Fine-tuning for Bird Sound Classification: An Empirical Study

David Stein[1] and Björn Andres[1,2]

[1] TU Dresden
{david.stein1,bjoern.andres}@tu-dresden.de
[2] Center for Scalable Data Analytics and AI Dresden / Leipzig

Abstract. In biodiversity research scientists now routinely acquire audio recordings of vocalizing bird species and are then faced with the task of identifying the species audible in these recordings. Here, we analyze the accuracy (precision, recall and F_1 score) of several deep networks, in conjunction with pre-training and data augmentation techniques, for classifying audio recordings of twelve bird species under multiple data scarcity settings.

Keywords: Bird Sound Recognition · Classification · Fine-tuning

1 Introduction

A well-established indicator of ecosystem health is the variety of bird species [10]. Traditionally, ornithologists identify and count bird species in the study areas. This approach is time-consuming and expensive. Another approach is the combination of passive acoustic monitoring [7, 17, 27, 35] and automated bird sound classification [24] which facilitates the implementation of monitoring systems. These systems can enable long-term measurements of biodiversity and can assist ornithologists. The latter approach is shown to be effective, e.g., in [25, 43]. The annual BirdCLEF Challenge [11, 19, 20, 22, 23] fosters innovation in model and algorithm development for bird sound classification, the most effective of which are deep networks [11, 13, 21, 23–25, 32, 34].

Inferring decisions with a deep network from data distributed unlike the data it was trained on can result in lower accuracy. Thus, a fine-tuning of the parameters of the network can become necessary. This requires the annotation of additional data, which can be expensive.

In order to study this data efficiency problem in the context of passive acoustic monitoring of bird species, we make the following contributions: Firstly, we collect and annotate audio recordings of twelve bird species in a rather challenging environment. Secondly, we learn the parameters of several convolutional networks with a varying fraction of available training data, both from a random initialization and from an initialization learned on a large pre-training dataset. Thirdly, we compare the different models and analyze how much data is necessary to obtain satisfactory results. The source code to reproduce the experiments is available from [37]. The collected data and annotations are available upon request.

2 Related Work

The task of bird sound classification is studied in [11, 13, 21, 23–25, 32, 34], most notably BirdNET [25]. The authors of [39] evaluate transfer learning of hybrid CNN-LSTMs [1,2], attention-based CNN-LSTMs [47] and multi-head transformers [9,41] for the task of bird sound classification. An interpretable convolutional network for bird sound classification is proposed in [16], which computes prototypes of bird sounds.

Apart from the task of classification, the authors of [6, 28, 33, 38] study the task of clustering bird sounds by k means clustering [33], k nearest neighbor clustering [6], clustering with respect to all elements of three given clusters [28, Section 2.2] and correlation clustering [38].

Self-supervised representation learning for natural images is studied in [3–5, 14, 36, 46]. The authors of [29] benchmark the contrastive self-supervised methods SimCLR [5], Barlow Twins [46] and FroSSL [36] for the task of few-shot bird sound classification, showing competitive results compared to supervised learning approaches. A joint contrastive and generative pretext task for the task of audio representation learning is suggested in [12].

The authors of [42] propose a model for sound separation that separate multiple audio sources audible in the same sound recording. This approach has been shown to increase the accuracy of bird sound classification [8].

3 Models

3.1 Data Representation

We consider a finite, non-empty set S of sound recordings, a feature space $X \subseteq \mathbb{R}^n$ with $n \in \mathbb{N}$, a finite, non-empty set C of classes, and a label space $Y = \{0,1\}^C$, i.e., a multi-label classification problem. Moreover, we consider the features $x\colon S \to X$ and labels $y\colon S \to Y$. In our experiments, each recording $s \in S$ has a duration of 3 seconds and its features x_s are given by the pixel values of the mel spectrogram computed with a frame width of 1024 samples, an overlap of 768 samples and 128 mel bins, rescaled to 128×384 pixels.

3.2 Architectures

We employ the convolutional networks ResNet-{18, 34, 50} [15], EfficientNet (small) [40] and WideResNet50 [45]. We adjust these architectures by setting the number of input channels for the first convolutional layer to one. Apart from these architectures, we re-implement BirdNet [25]. In addition, we evaluate the BirdNet-Analyzer [18]. Each model defines a function $f_\theta\colon X \to \mathbb{R}^C$ where $\theta \in \Theta = \mathbb{R}^m$ are the $m \in \mathbb{N}$ parameters of the respective architecture.

