [30], SpecAugment [31], and spectrogram mixing [32]. These models are trained with a projector of dimension 128. For the CE loss, we only apply circular time shift and SpecAugment as augmentations, excluding Spectrogram Mixing, as it prevented the model from converging. The learning rate for CE and ProtoCLR is set to 5×10^{-4}, while for SupCon and SimCLR, we use a learning rate of 1×10^{-4}. We tune hyperparameters by monitoring *k*-NN accuracy on the POW dataset.

D. Results and Discussion

Table I presents the performance of different models on one-shot and five-shot bird sound classification tasks. ProtoCLR pre-trained on XCM consistently outperforms others in both tasks, with SupCon close behind. Additionally, ProtoCLR is more computationally efficient; for one training epoch with a batch size of 256, SupCon computes 80.4B MACs, while ProtoCLR computes only 28.3B. Notably, ProtoCLR significantly outperforms CE on average (75.0 vs. 46.2). On the other hand, SimCLR performs the worst on XCM, likely due to domain shift. Incorporating unsupervised domain generalization techniques [10] could enhance SimCLR's generalization ability.

Pre-training on the larger XCL dataset negatively impacts the performance of ProtoCLR, SupCon, and CE models. This may be due to XCL's broader class diversity (9,736 classes) compared to XCM (411 classes) that only contains the target classes, which leads to a loss of discriminative capacity on the target classes. XCL also has a higher coefficient of variation (1.43 v 0.43) and Gini coefficient (0.62 vs 0.24), indicating greater class imbalance, which ProtoCLR may be more sensitive to than SupCon.

When pretrained on XCL, SupCon and ProtoCLR outperform CE in one-shot learning but not in five-shot learning, where SupCon surpasses ProtoCLR. Additionally, SupCon

Model	Val	Test							Mean
	POW	**PER**	**NES**	**UHH**	**HSN**	**NBP**	**SSW**	**SNE**	
Random Guessing	2.08	0.75	1.12	3.70	4.76	1.96	1.23	1.78	2.17
One-Shot Classification									
BioLingual	39.6±4.6	33.8±1.5	41.2±4.0	59.2±4.6	50.3±9.8	44.8±3.1	39.9±4.0	41.6±0.4	43.8
Perch	39.4±3.0	41.7±0.8	45.0±3.6	59.3±8.1	46.5±7.3	48.1±2.9	40.5±4.7	40.7±8.6	45.1
XCM Pre-training									
CE	43.6±3.5	41.8±1.4	45.2±2.7	67.2±6.5	58.8±4.9	34.4±2.9	37.5±3.2	41.3±3.6	46.2
SimCLR	15.5±3.2	12.4±0.9	14.9±1.4	27.2±4.7	22.0±3.4	15.1±1.3	12.0±1.5	14.3±1.8	16.6
SupCon	69.3±4.4	75.3±2.1	72.8±4.0	81.2±8.4	74.8±7.1	58.4±3.9	62.5±4.5	62.6±4.1	69.6
ProtoCLR	**72.3**±4.8	**79.0**±2.2	**79.1**±2.9	**83.4**±6.2	**82.2**±6.2	**65.2**±3.5	**69.6**±3.5	**69.6**±5.9	**75.0**
XCL Pre-training									
CE	32.3±2.9	27.6±1.9	34.1±2.5	54.5±7.4	44.8±4.4	31.6±1.8	27.7±2.1	30.4±2.6	35.3
SimCLR	15.0±1.9	12.8±0.8	15.3±1.6	27.6±5.7	19.6±3.6	15.4±1.3	12.7±1.9	15.1±2.3	16.7
SupCon	35.4±5.5	36.1±2.1	37.1±3.5	58.7±4.2	45.1±5.2	60.7±4.1	40.0±3.0	44.4±4.6	44.7
ProtoCLR	30.7±3.2	28.9±1.5	32.3±2.6	51.5±4.5	39.4±4.4	50.6±3.4	33.5±3.5	38.6±4.1	38.2
Five-Shot Classification									
BioLingual	65.8±0.9	58.8±0.5	66.0±0.7	77.5±1.5	72.0±2.9	70.7±0.5	65.4±0.8	64.7±1.7	67.6
Perch	67.3±1.2	68.7±0.7	71.8±0.6	82.0±1.8	78.7±1.6	79.0±0.5	69.8±1.0	72.1±1.0	73.7
XCM Pre-training									
CE	79.8±0.9	80.7±0.6	82.6±0.8	91.2±1.4	88.6±1.0	63.3±1.0	74.9±1.1	78.1±1.4	79.9
SimCLR	28.8±1.4	27.3±0.7	30.0±1.2	47.1±3.5	38.3±1.7	28.3±1.1	25.0±1.2	28.6±2.1	31.7
SupCon	83.8±0.5	87.7±0.2	86.4±0.3	90.2±0.8	88.6±0.7	77.1±0.4	80.8±0.7	80.9±1.0	84.4
ProtoCLR	**87.5**±0.4	**89.9**±0.2	**89.2**±0.3	**91.8**±0.8	**92.2**±0.5	**81.0**±0.6	**85.0**±0.6	**85.8**±0.8	**87.8**
XCL Pre-training									
CE	65.8±1.3	65.3±0.7	71.3±1.0	86.6±1.8	79.1±1.7	64.6±1.3	64.1±1.1	67.5±1.6	70.5
SimCLR	30.0±1.4	28.3±0.9	31.0±1.0	49.5±3.4	38.0±1.7	30.2±1.0	26.7±1.4	29.2±1.8	32.8
SupCon	58.3±1.2	57.8±0.7	60.6±0.8	76.4±1.7	69.5±2.2	77.2±0.5	60.2±1.0	62.9±1.0	65.3
ProtoCLR	52.3±1.7	51.2±0.7	54.2±1.3	73.1±1.1	63.0±2.1	71.8±0.9	53.0±1.2	55.9±1.4	59.3

