AudioSAM: Automated Annotation of Bioacoustic Soundscapes in the Wild

Zachary Baker **Reece Shuttleworth** **Daniela Rus**
Antonio Torralba **Jacob Andreas** **Pratyusha Sharma**
MIT CSAIL
{zbaker, rshuttle, rus, jda, torralba, pratyusha}@csail.mit.edu

Abstract

Automated analysis of bioacoustic recordings is essential for monitoring biodiversity and ecosystem health, yet current methods struggle with the complexity of natural soundscapes and the scarcity of labeled data. We introduce a bioacoustic Masked Autoencoder (a self-supervised framework) designed to learn robust audio representations from large-scale, unlabeled recordings. Pretrained on over 15,000 hours of diverse terrestrial and marine audio, our model—a 1B-parameter Vision Transformer encoder paired with a 500M-parameter decoder—learns representations that generalize across species and habitats. When evaluated on multiple bioacoustic benchmarks, our model achieves state-of-the-art performance among foundation models in both vocalization detection and species classification tasks. We further demonstrate the benefits of combining supervised and unsupervised contrastive objectives for species-aware embeddings. Our contributions include: (1) a large-scale unified dataset of bioacoustic recordings, (2) a pretrained foundation model for bioacoustic analysis (which we call *AudioSAM*), and (3) evidence that self-supervised learning enables scalable, label-efficient monitoring of global biodiversity. More results and visuals can be found at LINK.

1 Introduction

Bioacoustic monitoring has emerged as a critical tool for ecological research, wildlife conservation, and biodiversity assessment Stowell et al. [2016], Marques et al. [2012]. By recording and analyzing animal vocalizations, researchers can track population dynamics, detect species presence, and monitor ecosystem health without physical intervention. However, the automated analysis of bioacoustic data presents significant challenges, particularly in natural environments where recordings contain diverse species, background noise, and complex acoustic events Stowell and Plumbley [2014], Mesaros et al. [2020].

In this paper, we introduce a specialized Masked Autoencoder for bioacoustic data, which we call *AudioSAM* (Audio Self-supervised Animal Model), that learns robust audio representations without reliance on labeled examples. Our approach builds upon recent advances in self-supervised audio representation learning, particularly Audio-MAE [Huang et al., 2022], while incorporating several innovations tailored to the challenges of bioacoustic analysis. We also leverage a diverse collection of bioacoustic recordings spanning terrestrial and marine environments to ensure our model learns representations applicable across varied ecological contexts.

Our primary contributions are as follows. **Model:** we release a 1 billion-parameter Vision Transformer (ViT) encoder trained on extensive bioacoustic data, capable of handling long audio sequences. **Unified Dataset:** we compile and release a collection of dozens of bioacoustic datasets with unified annotations, significantly expanding training data diversity. **Benchmark Results:** we demonstrate that our model achieves strong results on multiple bioacoustic benchmarks, outperforming prior foundation models in both detection and classification tasks.