4 Learning

The parameters of the models are learned by minimization of the logistic loss. In practice, we compute the parameters $\hat{\theta} \in \mathbb{R}^m$ heuristically by stochastic gradient descent with an adaptive learning rate. As software, we employ AdamW [26] with mini-batches $B \subseteq S$. As hardware, we use a single NVIDIA A100 GPU with 16 AMD EPYC CPU 7352 cores, equipped with 32 GB of RAM.

We set the initial learning rate to 10^{-4}, the batch size to 64, and the maximum number of steps to 400,000. We terminate the optimization as soon as the F_1 score on the training set has not improved for 2,000 steps. The class-distribution within each batch is on average uniform, i.e., we use over-sampling to counteract class imbalance.

5 Experiments

5.1 Dataset

In this section, we describe the fine-tuning and pre-training datasets. We optionally apply data augmentation during learning, more specifically, horizontal and vertical roll, SpecAugment [30], and the addition of white noise, pink noise or environmental noise from ESC-50 [31]. We apply each type of augmentation with a probability of 20%. Thus, a sample can be augmented by multiple techniques but at most one type of noise. For further detail, we refer to the source code [37].

Fine-tuning Dataset. We collect recordings at a sampling rate of 32 kHz with a total of 60 AudioMoths, a low cost recorder [17], deployed at 75 different sites. Using BirdNET [25] we filter 3s-chunks of audio for the presence of one of 12 bird species. These chunks are validated and, if possible, corrected by ornithologists. This defines a dataset $\mathcal{D}$; see Table 1 and Figure 1. We partition $\mathcal{D}$ into five stratified folds $\{\mathcal{D}_i\}_{i \in \{0,1,2,3,4\}}$. Each fold defines a training and a test

Bird Species	# Sites	n_c
Alauda arvensis	46	1056
Lanius collurio	12	606
Motacilla flava	41	491
Passer montanus	20	268
Saxicola torquatus	13	340
Sylvia curruca	10	636
Anthus pratensis	25	367
Emberiza calandra	21	660
Emberiza citrinella	40	626
Linaria cannabina	20	480
Saxicola rubetra	6	117
Sylvia communis	24	611

Table 1: Above, we report for the fine-tuning dataset $\mathcal{D}$ and for each species $c \in C$ the number of sites the species has been detected at and the number of samples n_c.

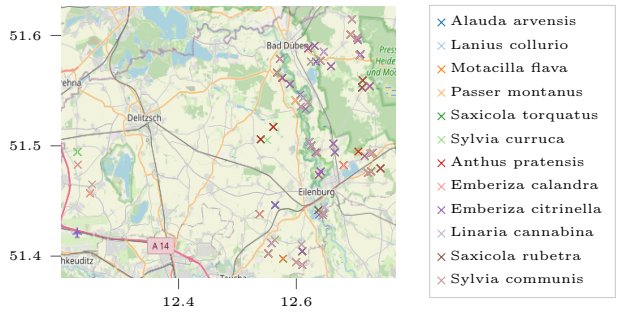

Fig. 1: Above, we show for the fine-tuning dataset $\mathcal{D}$ and for each species $c \in C$ the geographic coordinates of the sites the species has been detected at.

set, $\text{TRAIN}_i = \mathcal{D} \backslash \mathcal{D}_i$ and $\text{TEST}_i = \mathcal{D}_i$. Let n_c be the number of instances in TRAIN_i containing the bird species $c \in C$. With respect to an additional parameter $f \in [0, 1]$, we define the training datasets $\text{TRAIN}_{i,f}$, which is a randomly sampled subset of TRAIN_i and contains $\lfloor f \cdot n_c \rfloor$ samples of bird species c. We sample in such a way that for any $f' \geq f$, $\text{TRAIN}_{i,f} \subseteq \text{TRAIN}_{i,f'}$.

Pre-training Dataset. From Xeno-Canto [44], we collect 317,656 recordings of a total of 8,968 bird species of quality A or B, each recording containing only a single species. We resample the files to 44,100 Hz and split them into 3s-chunks. We apply the signal detector proposed in [25] to exclude non-salient chunks. This defines the dataset PRE-TRAIN consisting of 1,026,539 audio chunks.

Metrics. In order to measure the accuracy of decisions $\hat{y} \colon S \to \{0, 1\}^C$, we compute the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for every class $c \in C$ individually. From these, we compute the macro-averaged precision, recall and F_1 score.

