This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TaxaBind: A Unified Embedding Space for Ecological Applications

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, Nathan Jacobs
Washington University in St. Louis
{s.sastry, k.subash, a.dhakal, aadeel, jacobsn}@wustl.edu

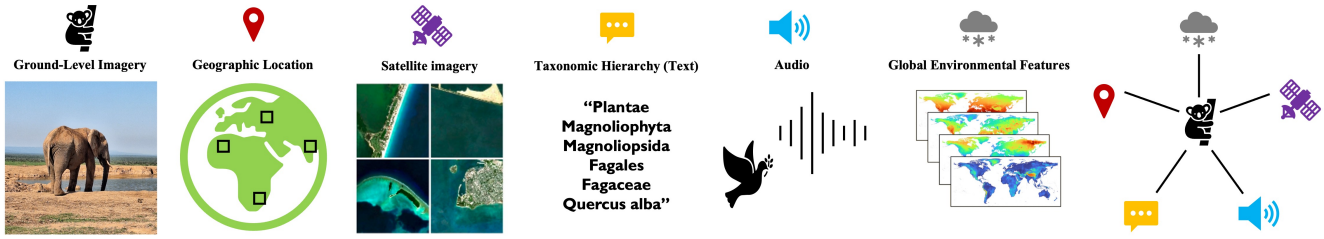


Figure 1. **TaxaBind Framework**. To create a unified embedding space consisting of different modalities, we exploit ground-level images of species as the binding modality. We use various ground-level image-paired datasets to train modality-specific encoders. Ultimately, the encoders support embedding arithmetic and exhibit emergent properties and zero-shot capabilities.

Abstract

We present TaxaBind, a unified embedding space for characterizing any species of interest. TaxaBind is a multimodal embedding space across six modalities: ground-level images of species, geographic location, satellite image, text, audio, and environmental features, useful for solving ecological problems. To learn this joint embedding space, we leverage ground-level images of species as a binding modality. We propose multimodal patching, a technique for effectively distilling the knowledge from various modalities into the binding modality. We construct two large datasets for pretraining: iSatNat with species images and satellite images, and iSoundNat with species images and audio. Additionally, we introduce TaxaBench-8k, a diverse multimodal dataset with six paired modalities for evaluating deep learning models on ecological tasks. Experiments with TaxaBind demonstrate its strong zero-shot and emergent capabilities on a range of tasks including species classification, cross-model retrieval, and audio classification. The datasets and models are made available at https://github.com/mvrl/TaxaBind.

1. Introduction

Fine-grained species classification is a challenging task in computer vision, which is often necessary for ecologists to automatically label images of rare species. A related and arguably a more important task is species distribution mapping which aims to map the presence of a given species of interest. Until now, both tasks were addressed using separate frameworks and methodologies, often requiring different datasets. In this work, we propose learning a unified embedding space over six modalities that is useful for several downstream ecological tasks including but not limited to species distribution mapping, fine-grained classification, and audio classification.

The presence of a particular species at a given geographic location can reveal several important characteristics of that species. Previous studies attempted to implicitly learn the relationship between geographic location and the presence of species by considering either environmental features [1] or satellite images [2–4] describing the location. This leads to learning an effective representation of any geographic location which is useful for species distribution mapping. However, this type of modeling often overlooks important species attributes, such as their taxonomic hierarchy or audio signatures.

Recent works such as BioCLIP [5] and ArborCLIP [6] have demonstrated impressive zero-shot species classification capabilities. However, these frameworks are restricted to image and text modalities, ignoring crucial geographic, audio, and habitat characteristics of species. Multimodal embedding frameworks like ImageBind [7] and GRAFT [8] have shown that it is possible to learn a joint representation space by aligning all available modalities to the ground-level image modality. This allows for training modality-specific encoders using only image-paired datasets. One

potential downside of such methods is that they perform locked tuning with the ground-level image modality. This means that the ground-level image encoder is kept frozen, while the other modalities are trained to project to the existing learned space of the ground-level image modality. This can lead to sub-optimal performance since task-specific unique information of each modality is lost [9].

To this end, we propose multimodal patching, building upon patching [10], a framework to distill knowledge from various modalities while still preserving the original embedding space of the binding modality. We show that multimodal patching can improve zero-shot classification performance of the binding modality. We create a joint embedding space containing six modalities (Figure 1). To facilitate future research and evaluation of ecological models, we present TaxaBench-8k, a truly multimodal dataset containing *six paired modalities*. The contributions of our work are fivefold:

1. *Multimodal Patching*. We propose a simple yet effective patching technique that improves over the ImageBind framework.

2. *Multimodal Models for Ecological Applications*. We propose modality-specific encoders that can handle various ecological tasks over six modalities: ground-level image, geographic location, satellite image, text, audio, and environmental features.

3. *Multimodal datasets*. We compiled two large-scale novel cross-view datasets: i) iSoundNat: ground-level images of species with their corresponding audio; and ii) iSatNat: ground-level images of species with their corresponding satellite imagery.