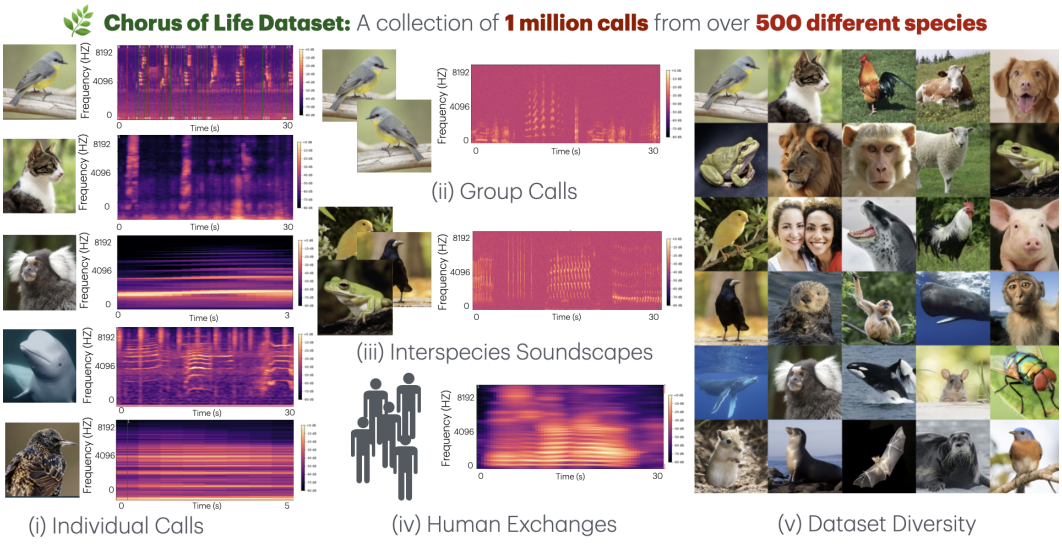

Figure 1: Overview of the *Chorus of Life* dataset. This dataset consists of over 1 million calls from 500+ species, showcasing a wide range of acoustic contexts. **(i)** Individual Calls: examples of isolated vocalizations from different species, represented as spectrograms. **(ii)** Group Calls: recordings of multiple individuals of the same species vocalizing together, highlighting overlapping patterns. **(iii)** Interspecies Soundscapes: complex acoustic environments where calls from multiple species co-occur, mimicking real-world habitats. **(iv)** Human Exchanges: human speech interactions included in the dataset to support human-animal communication studies. **(v)** Dataset Diversity: visual representation of species diversity, including birds, mammals, marine animals, and humans.

2 Related Work

Species-specific supervised models. Early approaches to automatic bioacoustic classification were often tailored to a single taxon or task. For example, *BirdNET* Kahl et al. [2021] is a highly successful supervised model for bird-song identification, trained on large annotated datasets (e.g., Xeno-Canto). More recently, *Perch* Ghani et al. [2023] extended supervised bioacoustic modeling to a broader range of species; Perch 1.0 (2023) and its update Perch 2.0 Tang et al. [2025] are large convolutional networks that serve as strong pretrained classifiers for various animal sounds. While effective within their domains, these models rely on extensive labeled data and tend to struggle when generalizing beyond the species or acoustic conditions they were trained on.

Self-supervised foundation models. To overcome the limitations of labeled-data scarcity, several works have explored self-supervised learning for bioacoustics. The Earth Species Project introduced *AVES* Hagiwara [2022], the first self-supervised foundation model for animal vocalizations, and later *BirdAVES* Earth Species Project [2024], a 316M-parameter HuBERT-based model specializing in bird calls (achieving over 20% improvement on bird sound tasks compared to AVES). Concurrently, Rauch et al. proposed Bird-MAE Rauch et al. [2025], which applies an Audio-MAE pretraining approach on a large corpus of avian recordings (BirdSet). While both Bird-MAE and our approach leverage masked spectrogram reconstruction, AudioSAM differs in several key aspects: it is trained on a much broader taxonomic range beyond birds (including mammals, marine animals, and amphibians), uses a substantially larger ViT encoder (1B vs. 300M parameters), and draws on a larger, more diverse corpus (15,000 hours vs. 4,000 hours). Whereas Bird-MAE focuses on bird-specific detection, AudioSAM is designed for general-purpose transfer across taxa and habitats. Beyond broad performance, some works have also prioritized interpretability—e.g., *AudioProtoPNet* Heinrich et al. [2025] uses a prototypical part-learning mechanism to classify bird calls with human-interpretable audio prototypes. These efforts mark a shift toward general-purpose yet domain-tuned models that leverage large unlabeled corpora across the animal kingdom.