5.2 Experiments

We consider three initializations of the parameters θ of the models defined in Section 3.2: (1) randomly distributed parameters, (2) parameters learned from PRE-TRAIN without augmentation, and (3) parameters learned from PRE-TRAIN with augmentation. We learn the initializations (2) and (3) as described in Section 4. For all initializations, we learn the parameters on the datasets $\text{TRAIN}_{i,f}$, with and without augmentation, and apply it to the independent dataset TEST_i, for every $i \in \{0, 1, 2, 3, 4\}$ and for every $f \in \{0.1j \mid 1 \leq j \leq 10\}$. After completion of the optimization and for every model, we infer independent decisions $\hat{y}^s \in \{0, 1\}^C$ for every recording $s \in S$ and class $c \in C$ by asking whether $f_\theta(x^s)_c \geq 0$ ($\hat{y}^s_c = 1$) or $f_\theta(x^s)_c < 0$ ($\hat{y}^s_c = 0$).

In Figure 2 we show the mean, minimum and maximum F_1 score over all folds, $0 \leq i \leq 4$, as a function of the fraction of samples f, and for models initialized

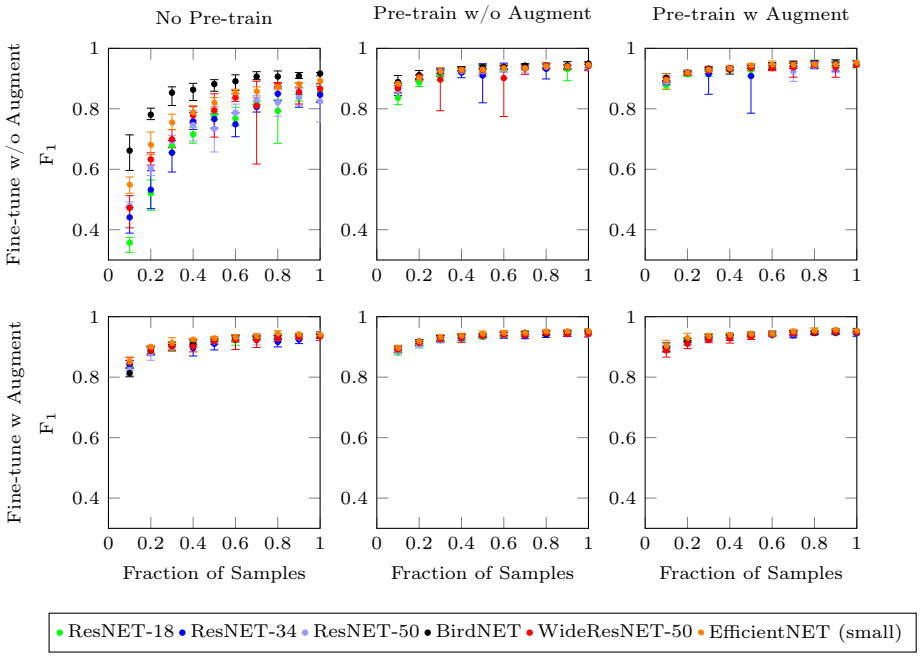

Fig. 2: We depict above the F_1 score of the models ResNET-{18, 34, 50}, BirdNET, WideResNET-50 and EfficientNet (small) as a function of the fraction of samples per species. We show these for models initialized with randomly distributed parameters (left) and from parameters learnt on PRE-TRAIN without (middle) and with (right) augmentations. Also, we report these metrics without (top) and with (bottom) augmentation during fine-tuning. Each datapoint shows the mean, minimum and maximum of the F_1 score across the five test folds.

with random parameters (Column 1), with parameters learned from PRE-TRAIN without augmentation (Column 2), with parameters learned from PRE-TRAIN with augmentation (Column 3); and with parameters fine-tuned on $\text{TRAIN}_{i,f}$, without augmentation (Row 1) and with augmentation (Row 2). In Table 2, we report for each model at $f = 1.0$ the mean of the macro-averages of the precision, the recall and the F_1 score across the five test folds. In Table 3, we report for $f = 1.0$ and the model with the highest mean F_1 score the precision, recall and F_1 score for each species of bird.

