BirdSet

A Large-Scale Dataset for Audio Classification in Avian Bioacoustics

Lukas Rauch[1a], Raphael Schwinger[2], Moritz Wirth[1,3], René Heinrich[1,3], Denis Huseljic[1] Marek Herde[1], Jonas Lange[2], Stefan Kahl[4], Bernhard Sick[1], Sven Tomforde[2], Christoph Scholz[1,3]

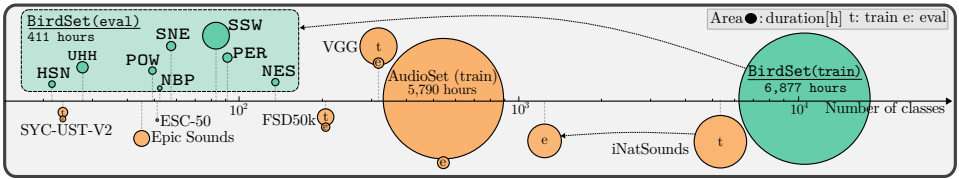
[1]University of Kassel [2]Kiel University [3]Fraunhofer IEE [4]TU Chemnitz

[a]lukas.rauch@uni-kassel.de

ICLR 2025

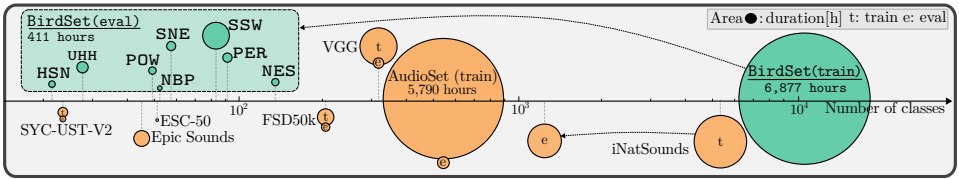
Motivation

- **General audio classification is challenged** by noise, label imbalance, domain shifts, and overlapping sources. **AudioSet has its limitations**.



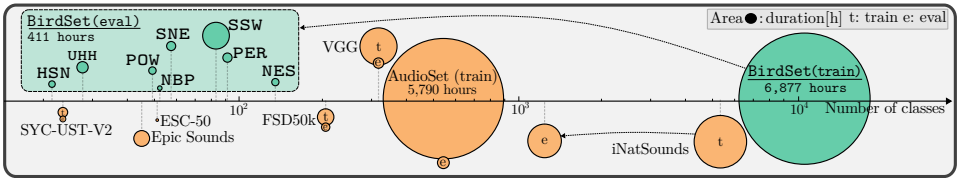
Motivation

- **General audio classification is challenged** by noise, label imbalance, domain shifts, and overlapping sources. **AudioSet has its limitations**.
- **Avian bioacoustics is a rich testbed:** it combines complex, multi-label signals with diverse acoustic environments and recording conditions.



Motivation

- **General audio classification is challenged** by noise, label imbalance, domain shifts, and overlapping sources. **AudioSet has its limitations**.
- **Avian bioacoustics is a rich testbed:** it combines complex, multi-label signals with diverse acoustic environments and recording conditions.
- **No large-scale, curated benchmark exists** for avian bioacoustics, limiting progress and comparability in the field.



Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.

Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.
2. `BirdSet` **enables a wide range of tasks in audio classification**, such as classification under domain shift and label noise.

Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.
2. `BirdSet` **enables a wide range of tasks in audio classification**, such as classification under domain shift and label noise.
3. The dataset includes **large-scale training data and carefully curated test datasets** across multiple real-world recording sites.

Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.
2. `BirdSet` **enables a wide range of tasks in audio classification**, such as classification under domain shift and label noise.
3. The dataset includes **large-scale training data and carefully curated test datasets** across multiple real-world recording sites.
4. We identify **key challenges in bird sound classification** and define structured evaluation use cases.

Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.
2. `BirdSet` **enables a wide range of tasks in audio classification**, such as classification under domain shift and label noise.
3. The dataset includes **large-scale training data and carefully curated test datasets** across multiple real-world recording sites.
4. We identify **key challenges in bird sound classification** and define structured evaluation use cases.
5. We **benchmark several deep learning models** across different scenarios.