4. *TaxaBench-8k*. We present TaxaBench-8k, a benchmarking dataset containing six paired modalities for evaluating multimodal ecological models.

5. We demonstrate our models' effectiveness and emergent properties on several benchmarking and zero-shot tasks.

2. Related Work

2.1. Multimodal Self-Supervised Learning

Multimodal self-supervised methods using contrastive learning (CL) objectives have shown impressive results across diverse tasks. These methods encompass advancements in CL frameworks, integration of multiple modalities into unified embedding spaces, and innovations in training strategies to further enhance model performance. Out of many notable works in this area, CLIP [11] utilizes symmetric InfoNCE loss [12]; SupCon [13] utilizes label information in its contrastive objective; SigLIP [14] employs binary classification loss. Frameworks such as ImageBind [7], Sat2Cap [15], GRAFT [8], GeoCLAP [16], and GeoBind [17] demonstrate the integration of additional modalities such as audio, satellite imagery, or metadata into CL-trained multimodal embedding spaces. The core idea of these strategies involves utilizing a pretrained image-text embedding space and learning to project all other modalities into this space. However, this simple yet effective strategy can suffer from information collapse [9]. Recent innovations in training strategies include LiT [18], which enhances zero-shot performance by freezing the vision encoder while training the text encoder; OmniVec [19] and OmniVec2 [20], which improve performance through modality-specific encoders and shared backbones for multimodal multitask learning; factorized contrastive learning [9] which focuses on preserving unique modality-specific information; and Patching [10], which boosts performance by interpolating weights between pre-trained and fine-tuned models. In this work, we generalize the concept of patching to more than two modalities. We call this multimodal patching which improves over the training strategy of ImageBind.

2.2. Multimodal Learning for Ecology

Multimodal learning for ecological applications, such as species distribution modeling (SDM) and fine-grained visual classification (FGVC) of species, has recently advanced significantly. These advancements are driven by the availability of large-scale multimodal datasets [5, 6, 21, 22] and novel multimodal learning frameworks [2, 4, 23, 24]. Methods like BioCLIP [5] and ArborCLIP [6] have demonstrated the utility of combining images of species with their corresponding taxonomic text descriptions. BioCLIP is a foundational model for the tree of life, trained with a CLIP-like contrastive loss between image representations of different species and taxonomic descriptions. ArborCLIP is a recent model contrastively trained on a large multimodal dataset containing over 134 million images of diverse species paired with their detailed taxonomic descriptions. Although such CLIP-based models have shown state-of-the-art performance in zero-shot FGVC, they are limited by the number of modalities they can consume and the tasks they can solve. On the other hand, the multimodal fusion of remote sensing data has resulted in unprecedented performance in SDM under presence-only [1, 4, 24, 25] and presence-absence [3, 26] settings. Most of these methods are general-purpose ecological predictors that effectively learn multimodal features that can vary across space. In these frameworks, geolocations are represented either by learning an implicit neural function [1, 23, 24] or by representations derived from satellite imagery [3, 4, 26]. Other CL-based methods such as SatClip [27] and GeoCLIP [28]

aim to learn general-purpose location representations which can then eventually be utilized for downstream ecological tasks. In our work, we combine these parallel lines of work into a single general-purpose framework, TaxaBind, that can consume multiple modalities and solve numerous ecology-related tasks.

3. Dataset

In this work, we train and evaluate our models using various large-scale multimodal datasets. We built three datasets to advance multimodal learning in the field. Here we provide the details about the datasets.

Training datasets. When it comes to applying deep learning in ecology, there is a lack of high-quality multimodal datasets, especially the ones with paired ground-level images of species. To bridge the gap, we constructed two large-scale datasets: `iSatNat` and `iSoundNat` (Table 1). iSatNat consists of pairs of ground-level and satellite imagery while iSoundNat contains pairs of ground-level images and audio recordings of species. We begin with the iNat-2021 dataset [21] which contains 2.7M images of species alongwith metadata including geolocation information. We built iSatNat by collecting **2.7M** Sentinel-2 cloudless [29] imagery corresponding to each ground-level image of species in the iNat-2021 dataset. iSoundNat was built by collecting ground-level images of species from the iNaturalist platform which contained audio recordings. We specifically downloaded *research grade* observations from the platform. This resulted in a total of **88,130** pairs of images and audio. For more details about the datasets, please refer to the appendix. We leverage BioCLIP's TreeofLife-10M [5] dataset for image-text pretraining. To simplify our experimentation, we use pretrained BioCLIP vision and text encoders. For training the location encoder, we use the geolocation information present in the iNat-2021 dataset corresponding to each image. We use the environmental variables from WorldClim-2.1 for training the environmental encoder. To do this, we extract the bioclimatic variables corresponding to the geolocations present in the iNat-2021 dataset.