From Figure 2 (Column 1), we observe: On the one hand, learning without augmentations during fine-tuning (Row 1) leads to unstable results for the F_1 score across the five test folds, e.g., the difference between the minimum and maximum F_1 score across the five folds for the BirdNET architecture at $f = 0.1$ can be as high as 10%. On the other hand, applying augmentations during fine-tuning (Row 2) leads to more consistent F_1 scores across the five folds for all considered models and all fractions of samples f. Moreover, when learning

without data augmentations (Row 1) no model except the BirdNET architecture shows saturation of the F_1 score, even when using all data samples, $f = 1.0$. In contrast, learning with data augmentations (Row 2) leads to almost saturated F_1 scores, even for $f \geq 0.3$. Lastly, we observe an increase in the F_1 score when using augmentations as opposed to not using augmentations, which is higher for smaller datasets used for fine-tuning, i.e., a low f.

	Pre-train	FT Aug.	Precision	Recall	F_1
BirdNet-Analyzer (conf. = 0.1) [18]					
0.	-	-	88.5	81.0	83.3
ResNet-18					
1.	no	no	93.1	79.4	85.3
2.	no	yes	94.8	92.7	93.7
3.	yes	no	**95.9**	92.6	94.2
4.	yes	yes	95.5	93.7	94.5
5.	yes + Aug.	no	95.8	**93.8**	**94.7**
6.	yes + Aug.	yes	95.5	93.7	94.5
ResNet-34					
1.	no	no	92.2	79.2	84.7
2.	no	yes	95.0	92.7	93.8
3.	yes	no	95.5	93.6	94.5
4.	yes	yes	95.1	93.7	94.3
5.	yes + Aug.	no	**95.8**	**94.3**	**95.0**
6.	yes + Aug.	yes	95.4	93.7	94.5
ResNet-50					
1.	no	no	90.1	78.3	82.4
2.	no	yes	94.8	92.9	93.7
3.	yes	no	**95.5**	92.9	94.1
4.	yes	yes	94.9	94.0	94.4
5.	yes + Aug.	no	**95.5**	93.9	94.7
6.	yes + Aug.	yes	95.4	**94.5**	**94.9**

	Pre-train	FT Aug.	Precision	Recall	F_1
BirdNet [25]					
1.	-	no	92.5	91.1	91.7
2.	-	yes	94.1	93.4	93.6
3.	yes	no	**94.7**	95.3	95.0
4.	yes	yes	94.3	95.9	95.1
5.	yes + Aug.	no	94.5	95.2	94.8
6.	yes + Aug.	yes	94.5	**96.0**	**95.2**
WideResNet-50					
1.	no	no	92.1	82.6	86.6
2.	no	yes	94.9	92.4	93.5
3.	yes	no	95.5	93.2	94.2
4.	yes	yes	95.2	93.6	94.3
5.	yes + Aug.	no	95.6	93.9	94.6
6.	yes + Aug.	yes	**95.7**	**94.1**	**94.8**
EfficientNet (small)					
1.	no	no	91.4	87.4	89.2
2.	no	yes	94.5	93.7	94.0
3.	yes	no	95.2	93.9	94.4
4.	yes	yes	95.1	95.2	95.1
5.	yes + Aug.	no	**95.5**	94.8	95.1
6.	yes + Aug.	yes	95.4	**95.3**	**95.3**

Table 2: Above, we report the median of the macro-averaged precision, recall and F_1 score across the five test folds for $f = 1$. We mark in boldface the best entry for a specific model, and underline the best entry across all models.

Bird Species	Precision	Recall	F_1
Alauda arvensis	90.7	90.4	90.6
Lanius collurio	99.3	98.5	98.9
Motacilla flava	92.6	93.5	93.0
Passer montanus	95.9	95.4	95.6
Saxicola torquatus	95.1	90.7	92.8
Sylvia curruca	99.2	99.5	99.4
Anthus pratensis	96.8	97.6	97.1
Emberiza calandra	92.0	94.5	93.2
Emberiza citrinella	90.9	92.2	91.5
Linaria cannabina	98.9	98.3	98.6
Saxicola rubetra	97.4	95.7	96.5
Sylvia communis	95.7	97.5	96.6

Table 3: Above, we report the mean of the precision, recall and F_1 score across the five test folds per species and for the overall best performing model Efficient-Net (small).

In Figure 2 (Columns 2-3), we observe: Initializing the parameters of the models with parameters learned from PRE-TRAIN, with or without augmentation, leads to the F_1 score being more consistent across the five test folds, irrespective of whether or not augmentations are applied during fine-tuning. For $f \geq 0.3$, we obtain almost saturated F_1 scores with only marginal improvements for increasing the fine-tuning dataset size, i.e., increasing f. Applying augmentations during fine-tuning for both initializations leads to marginal improvements in the F_1 score for both model initializations.

