SELF-SUPERVISED LEARNING FOR AUTOMATED SPECIES DETECTION FROM PASSIVELY RECORDED SOUNDSCAPES

Toward Self-Supervised Avian Diversity Monitoring

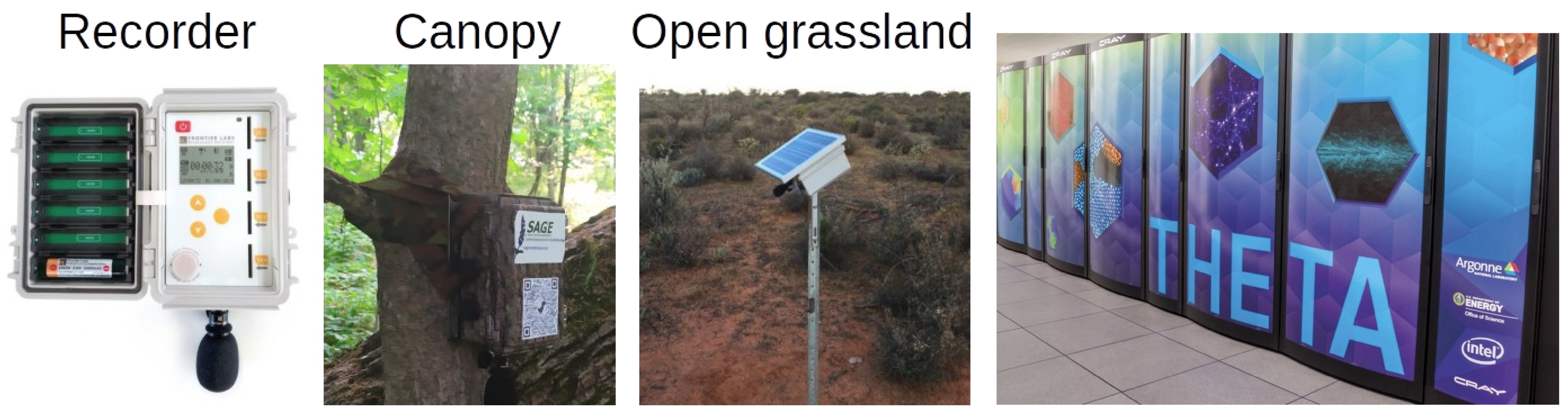
D. Dematties*, B. A. Raut, R. Sankaran and N. J. Ferrier
Argonne National Laboratory. *ddematties@anl.gov

ABSTRACT

- **Monitoring biodiversity** is more important than ever in human history, but it is expensive, not reliable neither scalable.
- Machine Learning could bring a solution but it still requires big data sets all labeled by human experts [4,5].
- Here, we propose **Self-Supervised learning** for studying biodiversity through Avian Diversity Monitoring.
- **High Performance Computing** resources are demanded since we need to process great volumes of data on computationally demanding models.
- We use a joint embedding configuration [2] to acquire features from ∼190 hours of audio.
- These algorithms are able to **discern different birdsongs**, as well as identify noise in various spectrograms at scale **with limited human intervention**.

DATA COLLECTION AND COMPUTING SETUP

- ∼2.3TB of PCM, 16 bit, mono 44100 Hz audio data collected by using nine recorders (2 open grassland and 7 canopy in the left figure).
- Recordings were conducted at Morton Arboretum, located in Illinois, US. from May 24th to September 1st 2021.
- We utilized a ThetaGPU DGX A100 node with eight NVIDIA A100 Tensor Core GPUs. We trained a Vision Transformer (ViT) [3] on ∼230k images during 50 epochs.