TaxaBench-8k. For evaluation and further research, we constructed a truly multimodal dataset containing six paired modalities: ground-level image, geographic location, satellite image, text, audio, and environmental features. We begin with the test split of our iSoundNat dataset which contains **8,813** image and audio pairs. For each sample, we downloaded Sentinel-2 level 2A imagery corresponding to the geolocation information present with that sample. Similarly, we extracted the environmental features from WorldClim-2.1 corresponding to the geolocations of the samples. Our TaxaBench-8k dataset is multifaceted and can be used to evaluate ecological models for various tasks such as cross-modal retrieval, species classification, and audio classification.

iSoundNat	Train	Val	Test
#samples	74,910	4,407	8,813
#species	6,925	1,482	2,225
iSatNat	Train	Val	Test
#samples	2.55M	134k	100k
#species	10k	10k	10k

Table 1. The number of samples and unique species categories in our datasets.

Evaluation datasets. We evaluate our models on a range of downstream ecological tasks. We assess the effectiveness of each encoder through these downstream tasks. The details of the datasets are mentioned in Section 5 with additional details present in the appendix.

4. Method

We aim to learn a unified embedding space that can uniquely characterize a given taxon. This is done by aligning all available modalities to the ground-level images of species. For this, we utilize the BioCLIP [5] embedding space as a teacher and learn to project the embeddings from all the modalities to this space. We propose multimodal patching to further distill task-specific unique knowledge into the modalitiy-specific encoders. This enables embedding arithmetic which is useful for zero-shot classification and cross-modal retrieval tasks. Further, the modalities exhibit emergent properties with each other when only trained with corresponding ground-level image pairs. Below we discuss the contrastive learning approach used for training, multimodal patching, and the implementation details of our framework.

4.1. Contrastive Learning

InfoNCE [12] is a widely used loss function for learning an embedding space with similar and dissimilar examples within a single mini-batch. This loss function works by treating paired examples as positives while considering all other pairs as negatives. These pairs can be constructed using augmentation techniques or can originate from different modalities, such as images and text. However, for our problem, we note that a single mini-batch may contain multiple instances of the same species category. Considering examples from the same species category as negative may lead to sub-optimal performance. Hence, we utilize the concept of CLIP loss [11] and combine it with SupCon loss [13] as the species category information is available for each ground-level image. Consider the pair of modalities

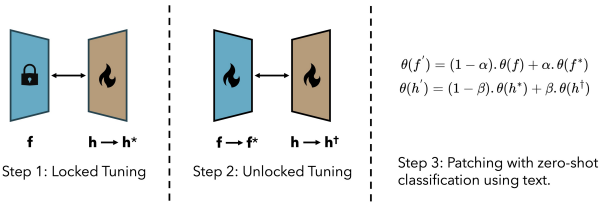


Figure 2. **Multimodal Patching**. For distilling unique information from different modalities, we patch the encoders using zero-shot classification with text. Note that since the network f is shared across all modalities, it is patched using techniques like sequential patching or parallel patching.

($\mathcal{G}$, $\mathcal{M}$), where $\mathcal{G}$ denotes the ground-level image modality while $\mathcal{M}$ denotes some other modality (e.g. satellite imagery). Consider a dataset with aligned observations of the two modalities and a mini-batch of examples during training $\{g_i, m_i\}_{i=1,...N}$. Using deep networks f and h, one can obtain normalized embeddings of the respective modalities as $z_i = f(g_i)$ and $y_i = h(m_i)$. Let $N(i)$ denote the set of all plausible indices in a mini-batch and $P(i)$ denote the set of indices of examples with the same species category as that of the i^{th} example. Then, the deep networks are optimized using the following combined loss:

$$\mathcal{L}_{\mathcal{G}\to\mathcal{M}} = \sum_{i\in N(i)} \frac{-1}{|P(i)|} \sum_{j\in P(i)} \log \frac{\exp(z_i \cdot y_j/\tau)}{\sum_{n\in N(i)} \exp(z_i \cdot y_n/\tau)}$$

$$\mathcal{L}_{\mathcal{M}\to\mathcal{G}} = \sum_{i\in N(i)} \frac{-1}{|P(i)|} \sum_{j\in P(i)} \log \frac{\exp(y_i \cdot z_j/\tau)}{\sum_{n\in N(i)} \exp(y_i \cdot z_n/\tau)}$$

$$\mathcal{L}_{\mathcal{G},\mathcal{M}} = \frac{\mathcal{L}_{\mathcal{G}\to\mathcal{M}} + \mathcal{L}_{\mathcal{M}\to\mathcal{G}}}{2} \quad (1)$$

where τ is a scalar temperature parameter that controls the sensitivity of the predicted softmax distribution. The loss function seeks to cluster positive examples from the modalities in the embedding space while pushing negative examples farther away.

4.2. Multimodal Patching

We describe multimodal patching as shown in Figure 2. The overall framework consists of three steps. The pseudo-code for multimodal patching is presented in Algorithm 1. Let f be a pretrained binding modality encoder and h be an encoder of a different modality. First, we perform locked tuning of h to obtain h^*. In this step, f is utilized as a teacher, and embeddings obtained from h are learned to project to the existing embedding space of f. In the second step, we perform full finetuning of f and h to obtain f^* and $h^\dagger$ respectively. The encoders f^* and $h^\dagger$ no longer preserve the original embedding space of f. Finally, in the third step, we perform patching [10] of h by linearly interpolating the weights between h^* and $h^\dagger$. The interpolation

Algorithm 1 Multimodal Patching

Require: Binding Modality $\mathcal{G}$, Set of Modalities $\mathcal{M}$, Binding Modality Encoder f, Set of Encoders of $\mathcal{M}$ - $\{h_i\}_{i=1,...|\mathcal{M}|}$, Patching Task $\mathcal{P}$

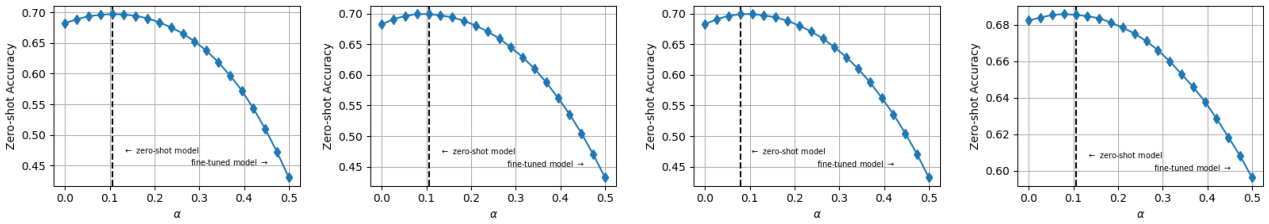
$\theta(z) \leftarrow \theta(f)$
for each h_i in $\{h_i\}_{i=1,...|\mathcal{M}|}$ **do**
  $h_i^* \leftarrow$ locked_tuning(f, h_i)
  $f^*, h_i^\dagger \leftarrow$ unlocked_tuning(f, h_i)
  $\alpha \leftarrow \arg\max_\alpha \mathcal{P}[(1-\alpha).\theta(z) + \alpha.\theta(f^*)]$
  $\beta \leftarrow \arg\max_\beta \mathcal{P}[(1-\beta).\theta(h_i^*) + \beta.\theta(h_i^\dagger)]$
  $\theta(z) \leftarrow (1-\alpha).\theta(z) + \alpha.\theta(f^*)$
  $\theta(h_i') \leftarrow (1-\beta).\theta(h_i^*) + \beta.\theta(h_i^\dagger)$
end for
$\theta(f') \leftarrow \theta(z)$
return $f', \{h_i'\}_{i=1,...|\mathcal{M}|}$

weights are determined by analyzing the performance of the patched models on a patching task $\mathcal{P}$. We utilize zero-shot classification with text as the patching task. This patching task helps to preserve the original embedding space of f and enables emergent capabilities. These steps are repeated for all available modalities. Since f is shared across the modalities, we perform sequential patching of f across all the modalities as described in Algorithm 1. Our strategy modifies the weights of f while preserving the original embedding space, unlike ImageBind, where f is frozen.

4.3. Implementation Details

We aim to learn an embedding space consisting of six modalities. To do this, we employ modality-specific encoders to encode each of the modalities into a joint embedding space. This setup proves to be computationally feasible and allows for precomputing embeddings. We use the transformer architecture for encoding ground-level images, text, satellite images, and audio. We use pretrained BioCLIP [5] (ViT-B/16) vision and text encoders for the ground-level images and text respectively. We use pretrained CLIP [11] (ViT-B/16) vision encoder for the satellite images and a pretrained CLAP [30] encoder for audio. We use the architecture described in GeoCLIP [28] to encode the geographic location. It consists of an Equal Earth Projection (EEP) operation, a random Fourier feature transform, and an MLP network. We use a ResNet-style MLP [1] to encode the environment features for a given geographic location.

We use a native embedding size of 512 for all the encoders as used in BioCLIP. All the encoders are trained independently using the BioCLIP vision encoder. Images are normalized and resized to (224, 224) pixels before feeding them to their respective encoders. For training the location encoder, we additionally sample pseudo-negative



(a) (b) (c) (d)

Figure 3. **Patching improves zero-shot classification performance with text**. We evaluate the zero-shot classification accuracy of the ground-level image encoder with different values of α on iNat-2021. We observe performance improvements in all the cases.

Model	Modality	Birds525	CUB-200-2011	BioCLIP-Rare	iNat-2021	TaxaBench-8k
BioCLIP [5]		82.92	77.51	34.52	68.24	32.88
ArborCLIP [6]		65.84	**82.41**	27.58	68.00	31.34
TaxaBind		**83.74**	78.22	**35.84**	**70.09**	**34.45**
ImageBind [7]	+	-	-	-	71.02	36.40
	+	-	-	-	72.62	36.30
	+	-	-	-	71.96	**36.59**
	+	-	-	-	-	35.91
TaxaBind	+	-	-	-	**72.73**	**36.59**
	+	-	-	-	**73.20**	**37.54**
	+	-	-	-	**72.02**	36.51
	+	-	-	-	-	**36.27**

Table 2. Zero-shot classification performance on various fine-grained species classification datasets using the taxonomic description of species.

locations as a way to train on locations absent in the training dataset [1]. For zero-shot classification using text, we consider the entire taxonomic description of a species. For each audio sample, we average across the channel dimension to get single-channel audio. Then we convert each single channel audio sample into mel-spectrogram features using the default CLAP settings: {`feature_size=64`, `sampling_rate=48000`, `hop_length=480`, `max_length_s=10`, `fft_window_size=1024`} in the `HuggingFace`-wrapper: `ClapProcessor` for the pre-trained CLAP model `clap-htsat-fused`. For each training run, we use 2 NVIDIA H100 GPUs, a batch size of 256, and a gradient accumulation of 8.

5. Experiments

We conduct several empirical evaluations of our modality-specific encoders against state-of-the-art methods across a range of ecology-related tasks. All details about each dataset used for evaluation are mentioned in appendix. Additionally, to evaluate the effectiveness of our proposed multimodal patching strategy, we compare it against the training recipe of ImageBind [7], where the binding modality encoder is kept frozen during the training. The ImageBind training recipe is equivalent to restricting our multimodal patching strategy to its first step. In the experiments, we show that adding the embeddings from multiple modalities leads to improved performance compared to using embeddings of a single modality. Below we discuss various experimental results.

Multimodal Patching. We first show that patching improves zero-shot image classification of the binding modality encoder. In Figure 3, we present the zero-shot image classification performance of our model when patched with varying values of α. It is seen that for certain values of α, our model exhibits improved zero-shot performance than the base pretrained model of ImageBind. This behavior is observed across all the modalities independently. For all the modalities, the optimal value of α is close to 0.1, indicating that models patched using higher values of α (close to 1) move significantly away from the embedding space of BioCLIP. We further show in Table 7 that multimodal sequential patching has an additive effect on our model. These ex-

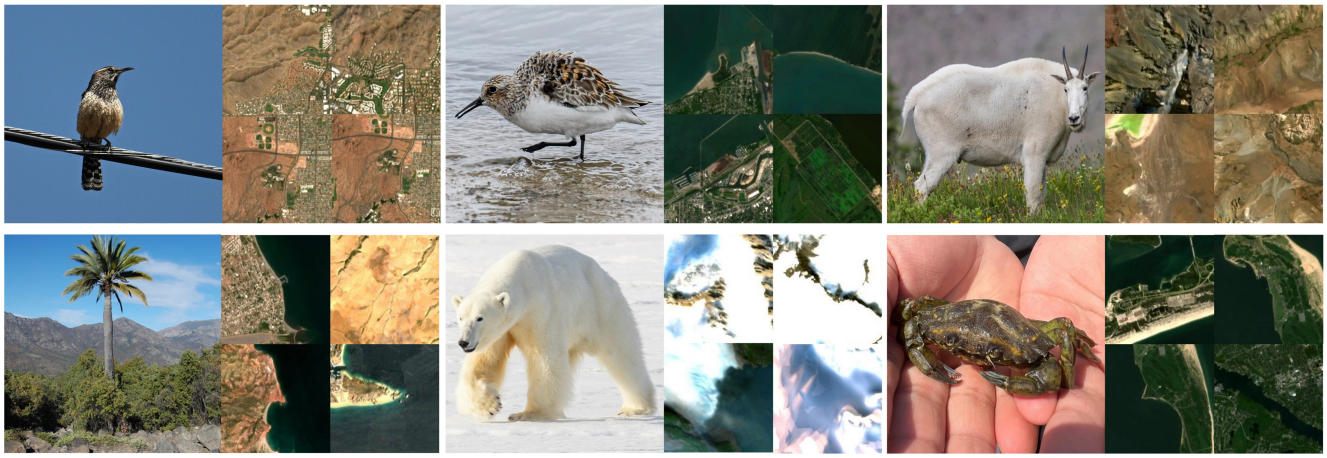


Figure 4. **Species image to satellite image retrieval task**. For each example, we show the top 4 most similar satellite images retrieved by our model from a gallery of 100k satellite images in the iSatNat-test set.

periments empirically demonstrate that ImageBind's training method fails to effectively distill knowledge from various modalities and that it is possible to improve the performance of pretrained binding modality encoder.

Zero-shot image classification. We evaluate TaxaBind's image encoder on zero-shot image classification task using the taxonomic description of species. We use BioCLIP [5] and ArborCLIP [6] as image-only baseline models. For multimodal evaluation, we compare against our ImageBind-trained baseline. In this setting, the embeddings obtained from the image encoder are added to those obtained from a different modality encoder and then used for zero-shot classification. For image-only experiments, we use Birds525 [31], CUB-200-2011 [32] and BioCLIP-Rare [5]. For multimodal experiments, we use iNat-2021 [21] and TaxaBench-8k. The results presented in Table 2 show that our model outperforms the baseline models in both settings across 4 out of 5 datasets. It is worth noting that our model always outperforms BioCLIP, which indicates the benefit of distilling multimodal information into the ground-level image encoder. Finally, the addition of ground-level and satellite image embedding leads to the best zero-shot classification performance.

Cross-modal retrieval. To demonstrate the emergent capabilities of our models, we evaluate on the task of cross-modal retrieval. We use TaxaBench-8k, which has a gallery size of 8,813 data points. In each retrieval setting, we selected the pair of modalities which were not explicitly trained together. The results, shown in Table 3, highlight the superior performance of TaxaBind compared to both a random baseline and the ImageBind framework. We further present six qualitative examples of species image to satellite image retrieval results in Figure 4. Here, we use a gallery of 100k satellite images and retrieve the top 4 most similar satellite images given a ground-level species image. The retrieved images show habitat characteristics correlated with those of the query species. This suggests that our models have learned to capture fine-grained habitat and climate-related information about the species. We present additional quantitative retrieval results in the appendix.

Method	Modality	R@1	R@5	R@10
Random Baseline	-	0.01	0.05	0.11
ImageBind [7]	sat → loc	**8.79**	**22.72**	**30.84**
	loc → sat	9.32	24.16	32.24
	sat → audio	1.94	6.68	10.56
	audio → sat	1.86	5.33	9.05
TaxaBind	sat → loc	8.43	21.72	30.42
	loc → sat	**9.62**	**24.60**	**33.42**
	sat → audio	**2.05**	**7.03**	**11.05**
	audio → sat	**2.36**	**5.96**	**9.50**

Table 3. **Emergent capabilities**. Cross-modal retrieval results on the TaxaBench-8k dataset. The gallery consists of 8,813 samples in each experiment.

Species audio classification. We assess the effectiveness of our audio encoder on the task of bird species audio classification. We used pretrained CLAP [30] and ImageBind's audio encoder as baselines and tested them on three datasets: BirdCLEF-2022, BirdCLEF-2023, and BirdCLEF-2024. These datasets are part of the LifeCLEF series [36, 37], which aims to identify bird species based on their soundscape. For each dataset, we perform linear probing over obtained audio embeddings from each audio encoder. The same audio preprocessing pipeline is used as described in Section 4.3. We report the top-1 accuracy of

Model	Modality	BirdCLEF-2022 (%)	BirdCLEF-2023 (%)	BirdCLEF-2024 (%)
CLAP [30]	🔊	42.33	32.85	39.72
ImageBind	🔊	47.11	37.46	45.04
TaxaBind	🔊	**52.60**	**42.19**	**49.31**
ImageBind	🔊+📍	60.22	44.04	51.64
TaxaBind	🔊+📍	**65.07**	**46.97**	**56.24**

Table 4. Top-1 linear probing results on the task of bird species audio classification.

Model	Modality	Geo Feature (R2)	Ecoregions (%)	Biomes (%)	GeoPlant (MSE)
GeoCLIP [28]	📍	75.07	72.18	69.69	0.0472
SatClip [27]	📍	74.91	70.08	68.54	0.0594
SINR [1]	📍	**75.90**	68.04	70.22	0.0467
ImageBind	📍	74.51	73.74	**71.73**	0.0449
TaxaBind	📍	74.55	**73.75**	**71.73**	**0.0420**

Table 5. Linear probing results on various geo-aware ecological tasks to demonstrate the effectiveness of our location encoder.

Model	Modality	SatBird-Kenya (MSE)	SatBird-USA-sum (MSE)	SatBird-USA-win (MSE)
CLIP [11]	🛰	0.0832	0.0715	0.0755
RVSA [33]	🛰	0.0842	0.0742	0.0791
ScaleMAE [34]	🛰	0.1200	0.1011	0.0994
SatMAE++ [35]	🛰	0.0953	0.0854	0.0867
Sat2Cap [15]	🛰	0.0836	0.0734	0.0770
Imagebind	🛰	0.0753	0.0662	0.0681
TaxaBind	🛰	**0.0721**	**0.0632**	**0.0661**
ImageBind	🛰+📍	0.0730	0.0648	0.0669
TaxaBind	🛰+📍	**0.0710**	**0.0614**	**0.0642**

Table 6. Linear probing results on species encounter rates prediction using satellite imagery as input on the SatBird dataset.

the models in Table 4. The results demonstrate the effectiveness of our audio encoder in representing the soundscape of species and characterizing the fine-grained differences between them. In the multimodal setting, we add the embeddings of audio and geographic location and then perform linear probing. It is observed that combining the embedding results in an improved performance.

Effectiveness of location encoder. We assess the capability of our location encoder to reason about the ecological traits of a given geographic location. This is important in several downstream applications such as species distribution modeling or climate prediction. To do this, we consider four tasks: 1) geo-feature regression [1]; 2) ecoregion classification [38]; 3) biome classification [38]; and 4) presence-absence prediction of plant species [26]. We use GeoCLIP [28], SatCLIP [27] and SINR [1] as the baseline location encoders. Results reported in Table 5 show that our model outperforms the baseline models in ecoregion, biome, and plant species prediction tasks while also achieving competitive results on the geo-feature regression task. This highlights the effectiveness of our location encoder in addressing a range of geospatial ecological tasks based solely on geographic location input.

Effectiveness of satellite image encoder. Satellite imagery can provide important visual characteristics of a location which can be used for addressing fine-grained ecological problems. We evaluate our satellite image encoder to predict bird species encounter rates. We employ the SatBird dataset [3] and follow their training procedure. We perform linear probing on the embeddings obtained from the satellite image encoder. We compare against several state-of-the-art satellite image encoders and report the results in

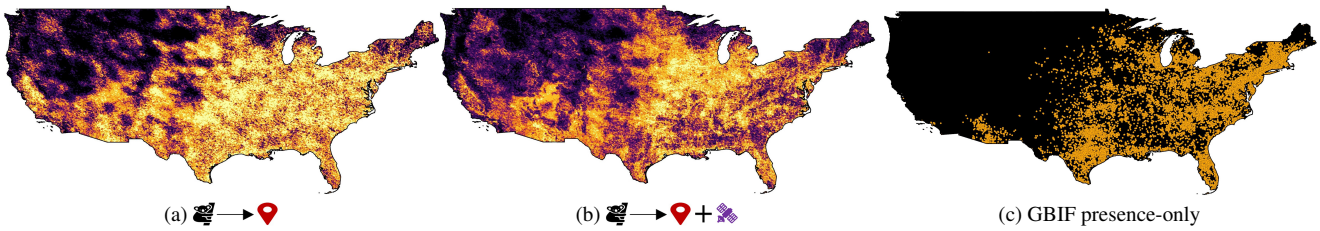


Figure 5. **Zero-shot Species Distribution Map**. We create a species distribution map of *Cardinalis cardinalis* using a query ground-level image and combination of various modalities across the USA.

Table 6. TaxaBind outperforms all the baselines considered for this task. This empirical evidence demonstrates that our model has effectively learned to extract fine-grained ecological traits from satellite imagery.

Species range mapping. We utilize our models to create fine-grained species distribution maps. This is done by computing the similarity between the embeddings of a query ground-level image of species and the embeddings obtained from various modalities for a particular region of interest. The benefit of this approach is that the embeddings for a region can be pre-computed and stored for real-time applications. In Figure 5, we show species distribution maps generated for *Cardinalis cardinalis* over the USA. We downloaded satellite imagery using a dense grid draped over the USA. Qualitatively, we observe that the generated maps are accurate and that using satellite imagery helps to create more fine-grained maps.

5.1. Ablation Studies

The results reported previously demonstrated that our proposed multimodal patching strategy outperforms the ImageBind training strategy. We now study the performance of using different patching strategies below.

Single-modality patching. In Figure 3, we presented the results achieved by the binding modality encoder when patched by training with each modality independently. Now, we investigate whether our sequentially patched model can outperform these single-modality-specific patched models. In Table 7, it is noticed that our model outperforms each of the other patched models. The sequential patching method has an *additive effect* and can distill knowledge from various modalities.

Multimodal patching type. In addition to the sequential patching technique, the parallel patching technique can be used to simultaneously patch the binding modality encoder with all the modalities. This involves averaging the weights of all the single-modality patched models and then determining the best interpolation weight using the patching task. Table 7 demonstrates that parallel patching yields inferior results compared to the sequential patching method, and it even performs worse than models patched using geographic location or satellite images. This indicates that averaging the weights of the individually patched models is not optimal. It also suggests that modalities such as geographic location and satellite imagery are more crucial for the downstream zero-shot classification task.

Method	iNat-2021	TaxaBench-8k
	69.70	33.31
	69.93	34.04
	68.84	33.30
	68.61	33.44
parallel patching	69.66	33.64
sequential patching	**70.09**	**34.45**

Table 7. The Table shows the additive effect of sequential patching, improving upon single-modality patching.

6. Conclusion

In this work, we presented TaxaBind, a unified embedding space for ecological applications covering six modalities: ground-level images, geographic location, satellite images, text, audio, and environmental features. We introduced multimodal patching, a technique to distill knowledge from multiple modalities into a binding modality, improving upon existing methods like ImageBind. We curated two datasets, iSatNat and iSoundNat, to train our models, and introduced TaxaBench-8k, a multimodal dataset for evaluating ecological models. Our extensive experiments demonstrated TaxaBind's effectiveness on various tasks such as zero-shot species classification, cross-modal retrieval, and audio classification, outperforming state-of-the-art methods. Through our multimodal framework, we showed the effectiveness of combining multiple modalities for addressing downstream ecological tasks. We emphasize that our models are general purpose and could potentially be used for other ecology and climate-related applications such as deforestation mapping, tree canopy height mapping, etc. In the future, we plan to explore different techniques for integrating multimodal data in ecology.