In Table 2, we observe: Both using augmentations during fine-tuning and initializing the models with parameters learned from PRE-TRAIN lead to increasing performance of the model with respect to the F_1 score. With an F_1 score of 95.3%, the EfficientNet architecture initialized with parameters learned from PRE-TRAIN with augmentations and fine-tuned with augmentations (Row 6) is the best performing model in this study. Interestingly, from a comparision of ResNet-18 (Row 3), BirdNet (Row 6) and EfficientNet (Row 6), we observe that the EfficientNet architecture is neither the best performing model with respect to precision (by a margin of 0.5%) or recall (by a margin of 0.7%). Hence, the EfficientNet architecture offers the best tradeoff between precision and recall in this study.

In Table 3, we observe: On the one hand, even though Alauda arvensis is the majority species in the dataset (cfg. Table 1), the EfficientNet architecture performs worst for this species with respect to the F_1 score. On the other hand, the model's F_1 score for the minority class Saxicola rubetra is substantially higher than for the species Emberiza citrinella, Emberiza calandra and Motacilla flava for which we have more training data.

6 Conclusion

We examine six neural network architectures for bird sound classification empirically, in different data scarcity settings. For those audio recordings of 12 bird species considered in the experiments, we conclude that both data augmentation and initialization with parameters learned from a pre-training datasets stabilizes the accuracy of the respective models across different test folds and improves accuracy across all data scarcity settings. With all the available training data, the most accurate architecture is EfficientNet. With only 10% of the training data, i.e., between 10-80 recordings per species, we still obtain satisfactory results for the classification of the 12 species considered in this study. Adjusting the parameters to the data distribution increases the accuracy compared to a non-adjusted, pre-trained model by a margin of 12%.

Acknowledgement. The authors acknowledge funding by the Federal Ministry of Education and Research of Germany, from grant 16LW0079K. We would like to thank our colleagues at the Chair of Computational Landscape Ecology as well as Stefan Hannabach and Michael Bokämper for their help in collecting the data needed for this study.

References

1. Ashraf, M., Abid, F., Din, I.U., Rasheed, J., Yesiltepe, M., Yeo, S.F., Ersoy, M.T.: A hybrid cnn and rnn variant model for music classification. Applied Sciences **13**(3) (2023). https://doi.org/10.3390/app13031476
2. Ayadi, S., Lachiri, Z.: A combined cnn-lstm network for audio emotion recognition using speech and song attributs. In: ATSIP (2022). https://doi.org/10.1109/ATSIP55956.2022.9805924
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020), https://proceedings.mlr.press/v119/chen20j.html
6. Clementino, T., Colonna, J.: Using triplet loss for bird species recognition on BirdCLEF 2020. In: CLEF (working notes) (2020)
7. Darras, K., Batáry, P., Furnas, B., Celis-Murillo, A., Van Wilgenburg, S.L., Mulyani, Y.A., Tscharntke, T.: Comparing the sampling performance of sound recorders versus point counts in bird surveys: A meta-analysis. Journal of Applied Ecology **55**(6), 2575–2586 (2018). https://doi.org/10.1111/1365-2664.13229
8. Denton, T., Wisdom, S., Hershey, J.R.: Improving bird classification with unsupervised sound separation. In: ICASSP (2022). https://doi.org/10.1109/ICASSP43922.2022.9747202
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021), https://openreview.net/forum?id=YicbFdNTTy
10. Fitzpatrick, J.W., Lovette, I.J.: Handbook of bird biology. John Wiley & Sons (2016)
11. Goëau, H., Kahl, S., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification. In: CLEF (2018)
12. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: Self-supervised audio spectrogram transformer. In: AAAI (2022). https://doi.org/10.1609/aaai.v36i10.21315
13. Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., Ferres, J.L.: Comparing recurrent convolutional neural networks for large scale bird species classification. Scientific reports **11**(1), 17085 (2021). https://doi.org/10.1038/s41598-021-96446-w
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016). https://doi.org/10.1109/CVPR.2016.90
16. Heinrich, R., Sick, B., Scholz, C.: Audioprotopnet: An interpretable deep learning model for bird sound classification (2024), https://arxiv.org/abs/2404.10420
17. Hill, A.P., Prince, P., Snaddon, J.L., Doncaster, C.P., Rogers, A.: Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment. HardwareX **6**, e00073 (2019). https://doi.org/10.1016/j.ohx.2019.e00073