Dataset	Num Calls	Duration	Dataset	Num Calls	Duration
AudioSet Gemmeke et al. [2017]	0	5,800 h	Giant Otters Mumm and Knörnschild [2014]	441	1 h
Animal Sounds Jayaya [2025]	809	1 h	Hainan Gibbons Dufourq et al. [2020]	1233	104 h
Anuraset Cañas et al. [2023]	16089	27 h	Hawaii Birds Navine et al. [2022]	59583	51 h
Bengal Finch Nicholson et al. [2017]	1215	5 h	HICEAS Yano et al. [2018]	796	13 h
BIRDeep Márquez-Rodríguez et al. [2024]	3749	9 h	Macaques Fukushima et al. [2015]	7285	1 h
BirdVox Lostanlen et al. [2018]	35402	18 h	Infant Marmosets Sarkar and Magimai.-Doss [2023]	169318	59 h
Domestic Canaries Belzner et al. [2009]	14407	4 h	Multimodal Birds Kumar et al. [2023]	6524	4 h
Columbia/CR Álvaro Vega-Hidalgo et al. [2023]	7338	35 h	Northeast US Kahl et al. [2022a]	50760	285 h
DARPA dar	1718	5 h	Orca Sounds Internet Archive	398	1 h
Avian Dawn Weldy et al. [2024]	41183	132 h	Pig Sounds Briefer et al. [2022]	6887	1 h
DCASE	7206	18 h	Rainforest Yassin et al. [2020]	1216	21 h
Fruit Bats Prat et al. [2017]	90000	38 h	Rodent Sounds Tachibana [2019]	4576	1 h
ENA Birds Chronister et al. [2021]	16052	7 h	Sierra Nevada Clapp et al. [2023]	10976	17 h
ESC Piczak	400	1 h	Southwest Amazon Hopping et al. [2022]	16482	22 h
Rook Birds Martin et al. [2022]	21662	29 h	SSW Van Horn et al. [2022]	3861	11 h
			Watkins Marine Sounds Sayigh et al. [2016]	15152	30 h
			Western US Kahl et al. [2022b]	20147	33 h
			Sperm Whales	14764	250 h

Table 1: Curated datasets with the number of annotated calls (each continuous annotation is counted as one call) and total duration.

Audio–language models. Multi-modal foundation models have also emerged for bioacoustics. *BioLingual* Robinson et al. [2023] constructs joint audio-text representations by training on a curated "AnimalSpeak" dataset of over one million audio-caption pairs (combining metadata from sources like AudioSet and xeno-canto). This contrastive audio–language pretraining enables zero-shot recognition of calls via text prompts and free-form text queries for sound retrieval. Similarly, *SALMONN* Tang et al. [2023] and *NatureLM-audio* Robinson et al. [2024] combine audio encoders with large language models, allowing cross-modal understanding of animal sounds (e.g., describing calls or performing call-type classification without explicit training on those specific classes). Such audio–language models represent a parallel avenue toward generalizable bioacoustic AI, using natural language supervision to achieve cross-species generalization.

Collectively, these studies illustrate an evolution from specialized, single-taxon classifiers to general-purpose, large-scale models capable of cross-domain transfer. Our work builds on these advances by integrating the strengths of self-supervised foundation models. In doing so, we aim to provide a unified framework that can robustly detect, classify, and interpret animal vocalizations across a wide range of taxa and real-world conditions.

3 Dataset Creation

We curated a comprehensive pretraining dataset alongside a large collection of labeled evaluation datasets by combining recordings from numerous sources covering diverse taxonomic groups and ecological environments. The primary datasets incorporated are summarized in Table 1. In total, we gathered more than 7,000 hours of audio and over 1 million annotated vocalization events across 30 genera and 500+ species.