Contributions

1. We introduce `BirdSet` on Hugging Face, **a large-scale dataset** with over half a million global bird recordings and diverse evaluation sets.
2. `BirdSet` **enables a wide range of tasks in audio classification**, such as classification under domain shift and label noise.
3. The dataset includes **large-scale training data and carefully curated test datasets** across multiple real-world recording sites.
4. We identify **key challenges in bird sound classification** and define structured evaluation use cases.
5. We **benchmark several deep learning models** across different scenarios.
6. A s**tandardized codebase ensures reproducibility and lowers the entry barrier** for new researchers.

Challenge 1: Datasets

- Focal recordings (directed mic): good for training, weakly labeled.

Challenge 1: Datasets

- Focal recordings (directed mic): good for training, weakly labeled.
- Soundscapes (omni mic): realistic, noisy, strongly labeled — ideal for evaluation.

Challenge 1: Datasets

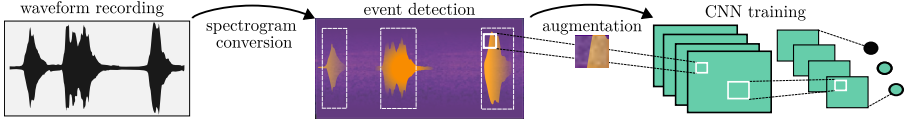
- Focal recordings (directed mic): good for training, weakly labeled.
- Soundscapes (omni mic): realistic, noisy, strongly labeled — ideal for evaluation.
- Research often mixes both or lacks structure, limiting comparability.

Challenge 1: Datasets

- Focal recordings (directed mic): good for training, weakly labeled.
- Soundscapes (omni mic): realistic, noisy, strongly labeled — ideal for evaluation.
- Research often mixes both or lacks structure, limiting comparability.
- **BirdSet**: large focal train set and diverse soundscape test sets.

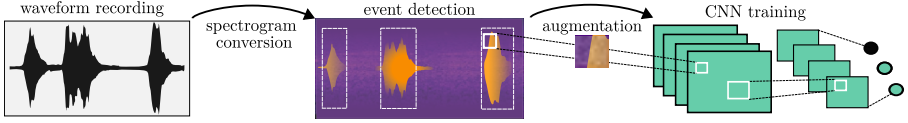
Challenge 2: Model Training

- Tasks vary: multi-class (focals) vs. multi-label (soundscapes).



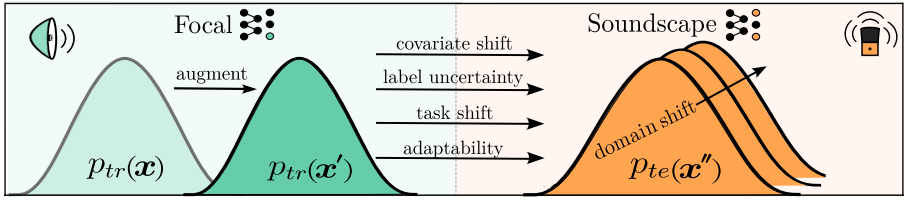
Challenge 2: Model Training

- Tasks vary: multi-class (focals) vs. multi-label (soundscapes).
- **BirdSet** provides a large scale multi-label benchmark.



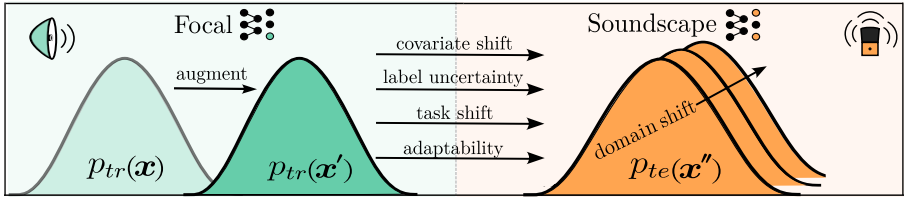
Challenge 3: Model Robustness

- Real-world bird audio varies by location, environment, and device.