18. Kahl, S.: BirdNET Analyzer, https://github.com/kahst/BirdNET-Analyzer
19. Kahl, S., Clapp, M., Hopping, W.A., Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments. In: CLEF (2020)
20. Kahl, S., Denton, T., Klinck, H., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2021: Bird call identification in soundscape recordings. In: CLEF (working notes) (2021)
21. Kahl, S., Hussein, H., Fabian, E., Schloßhauer, J., Thangaraju, E., Kowerko, D., Eibl, M.: Acoustic event classification using convolutional neural networks. In: Informatik 2017. Gesellschaft für Informatik, Bonn (2017). https://doi.org/10.18420/in2017_217
22. Kahl, S., Navine, A., Denton, T., Klinck, H., Hart, P., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings. In: CLEF (working notes) (2022)
23. Kahl, S., Stöter, F.R., Goëau, H., Glotin, H., Planque, R., Vellinga, W.P., Joly, A.: Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes. In: CLEF (2019)
24. Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-scale bird sound classification using convolutional neural networks. In: CLEF (working notes) (2017)
25. Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics **61**, 101236 (2021). https://doi.org/10.1016/j.ecoinf.2021.101236
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019), https://openreview.net/forum?id=Bkg6RiCqY7
27. Markova-Nenova, N., Engler, J.O., Cord, A.F., Wätzold, F.: Will passive acoustic monitoring make result-based payments more attractive? A cost comparison with human observation for farmland bird monitoring. Conservation Science and Practice **5**(9), e13003 (2023). https://doi.org/10.1111/csp2.13003
28. McGinn, K., Kahl, S., Peery, M.Z., Klinck, H., Wood, C.M.: Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Ecological Informatics **74**, 101995 (2023). https://doi.org/10.1016/j.ecoinf.2023.101995
29. Moummad, I., Serizel, R., Farrugia, N.: Self-supervised learning for few-shot bird sound classification. In: ICASSPW SASB (2024)
30. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In: Interspeech (2019). https://doi.org/10.21437/Interspeech.2019-2680
31. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. In: ACM Conference on Multimedia (2015). https://doi.org/10.1145/2733373.2806390
32. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. Signal Processing Letters **24**(3), 279–283 (2017). https://doi.org/10.1109/LSP.2017.2657381
33. Seth, H., Bhatia, R., Rajan, P.: Feature learning for bird call clustering. In: ICIIS (2018). https://doi.org/10.1109/ICIINFS.2018.8721418
34. Sevilla, A., Glotin, H.: Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: CLEF (working notes) (2017)
35. Shonfield, J., Bayne, E.M.: Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation & Ecology **12**(1) (2017). https://doi.org/10.5751/ACE-00974-120114

36. Skean, O., Dhakal, A., Jacobs, N., Giraldo, L.G.S.: Frossl: Frobenius norm minimization for efficient multiview self-supervised learning (2024), https://arxiv.org/abs/2310.02903
37. Stein, D.: Fine-tuning for bird sound classification: An empirical study: Code, https://github.com/dsteindd/finetuning-for-bird-sound-classification-an-empirical-study
38. Stein, D., Andres, B.: Correlation clustering of bird sounds. In: GCPR (2024). https://doi.org/10.1007/978-3-031-54605-1_33
39. Swaminathan, B., Jagadeesh, M., Vairavasundaram, S.: Multi-label classification for acoustic bird species detection using transfer learning approach. Ecological Informatics **80**, 102471 (2024). https://doi.org/10.1016/j.ecoinf.2024.102471
40. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019), http://proceedings.mlr.press/v97/tan19a.html
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NIPS (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
42. Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R.J., Wilson, K., Hershey, J.R.: Unsupervised sound separation using mixture invariant training. In: NeurIPS (2020), https://proceedings.neurips.cc/paper/2020/file/28538c394c36e4d5ea8ff5ad60562a93-Paper.pdf
43. Wood, C.M., Gutiérrez, R.J., Peery, M.Z.: Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. Ecology **100**(9) (2019). https://doi.org/10.1002/ecy.2764
44. Xeno-canto: Sharing wildlife sounds from around the world (2023), https://xeno-canto.org/about/xeno-canto
45. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
46. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: ICML (2021), http://proceedings.mlr.press/v139/zbontar21a.html
47. Zhang, Z., Xu, S., Zhang, S., Qiao, T., Cao, S.: Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing **453**, 896–903 (2021). https://doi.org/10.1016/j.neucom.2020.08.069