TaxaBind: Supplementary Material

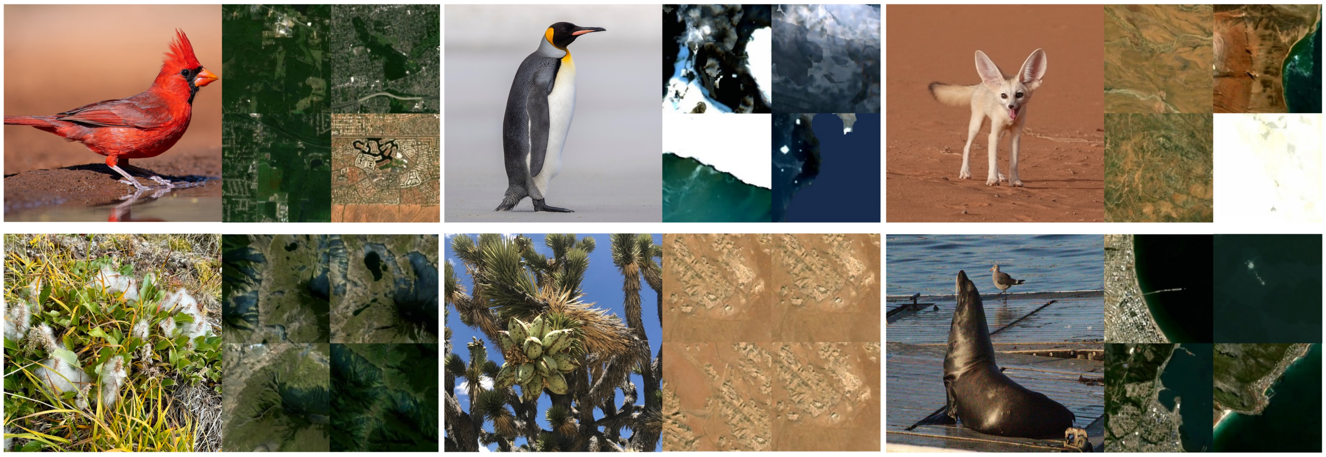


Figure 1. Additional results for species image to satellite image retrieval task. For each example, we show the top 4 most similar satellite images retrieved by our model from a gallery of 100k satellite images in the iSatNat-test set.

A. Retrieval Results

Here we provide additional cross-modal retrieval results of TaxaBind on our TaxaBench-8k dataset (Table 1). It is observed that the retrieval performance improves when embeddings from different modalities are added together. We also present six examples of ground-level image-to-satellite image retrieval (Figure 1). For each example, we present the top-4 most similar satellite images retrieved by our model. We also present range map predicted by our model for *Abies balsamea* in Figure 3.

B. Datasets

Here we provide additional details about the datasets used for training and evaluating our models.

B.1. Training Datasets

iSatNat. This dataset was built by collecting Sentinel-2 Level 2A imagery corresponding to each geolocation in the iNaturalist-2021 dataset. We used the training and validation splits of the dataset and dropped the testing split since it lacked the ground-truth labels. We used the validation split as the unseen testing set. We created a 90:10 split of the training dataset to create the final training and validation sets. Lastly, we applied a minimal filter to remove all the samples lacking geolocation entry.

iSoundNat. Using the iNaturalist platform, we filtered all the observations with audio recordings and ground-level images to date. We then removed all corrupted audio recordings and converted them to a common format (m4a). This resulted in a total of 88,130 observations. We then used a stratified sampling technique to split the dataset into 85:5:10 (train, validation, test) ratio. The spatial distribution of the dataset is shown in Figure 2.

WorldClim 2.1. This dataset consists of 19 bioclimatic variables and an additional elevation map. All the channels are scaled to a resolution of 5 arc minutes.

C. Evaluation Datasets

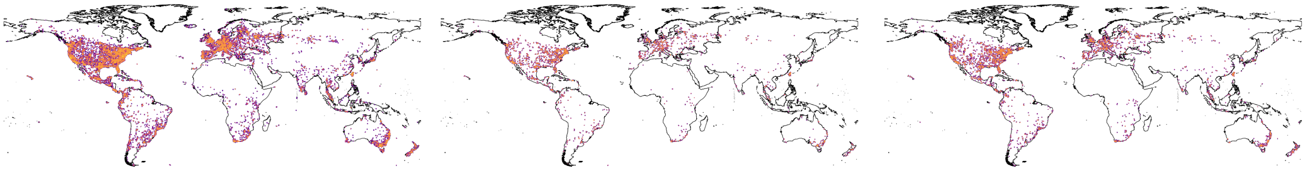
TaxaBench-8k. We extended the testing split of iSoundNat by downloading Sentinel-2 Level 2A imagery corresponding to each location in the split. This dataset is used for evaluating our models on zero-shot image classification and cross-modal retrieval. The spatial distribution of the dataset is shown in Figure 2c.

Birds525 [1]. This dataset consists of images of bird species across 525 categories. Each image features a single bird species. To evaluate our models, we use the testing split of the dataset which consists of 2,625 samples.

CUB-200-2011 [2]. This dataset consists of images of 200 bird species. We use the testing split of the dataset which contains 5,794 images.

Method	Modality	R@1	R@5	R@10
Random Baseline	-	0.01	0.05	0.11
TaxaBind	🛰️ → 🐨	1.87	7.23	12.02
	🐨 → 🛰️	1.34	5.42	9.26
	🔊 → 🐨	1.00	4.10	7.75
	🐨 + 💬 → 🛰️	2.03	8.16	13.66
	🐨 + 💬 → 🛰️ + 📍	2.39	8.80	14.74
	🌨️ → 🛰️	2.22	11.26	19.40
	🌨️ → 🛰️ + 📍	3.25	14.18	23.52

Table 1. Additional retrieval results on our TaxaBench-8k dataset.