We also apply extensive data augmentation to increase the effective size and diversity of the training data. Techniques include: (i) `Mixing` – overlaying multiple calls to simulate overlapping vocalizations; (ii) `Stitching` – concatenating short clips to emulate longer calls or choruses; (iii) `Amplitude Modulation` – varying gain to imitate calls at different distances; (iv) `Noise Addition/Reduction` – injecting or removing background noise; and (v) `Varying FFT window` – changing spectrogram time-frequency resolution to capture species with different temporal scales. These augmentations expose the model to a broader range of acoustic conditions during training.

4 Pretraining

To learn robust audio representations without reliance on labeled data, we implement a Masked Autoencoder (MAE) pretraining framework for bioacoustic spectrograms. Our approach is inspired by masked reconstruction methods in vision and audio Huang et al. [2022], He et al. [2021], adapted to the challenges of animal sound data.

Model	DCASE	ENA Birds	HICEAS	Rainforest	Gibbons	ESC-50	WTK (Watkins)
LLM (no audio)	0.000	0.001	0.210	0.000	0.013	0.020	0.041
SALMONN	0.005	0.004	0.097	0.002	0.005	0.320	0.041
BioLingual	0.036	0.109	**0.429**	0.004	0.018	0.307	0.041
NatureLM-audio	0.058	0.314	0.336	0.025	0.005	0.600	0.257
AudioSAM (Ours)	**0.282**	**0.902**	0.304	**0.111**	**0.041**	**0.719**	**0.431**

Table 2: Performance on multiple bioacoustic benchmarks (after fine-tuning our model on each task). Columns *DCASE*, *ENA Birds*, *HICEAS*, *Rainforest*, and *Gibbons* report F1 scores for vocalization event detection Robinson et al. [2024], while *ESC-50* and *WTK* report classification accuracy. Best score per column is **bolded**.

Model Architecture. The encoder processes only the visible (unmasked) portions of the input mel-spectrogram, greatly reducing computation during pretraining. We employ a ViT-style transformer encoder with self-attention to capture long-range acoustic patterns. A lightweight decoder then reconstructs the full spectrogram (including the masked patches) from the encoded latent representations and positional embeddings. The large capacity of the encoder (1B parameters) enables learning rich representations from nuanced bioacoustic signals.

Pretraining Data. We pretrain AudioSAM on a diverse collection of unlabeled recordings, primarily drawn from the unlabeled portion of AudioSet Gemmeke et al. [2017] supplemented with additional curated and augmented bioacoustic data (Section 3). In total, approximately 15,000 hours of audio spanning birds, mammals (land and marine), amphibians, insects, and abiotic environmental sounds are used. This broad coverage encourages the model to learn generic acoustic features that transfer across species and ecosystems.

5 Model Evaluation

Following the BEANS benchmark Claudino et al. [2023], we evaluate our model on a representative set of bioacoustic detection and classification tasks spanning diverse taxa and habitats. We trained our model on all datasets in Table 1, excluding those designated for downstream evaluation. In particular, the DCASE, ENA Birds, HICEAS, Rainforest, and Hainan Gibbons corpora were withheld from pretraining so that we could later fine-tune and evaluate on these as unseen benchmarks. This strategy tests the model's ability to transfer to new species and recording conditions that were not part of its unsupervised training data.

Vocalization Detection. For the detection task, we initialized the model with our self-supervised ViT encoder and then attached a lightweight binary event detector head. We fine-tuned this model on each benchmark's training set to predict vocalization events (start/end times) in audio. Table 2 summarizes the fine-tuning results, reported as F1 scores on the evaluation sets (for the detection benchmarks) and classification accuracy (for ESC-50 and Watkins). We also provide an example of qualitative detection results in Figure 2.

Overall, AudioSAM achieves the highest scores on most benchmarks compared to previous foundation models. Notably, it more than quadruples the F1 on DCASE (0.282 vs. 0.058) and substantially improves Rainforest and Gibbons detection, indicating superior generalization to new soundscape data. For ESC-50 and Watkins classification, our model also leads prior foundation models. However, we note that fully supervised models such as Perch 2.0 report higher accuracy on these tasks, reflecting the continued advantage of task-specific supervision in settings with abundant labels.