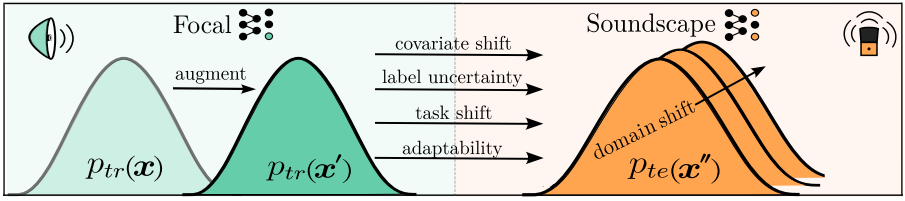
Challenge 3: Model Robustness

- Real-world bird audio varies by location, environment, and device.
- Four key obstacles: domain shift, label noise, task shift, adaptability.



Challenge 3: Model Robustness

- Real-world bird audio varies by location, environment, and device.
- Four key obstacles: domain shift, label noise, task shift, adaptability.
- **BirdSet** enables evaluation under domain/task shift and noisy labels.



BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.
  - **Dedicated training subsets:** 8 dedicated training subsets for each test sets.

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.
  - **Dedicated training subsets:** 8 dedicated training subsets for each test sets.
- **Test data (soundscapes):**

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.
  - **Dedicated training subsets:** 8 dedicated training subsets for each test sets.
- **Test data (soundscapes):**
  - 8 strongly labeled passive acoustic monitoring recordings from diverse regions.

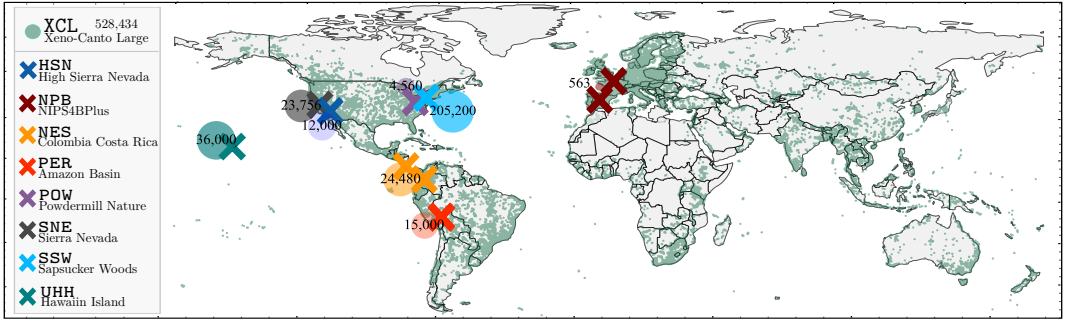
BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.
  - **Dedicated training subsets:** 8 dedicated training subsets for each test sets.
- **Test data (soundscapes):**
  - 8 strongly labeled passive acoustic monitoring recordings from diverse regions.
  - Designed for evaluation across different acoustic and species distributions.

BirdSet: Dataset Collection and Design

- **Curated, large-scale dataset** for multi-label classification of bird vocalizations.
- **Training data (focal recordings):**
  - **Xeno-Canto Large (XCL):** 530k recordings from 10,000 species.
  - **Dedicated training subsets:** 8 dedicated training subsets for each test sets.
- **Test data (soundscapes):**
  - 8 strongly labeled passive acoustic monitoring recordings from diverse regions.
  - Designed for evaluation across different acoustic and species distributions.
- **Fully public**, integrated via Hugging Face for easy access.

BirdSet: Overview



A Multi-Label Audio Classification Benchmark

- **Goal:** Benchmark models on real-world audio classification using **multi-label classification** of bird vocalizations.
- **Key challenges in the benchmark:**
  - Covariate shift (different train/test distributions)
  - Label uncertainty (weak vs. strong labels)
  - Task shift (multi-class $\rightarrow$ multi-label)
  - Class imbalance and subpopulation shift
- **Training scenarios:**
  - LT: Large-scale training on XCL
  - MT: Medium-scale training on XCM
  - DT: Dedicated fine-tuning on test-specific species

Conclusion & Takeaways

- BirdSet provides a **challenging, domain-specific dataset collection** to evaluate generalization in real-world audio settings.
- BirdSet introduces **evaluation use cases** for audio:
  - Multi-label classification under covariate shift, task shift, covariate shift
  - Event Detection
  - Self-supervised learning
  - Few-shot learning
- Benchmarked **supervised multi-label classification** with various models.
- **Next steps:** expand with different use cases and deeper analysis of robustness-related model behavior.