(a) Train

(b) Validation

(c) Taxabench-8k

Figure 2. Geographic locations of samples in the training, validation and testing splits of iSoundNat.

BioCLIP-Rare [3]. This dataset was used for the evaluation of BioCLIP. It contains 400 species categories not present in the TreeOfLife-10M dataset.

BirdCLEF-2022 [4]. This dataset contains audio recordings of rare bird species in Hawaii. We use the training split of the dataset which contains annotations. In total, there are 14,852 audio recordings across 141 categories. We use stratified sampling to split the dataset into 85:5:10 (train, validation, test) ratios.

BirdCLEF-2023 [5]. This dataset contains audio recordings of bird species in Kenya. We use the training split of the dataset which contains annotations. In total, there are 16,941 audio recordings across 264 categories. We use stratified sampling to split the dataset into 85:5:10 (train, validation, test) ratios.

BirdCLEF-2024 [6]. This dataset contains audio recordings of bird species in Western Ghats, India. We use the training split of the dataset which contains annotations. In total, there are 24,459 audio recordings across 182 categories. We use stratified sampling to split the dataset into 85:5:10 (train, validation, test) ratios.

Ecoregions. We use the ecoregion map from [7]. The map consists of 846 distinct categories of ecoregions. We randomly sample 100k points around the globe. We split the points into 85:5:10 (train, validation, test) ratio. Then we perform linear probing on our model and report top-1 classification accuracy on the testing split.

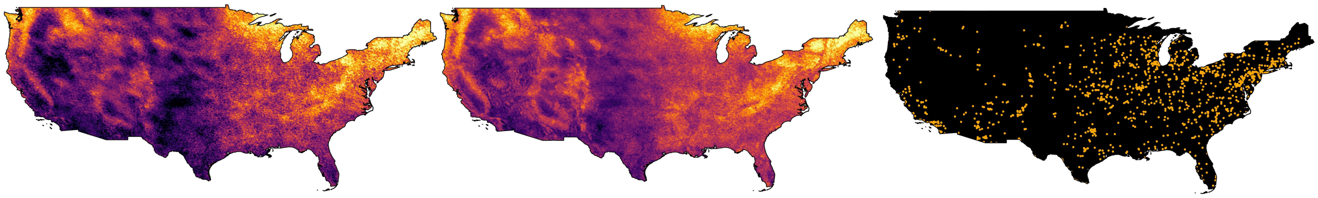
Biome. We use the biome map from [7]. The map consists of 14 distinct categories of biomes. We randomly sample 100k points around the globe. We split the points into 85:5:10 (train, validation, test) ratio. Then we perform linear probing on our model and report top-1 classification accuracy on the testing split.

GeoPlant [8]. This dataset consists of the presence-absence of plant species across different countries in Europe. The dataset additionally contains presence-only observations from GBIF. We use only the presence-absence split which contains 88,783 unique locations. We use stratified sampling using country labels to split the set into 85:5:10 (train, validation, test) ratios. For each location, the dataset contains the presence of multiple species. We use SatBird's [9] training procedure and predict the presence-absence of species at each of the locations.

SatBird [9]. This dataset contains the presence-absence checklist of bird species in two regions: the USA and Kenya. The dataset for the USA is further divided into two seasons: summer and winter. Each location in the dataset is associated with a multispectral Sentinel-2 satellite image.

D. Ethics and Limitations

The models built are a proof-of-concept for demonstrating the benefits of combining multiple modalities for solving ecological problems. Care must be taken when utilizing our models for real-life applications. Additional validation may be necessary before deploying our models. We recognize that the datasets used for training and evaluation may have some spatial bias. However, the goal of this work is not to specifically tackle the issue of spatial bias, but rather



(a) 🐦→📍 (b) 🐦→📍+🛰 (c) GBIF presence-only

Figure 3. Range map of *Abies balsamea* using a query ground-level image and combination of various modalities across the USA.

to utilize and understand patterns in multiple modalities. We observe that incorporating additional modalities into the framework helps to implicitly account for this bias in the data.

References

[1] G. Piosenka, "Birds 525 species- image classification," Apr 2023. 1

[2] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016. 1

[3] 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*, pp. 19412–19424, 2024. 2

[4] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings.," in *CLEF (Working Notes)*, pp. 1929–1939, 2022. 2

[5] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2023: Automated bird species identification in eastern africa.," in *CLEF (Working Notes)*, pp. 1934–1942, 2023. 2

[6] H. Klinck, Maggie, S. Dane, S. Kahl, T. Denton, and V. Ramesh, "Birdclef 2024," 2024. 2

[7] E. Dinerstein, D. Olson, A. Joshi, C. Vynne, N. D. Burgess, E. Wikramanayake, N. Hahn, S. Palminteri, P. Hedao, R. Noss, *et al.*, "An ecoregion-based approach to protecting half the terrestrial realm," *BioScience*, vol. 67, no. 6, pp. 534–545, 2017. 2

[8] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, P. Bonnet, and A. Joly, "Geoplant: Spatial plant species prediction dataset," *arXiv preprint arXiv:2408.13928*, 2024. 2

[9] M. Teng, A. Elmustafa, B. Akera, Y. Bengio, H. Radi, H. Larochelle, and D. Rolnick, "Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 2