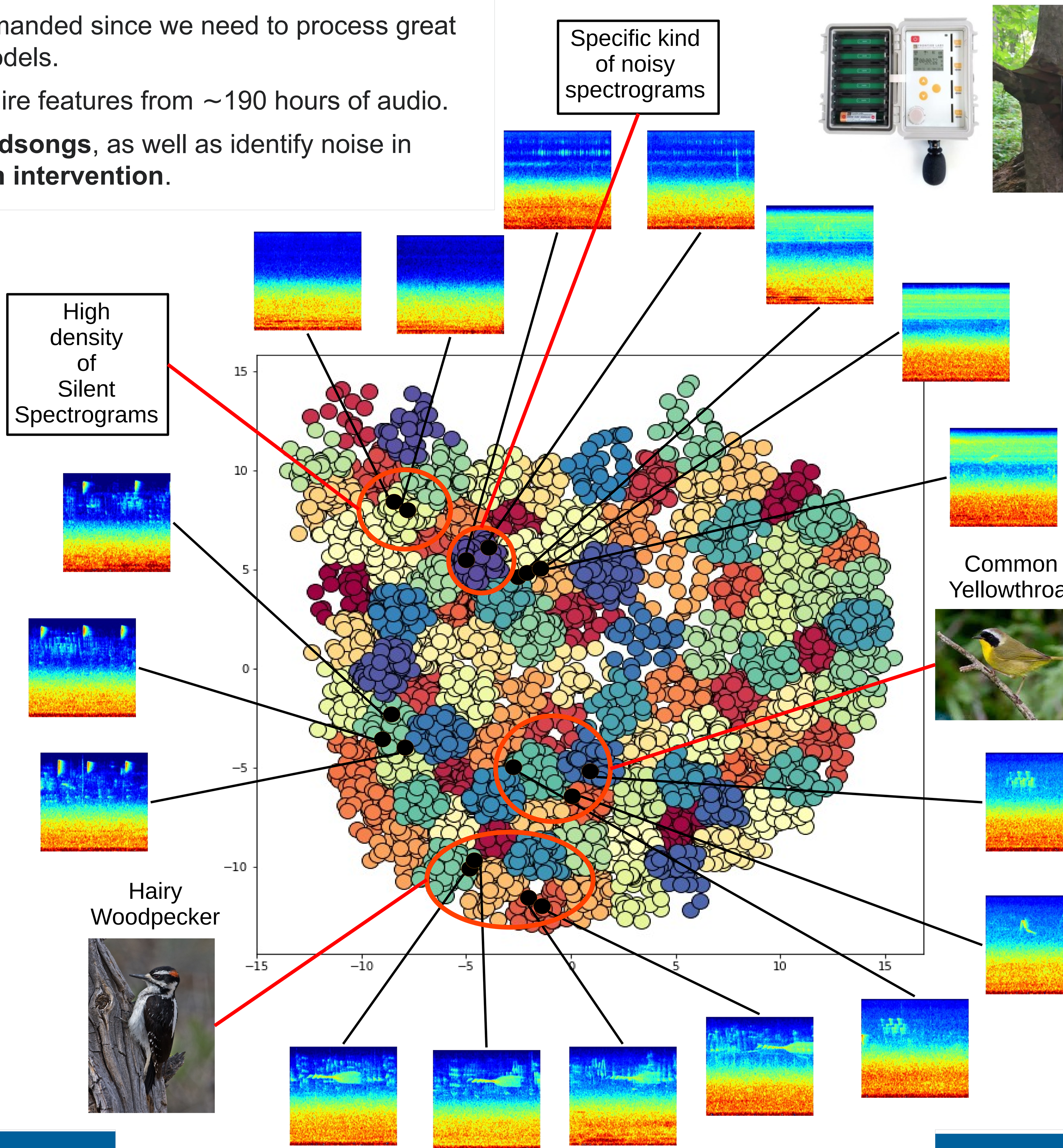


METHODS

- We applied a Joint Embedding architecture with a ViT backbone [2].
- The cloud of points from the Embeddings [2] is obtained by running inference with a trained model.
- A KMEANS clusterization with 100 clusters is conducted on a random sub-sample of 50k images.
- The dimensionality in the central figure has been reduced to 2 for visualization reasons. When using all the dimensions (384), birds can be separated with more accuracy.