References

[1] E. Cole, G. Van Horn, C. Lange, A. Shepard, P. Leary, P. Perona, S. Loarie, and O. Mac Aodha, "Spatial implicit neural representations for global-scale species mapping," in *International Conference on Machine Learning*, pp. 6320–6342, PMLR, 2023. 1, 2, 4, 5, 7

[2] S. Sastry, S. Khanal, A. Dhakal, D. Huang, and N. Jacobs, "Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7136–7145, 2024. 1, 2

[3] 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. 1, 2, 7

[4] J. Dollinger, P. Brun, V. Sainte Fare Garnot, and J. D. Wegner, "Sat-sinr: High-resolution species distribution models through satellite imagery," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 41–48, 2024. 1, 2

[5] 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. 1, 2, 3, 4, 5, 6

[6] C.-H. Yang, B. Feuer, Z. Jubery, Z. K. Deng, A. Nakkab, M. Z. Hasan, S. Chiranjeevi, K. Marshall, N. Baishnab, A. K. Singh, *et al.*, "Arboretum: A large multimodal dataset enabling ai for biodiversity," *arXiv preprint arXiv:2406.17720*, 2024. 1, 2, 5, 6

[7] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. 1, 2, 5, 6

[8] U. Mall, C. P. Phoo, M. K. Liu, C. Vondrick, B. Hariharan, and K. Bala, "Remote sensing vision-language foundation models without annotations via ground remote alignment," *arXiv preprint arXiv:2312.06960*, 2023. 1, 2

[9] P. P. Liang, Z. Deng, M. Q. Ma, J. Y. Zou, L.-P. Morency, and R. Salakhutdinov, "Factorized contrastive learning: Going beyond multi-view redundancy," *Advances in Neural Information Processing Systems*, vol. 36, 2023. 2

[10] G. Ilharco, M. Wortsman, S. Y. Gadre, S. Song, H. Hajishirzi, S. Kornblith, A. Farhadi, and L. Schmidt, "Patching open-vocabulary models by interpolating weights," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29262–29277, 2022. 2, 4

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021. 2, 3, 4, 7

[12] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *Advances in neural information processing systems*, 2018. 2, 3

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020. 2, 3

[14] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 2

[15] A. Dhakal, A. Ahmad, S. Khanal, S. Sastry, H. Kerner, and N. Jacobs, "Sat2cap: Mapping fine-grained textual descriptions from satellite images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 533–542, 2024. 2, 7

[16] S. Khanal, S. Sastry, A. Dhakal, and N. Jacobs, "Learning tri-modal embeddings for zero-shot soundscape mapping," in *British Machine Vision Conference (BMVC)*, Nov. 2023. 2

[17] A. Dhakal, S. Khanal, S. Sastry, A. Ahmad, and N. Jacobs, "Geobind: Binding text, image, and audio through satellite images," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2729–2733, 2024. 2

[18] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022. 2

[19] S. Srivastava and G. Sharma, "Omnivec: Learning robust representations with cross modal sharing," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1236–1248, 2024. 2

[20] S. Srivastava and G. Sharma, "Omnivec2-a novel transformer based network for large scale multimodal and multitask learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27412–27424, 2024. 2

[21] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018. 2, 3, 6

[22] T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet, and A. Joly, "Overview of geolifeclef 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data.," in *CLEF (Working Notes)*, pp. 1940–1956, 2022. 2

[23] Q. Diao, Y. Jiang, B. Wen, J. Sun, and Z. Yuan, "Metaformer: A unified meta framework for fine-grained recognition," *arXiv preprint arXiv:2203.02751*, 2022. 2

[24] S. Sastry, X. Xing, A. Dhakal, S. Khanal, A. Ahmad, and N. Jacobs, "Ld-sdm: Language-driven hierarchical species

distribution modeling," *arXiv preprint arXiv:2312.08334*, 2023. 2

[25] O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019. 2

[26] 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, 7

[27] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," *arXiv preprint arXiv:2311.17179*, 2023. 2, 7

[28] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization," *Advances in Neural Information Processing Systems*, vol. 36, 2023. 2, 4, 7

[29] EOX, "Sentinel-2 cloudless map of the world by EOX." 3

[30] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023. 4, 6, 7

[31] G. Piosenka, "Birds 525 species- image classification," Apr 2023. 6

[32] 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. 6

[33] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022. 7

[34] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023. 7

[35] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27811–27819, 2024. 7

[36] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, *et al.*, "Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 257–285, Springer, 2022. 6

[37] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, *et al.*, "Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 416–439, Springer, 2023. 6

[38] 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. 7