6 Conclusion

We introduced a scalable self-supervised framework for bioacoustic monitoring, enabling robust representation learning from complex, noisy soundscapes without reliance on manual annotation. By leveraging large-scale audio data and contrastive learning objectives, our approach yields strong performance in event detection and species identification across diverse ecosystems. In particular, AudioSAM's foundation model achieves competitive results on challenging benchmarks while using

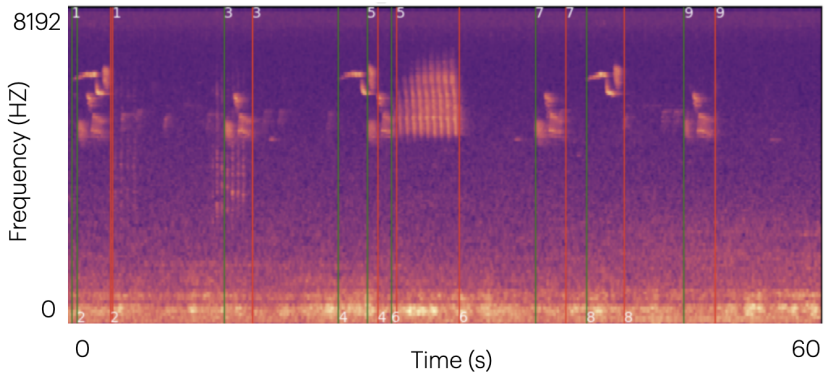

Figure 2: Sample model detection results on an animal vocalization audio clip. Green bars indicate predicted call start times and red bars indicate end times; each detected call is numbered. The model accurately identifies individual calls within a noisy soundscape.

far fewer labeled examples than previous supervised models. Future directions include integrating multi-modal environmental signals (e.g., visual or textual context) and exploring prompt-based or zero-shot recognition capabilities, as well as deploying lightweight versions of the model for real-time field applications. These steps will further advance automated, scalable biodiversity monitoring on a global scale.

Acknowledgments

We are deeply grateful to Project CETI for their collaboration and for granting access to datasets that made this research possible. We thank Etched for providing GPUs and compute support, and CSAIL for fostering an environment that enabled this work. We thank Pratyusha Sharma for her invaluable advising, guidance, and encouragement throughout the project. We also thank Martin Rodriguez for his prior work and for our many conversations that helped shape this research.

References

Dan Stowell, Mike Wood, Yannis Stylianou, and Herve Glotin. Bird detection in audio: A survey and a challenge, 2016.

Tiago Marques, Len Thomas, Stephen Martin, David Mellinger, Jessica Ward, David Moretti, Danielle Harris, and Peter Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2): 287–309, 2012.

Dan Stowell and Mark Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, page 488, 2014.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Bioacoustic challenges in dcase 2020, 2020. Dataset.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. URL `https://doi.org/10.1016/j.ecoinf.2021.101236`.

Uzayr Ghani, Brian Chitwood, Peggy Tang, Yanming Liao, and Masatoshi Hagiwara. Perch: A large-scale audio pretraining approach for bioacoustics. `https://github.com/earthspecies/perch`, 2023. Earth Species Project.

Peggy Tang, Uzayr Ghani, Brian Chitwood, and Masatoshi Hagiwara. Perch 2.0: Bioacoustics model for species identification. `https://hackernoon.com/perch-20-bioacoustics-model-for-species-identification`, 2025. Accessed: 2025-11-08.

Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision. *arXiv preprint arXiv:2210.14493*, 2022. URL https://arxiv.org/abs/2210.14493.

Earth Species Project. Introducing birdaves: Self-supervised audio foundation models for birds. https://www.earthspecies.org/blog/introducing-birdaves-self-supervised-audio-foundation-model-for-birds, 2024. Accessed: 2025-11-08.

Lukas Rauch, Aaron van den Oord, and Abdelrahman Mohamed. Can masked autoencoders also listen to birds? *arXiv preprint arXiv:2504.12880*, 2025. URL https://arxiv.org/abs/2504.12880.

René Heinrich, Christoph Scholz, and Bernhard Sick. Audioprotopnet: An interpretable deep learning model for bird sound classification. *Ecological Informatics*, 79:102446, 2025. URL https://doi.org/10.1016/j.ecoinf.2025.102446.

Daniel Robinson, Brian Chitwood, Yanming Liao, Peggy Tang, and Uzayr Ghani. Transferable models for bioacoustics with human language supervision. *arXiv preprint arXiv:2308.04978*, 2023. URL https://arxiv.org/abs/2308.04978.

Chengxi Tang, Wenyi Huang, Yuan Gong, et al. Salmonn: Towards generic hearing abilities for large audio-language models. *arXiv preprint arXiv:2310.13289*, 2023. URL https://arxiv.org/abs/2310.13289.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint arXiv:2411.07186*, 2024.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Christina Mumm and Mirjam Knörnschild. The vocal repertoire of adult and neonate giant otters (pteronura brasiliensis). *PLoS ONE*, 9:e112562, 11 2014. doi: 10.1371/journal.pone.0112562.

Maulana Akbar Dwi Jayaya. Animal sounds dataset. Kaggle, 2025. Accessed on March 5, 2025.

Emmanuel Dufourq, James Hansford, Amanda Hoepfner, Heidi Ma, Jessica Bryant, Christina Stender, Wenyong Li, Zhiwei Liu, Qing Chen, Zhaoli Zhou, and Samuel Turvey. Automated detection of hainan gibbon calls for passive acoustic monitoring, 09 2020.

Juan Sebastián Cañas, Maria Paula Toro-Gómez, Larissa Sayuri Moreira Sugai, Hernán Darío Benítez Restrepo, Jorge Rudas, Breyner Posso Bautista, Luís Felipe Toledo, Simone Dena, Adão Henrique Rosa Domingos, Franco Leandro de Souza, Selvino Neckel-Oliveira, Anderson da Rosa, Vítor Carvalho-Rocha, José Vinícius Bernardy, José Luiz Massao Moreira Sugai, Carolina Emília dos Santos, Rogério Pereira Bastos, Diego Llusia, and Juan Sebastián Ulloa. Anuraset: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring, 2023. URL https://arxiv.org/abs/2307.06860.

Amanda Navine, Stefan Kahl, Ann Tanimoto-Johnson, Holger Klinck, and Patrick Hart. A collection of fully-annotated soundscape recordings from the island of hawai'i. *Dataset on Zenodo, September*, 2022.

David Nicholson, Jonah E. Queen, and Samuel J. Sober. Bengalese Finch song repository, 10 2017. URL https://figshare.com/articles/dataset/Bengalese_Finch_song_repository/4805749.

Kymberly M Yano, Erin Marie Oleson, Jennifer L Keating, Lisa Taylor Ballance, Marie Chapla Hill, Amanda L Bradford, Ann N Allen, Trevor W Joyce, Jeffrey E Moore, and Annette Elizabeth Henry. Cetacean and seabird data collected during the hawaiian islands cetacean and ecosystem assessment survey (hiceas), july–december 2017, 2018.

Alba Márquez-Rodríguez, Miguel Angel Mohedano-Muñoz, Manuel Jesus Marin-Jimenez, Eduardo Santamaria-Garcia, Giulia Bastianelli, Pedro Jordano, and Irene Mendoza. Birdeepaudioannotations (revision 4cf0456), 2024. URL https://huggingface.co/datasets/GrunCrow/BIRDeep_AudioAnnotations.

Makoto Fukushima, Alex M Doyle, Matthew P Mullarkey, Mortimer Mishkin, and Bruno B Averbeck. Distributed acoustic cues for caller identity in macaque vocalization. *Royal Society open science*, 2(12):150432, 2015.

Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Birdvox-full-night: a dataset and benchmark for avian flight call detection. In *Proc. IEEE ICASSP*, April 2018.

Eklavya Sarkar and Mathew Magimai.-Doss. Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers? In *Proc. INTERSPEECH 2023*, pages 1189–1193, 2023. doi: 10.21437/Interspeech.2023-1968.

Sandra Belzner, Cornelia Voigt, Clive K Catchpole, and Stefan Leitner. Song learning in domesticated canaries in a restricted acoustic environment. *Proceedings of the Royal Society B: Biological Sciences*, 2009.

Sumit Kumar, B. Anshuman, Linus Rüttimann, Richard H.R. Hahnloser, and Vipul Arora. Balanced deep cca for bird vocalization detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094650.

Álvaro Vega-Hidalgo, Stefan Kahl, Laurel B. Symes, Viviana Ruiz-Gutiérrez, Ingrid Molina-Mora, Fernando Cediel, Luis Sandoval, and Holger Klinck. A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica, 2023.

Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the northeastern united states. *Dataset on Zenodo, August 2022a. URL https://doi. org/10.5281/zenodo*, 7079380, 2022a.

The darpa timit acoustic-phonetic continuous speech corpus.

Internet Archive. Orcas classification.

Matthew Weldy, Tom Denton, Abram Fleishman, Jaclyn Tolchin, Matthew McKown, Robert Spaan, Zachary Ruff, Julianna Jenkins, Matthew Betts, and Damon Lesmeister. Audio tagging of avian dawn chorus recordings in california, oregon and washington., 2024.

EF Briefer, CCR Sypherd, LMC Leliveld, M Padilla de la Torre, and C Tallet. The soundwel database: a labeled pig vocalization repository [data set]. *Scientific Reports*, 12:3409, 2022.

Bourhan Yassin, inversion, Mahreen Qazi, and Zephyr Gold. Rainforest connection species audio detection. `https://kaggle.com/competitions/rfcx-species-audio-detection`, 2020. Kaggle.

Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Scientific Data*, 4:sdata2017143, 10 2017. doi: 10.1038/sdata.2017.143.

Ryosuke O Tachibana. Dataset for usvseg performance test, 2019.

Lauren Chronister, Tessa Rhinehart, Aidan Place, and Justin Kitzes. An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information. *Ecology*, 102, 05 2021. doi: 10.1002/ecy.3329.

Mary Clapp, Stefan Kahl, Erik Meyer, Megan McKenna, Holger Klinck, and Gail Patricelli. A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range. *Dataset on Zenodo, January*, 2023.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL `http://dl.acm.org/citation.cfm?doid=2733373.2806390`.

W Alexander Hopping, Stefan Kahl, and H Klink. A collection of fully-annotated soundscape recordings from the southwestern amazon basin. 1. *Zenodo. URL: https://doi. org/10.5281/zenodo*, 7079124, 2022.

Killian Martin, Olivier Adam, Nicolas Obin, and Valérie Dufour. Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks. *Ecological Informatics*, 72, 12 2022. doi: 10.1101/2022.02.19.481011.

Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. In *European Conference on Computer Vision (ECCV)*, 2022.

Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics*, volume 27. AIP Publishing, 2016.

Stefan Kahl, Connor M Wood, Philip Chaon, M Zachariah Peery, and Holger Klinck. A collection of fully-annotated soundscape recordings from the western united states. *Dataset on Zenodo, September 2022c. URL https://doi. org/10.5281/zenodo*, 7050014, 2022b.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Leonardo Claudino, Yan Zhang, Julian Risch, Lisa Gill, and Dan Stowell. Beans: A benchmark for animal sound detection and classification, 2023. URL `https://arxiv.org/abs/2310.20502`.