TABLE I

TOP-1 ACCURACY FOR ONE-SHOT AND FIVE-SHOT CLASSIFICATION. ALL REPORTED RESULTS ARE THE AVERAGE OF TEN RUNS. FOR OUR MODELS, THE RESULTS ARE FURTHER AVERAGED OVER THREE CHECKPOINTS. THE HIGHEST ACCURACIES ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND HIGHEST ARE UNDERLINED.

Split	Dataset	# of recordings	# of 5s examples	# of classes
Train	XCL	528,434	1,401,478	9,736
	XCM	89,798	189,880	411
Val	POW	14,911	51,394	48
Test	PER	16,802	65,479	132
	NES	16,117	58,251	89
	UHH	3,626	12,980	27
	HSN	5,460	17,940	21
	NBP	24,327	76,446	51
	SSW	28,403	92,514	81
	SNE	19,390	64,978	56

TABLE II

BIRDSET CLASSIFICATION BENCHMARK.

performs comparably to BioLingual on both one-shot and five-shot tasks. Perch slightly outperforms CE on XCL, suggesting that incorporating taxonomic ranks as auxiliary tasks could be a promising direction for future research.

These results indicate that ProtoCLR and SupCon achieves strong few-shot performance especially in one-shot learning scenarios.

V. CONCLUSION

In this work, we addressed the challenge of domain generalization for bird sound classification in few-shot scenarios, focusing on the domain shift from focal to soundscape recordings. We proposed a new few-shot benchmark derived from the BirdSet dataset to evaluate generalization capabilities of models trained on focal recordings and tested on soundscape recordings. Additionally, we introduced ProtoCLR, an alternative to SupCon inspired by prototypical networks, with reduced computational complexity. Few-shot evaluation under two pre-training scenarios—one with recordings limited to test classes and another with a larger, more diverse set—showed that pre-training on the set containing only test classes leads to better transfer performance.

Future work could explore techniques to scale to larger class sets without compromising performance, as well as investigate domain adaptation methods, particularly source-free adaptation, which is suited for few-shot bioacoustics when source data is unavailable or too costly to retrain on.

References

[1] E. Browning, R. Gibb, P. Glover-Kapfer, and K. E. Jones, "Passive Acoustic Monitoring in Ecology and Conservation." 2017.

[2] D. Stowell, "Computational Bioacoustics with Deep Learning: A Review and Roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[3] W.-P. Vellinga and R. Planqué, "The Xeno-canto Collection and its Relation to Sound Recognition and Classification." in *CLEF (Working Notes)*, 2015.

[4] 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.

[5] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[6] I. Moummad, N. Farrugia, and R. Serizel, "Self-Supervised Learning for Few-Shot Bird Sound Classification," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 600–604.

[7] ——, "Regularized Contrastive Pre-training for Few-shot Bioacoustic Sound Detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1436–1440.

[8] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[10] X. Zhang, L. Zhou, R. Xu, P. Cui, Z. Shen, and H. Liu, "Towards Unsupervised Domain Generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4910–4920.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," *Advances in neural information processing systems*, vol. 30, 2017.

[12] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, J. Lange, S. Kahl, B. Sick, S. Tomforde, and C. Scholz, "BirdSet: A Dataset and Benchmark for Classification in Avian Bioacoustics," *arXiv preprint arXiv:2403.10380*, 2024.

[13] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin *et al.*, "Learning to detect an animal sound from five examples," *Ecological informatics*, vol. 77, p. 102258, 2023.

[14] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "Metaaudio: A Few-Shot Audio Classification Benchmark," in *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 219–230.

[15] S. Kahl, M. Clapp, W. A. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, and A. Joly, "Overview of BirdCLEF 2020: Bird Sound Recognition in Complex Acoustic Environments," in *CLEF 2020-Conference and Labs of the Evaluation Forum*, vol. 2696, no. 262, 2020.

[16] J. Hamer, E. Triantafillou, B. van Merrienboer, 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] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain Generalization with Adversarial Feature Learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[18] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards Domain Generalization using Meta-Regularization," *Advances in neural information processing systems*, vol. 31, 2018.

[19] R. Volpi and V. Murino, "Addressing Model Vulnerability to Distributional Shifts over Image Transformation Sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7980–7989.

[20] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[21] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[22] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," 2023.

[23] F. Michaud, J. Sueur, M. Le Cesne, and S. Haupert, "Unsupervised classification to improve the quality of a bird song recording dataset," *Ecological Informatics*, vol. 74, p. 101952, 2023.

[24] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, 2023.

[25] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf *et al.*, "BioCLIP: A Vision Foundation Model for the Tree of Life," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 412–19 424.

[26] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. Van Der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv preprint arXiv:1911.04623*, 2019.

[27] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable Models for Bioacoustics with Human Language Supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1316–1320.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[29] 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.

[30] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised Contrastive Learning of Sound Event Representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 371–375.

[31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[32] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.