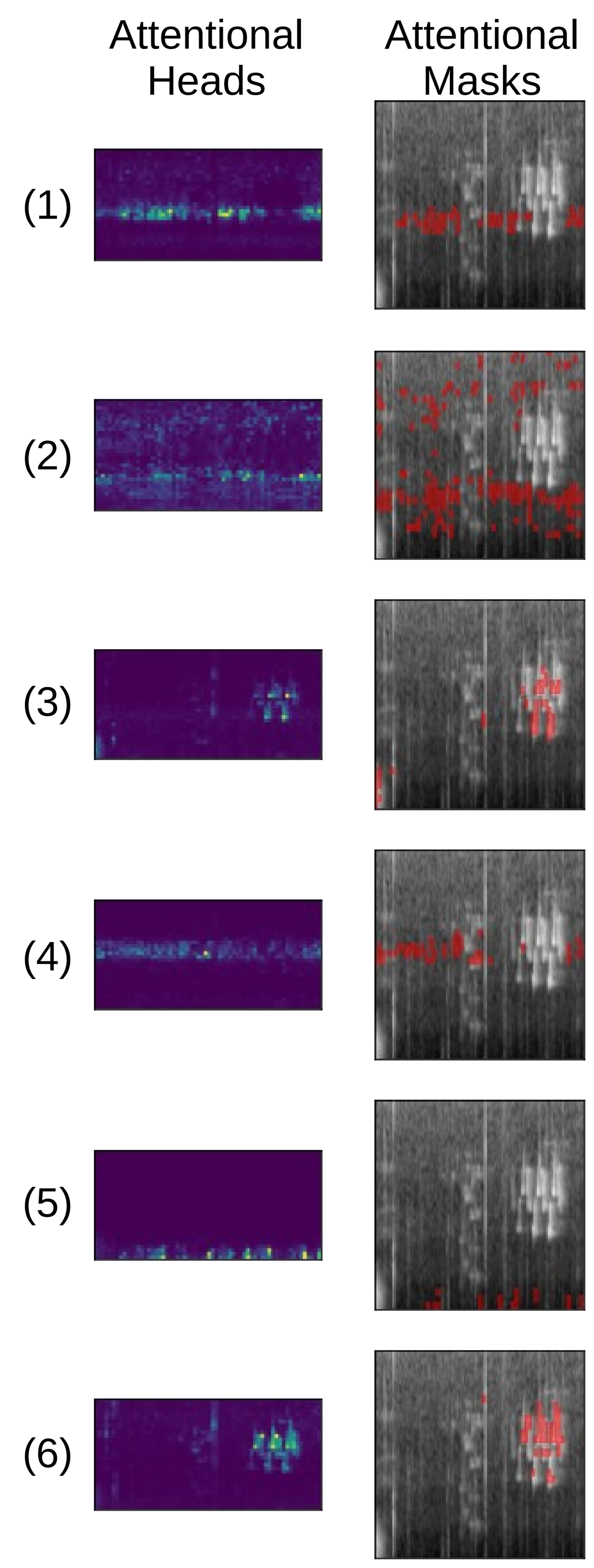
RESULTS

- It distributes the spectrograms in a multidimensional output space.
- Similar acoustic conditions are mapped to neighboring regions in the output space.
- For instance, some clusters contain silent audio spectrograms with similar noise conditions.
- Remarkably, there are some clusters populated with spectrograms containing similar bird songs.



RESULTS

- Different heads from the transformer highlight different semantic aspects in the Spectrograms.
- Third and sixth heads are mostly highlighting a birdsong inside the Spectrogram.
- On the right we find the unsupervised segmentation produced from the attentional maps on the Spectrograms.
- We were able to pinpoint the birdsong from the background noise using the Attentional Masks.



CONCLUSIONS

- **Self-Supervised learning is well suited for analyzing audio data,** making the labeling process for biologists and ornithologists tractable and scalable.
- Ornithologists—for instance—could easily **tease apart millions of spectrograms with different noise characteristics,** detecting anomalies as outliers in the output space.
- Ideally experts can **detect different birdsongs** in different regions of the output space, and label millions of examples per hour.

NEXT STEPS

- Process all the 2.3TB of audio and also test different ViTs sizes as well as different joint embedding strategies such as Variance Invariance Covariance Regularization [1].
- Compare the clusterization performance of several output spaces utilizing pseudo labels generated by trained supervised algorithms such as BirdNet [4].
- Use labeled datasets as a mean of validation of the clustering performance as well as testing the segmentation produced from attentional maps in order to detect informative features from background noise in spectrograms.

REFERENCES

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv´e J´egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[4] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61:101236, 2021.

[5] Yuren Sun, Tatiana Midori Maeda, Claudia Solis-Lemus, Daniel Pimentel Alarcon, and Zuzana Burivalova. Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks, 2021.

Dataset: https://zenodo.org/record/6964885

Code: https://zenodo.org/record/6964966

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY