GLOBAL-SCALE SPECIES MAPPING FROM CROWDSOURCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the geographical range of a species from in situ observational data is a challenging and important geospatial prediction problem. Given a set of locations indicating where a species has been observed, the goal is to learn a model that can predict how likely it is for the species to be present at any other location. While this is a well-studied problem, traditional approaches are unable to take advantage of more recently available large-scale datasets that cover many locations and species. We propose a new approach that jointly estimates the geographical ranges of tens of thousands of different species simultaneously. We develop a series of benchmark evaluation tasks that measure different aspects of the species range and spatial representation learning problems. We show that our approach scales both in terms of amount of training data and species, where adding more data enables the models to learn better spatial representations that generalize to other species. Despite being only trained on weakly supervised crowdsourced data, our models can approach the predictions of current expert-developed gold standard models.

1 INTRODUCTION

We are witnessing a dramatic decline in global biodiversity, which has severe ramifications for natural resource management, food security, and ecosystem services that are crucial to human health and well-being Watson et al. (2019). Effectively planning conservation interventions depends on the ability to understanding where different species live. In ecology, this problem is known as species distribution modeling (SDM) Elith & Leathwick (2009).

Estimating distributions for hundreds of thousands of species at a global scale is extremely challenging. The core technical question is how to estimate the distribution for a species of interest based on a limited number of sparse observations. Given a sufficient amount of *presence-absence* data – records of where a species has been observed to occur and where it has been confirmed to be absent – this problem can be approached using traditional methods from machine learning and statistics Beery et al. (2021). However, this is not scalable because presence-absence data is hard to collect at scale. *Presence-only* data – which consists only of locations where a species has been observed, with no confirmed absences – is much more abundant (though not without drawbacks, e.g. Hastie & Fithian (2013)). For instance, the iNaturalist community science platform iNa has collected over 70M observations across 300k species, all of which is presence-only data. Can we use large-scale presence-only data for SDM and geospatial representation learning broadly?

Deep learning is one of our current best tools for dealing with large-scale data. Deep neural networks also have a key advantage over many prior SDM methods because they can *jointly* learn the distribution of multiple species in the same model Chen et al. (2017); Tang et al. (2018); Mac Aodha et al. (2019). This enables these models to learn representations that *share* information across species, resulting in improved predictions Chen et al. (2017). However, with some exceptions Mac Aodha et al. (2019), the majority of these models need presence-absence data for training, which prevents them from scaling beyond a small number of species and regions. As a result, this question of data scale for species range mapping has not been explored in detail in the literature.

In this work, we propose a new approach for joint species range mapping. We show that our approach has desirable scaling properties with respect to training data, i.e. adding more data from diverse species enables us to learn better spatial representations that transfer well to other down stream tasks. To achieve this, we make the following contributions: (i) A novel approach for joint

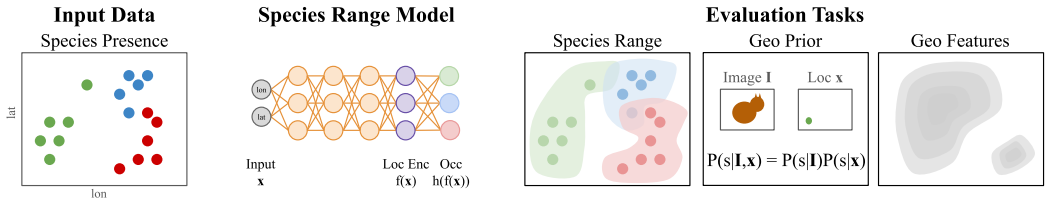


Figure 1: We show that sparse species observation data *(left)* can be used to train neural network-based models to learn representations of space *(middle)*. We evaluate these models and representations across a series of diverse down stream tasks *(right)*. These tasks include: species range estimation, assisting computer vision classifiers, and geographical feature prediction.

species range estimation from sparse, presence-only, crowdsourced data. (ii) A detailed comparison to other loss functions for joint species range estimation. (iii) A new evaluation framework that illustrates the benefit of our method across a range of tasks from species mapping to fine-grained visual categorization. Our intention is that this benchmark will serve as a valuable resource for future researchers who want to make progress on this task.

2 Related Work

Species distribution modeling (SDM) refers to a set of methods that aim to predict where (and sometimes when and in what quantities) species of interest are likely to be found Elith & Leathwick (2009). The literature on SDM is vast, and for the purpose of this work we restrict our discussion to methods which estimate the spatial range for a set of species, i.e. we do not explore other problem settings such as abundance estimation Potts & Elith (2006). For a recent review of SDM for computer scientists, please consult Beery et al. (2021).

Traditional approaches to SDM have focused on training conventional supervised learning models (e.g. logistic regression, decision tress, etc.) to learn a mapping between environmental features (e.g. altitude, average rainfall, etc.) and species presence Phillips et al. (2004); Elith et al. (2006). More recently, deep learning methods have been introduced that instead *jointly* represent multiple different species within the same model Chen et al. (2017); Botella et al. (2018); Tang et al. (2018); Mac Aodha et al. (2019). These models are typically trained on crowdsourced data, which can introduce additional challenges and biases that need to be accounted for during training Fink et al. (2010); Chen & Gomes (2019); Johnston et al. (2020); Botella et al. (2021). We build on the work of Mac Aodha et al. (2019), who proposed an implicit neural network-based approach that forgoes the need for environmental features (e.g. as required by Botella et al. (2018); Tang et al. (2018); Cole et al. (2020)) by learning to predict joint species presence from geographical location alone.

The core technical challenge addressed in this work, learning multi-label predictors from limited supervision, is related to a task of single positive multi-label learning (SPML) Cole et al. (2021); Verelst et al. (2022); Zhou et al. (2022). Here the objective is to train a model that is capable of making multi-label predictions at test time, despite having only ever observed one positive label per training instance (i.e. no confirmed negative training labels). Recently, Zhou et al. (2022) presented an approach based on the principle of entropy maximization that outperforms existing methods. Our work connects the SPML and species range mapping literature, demonstrating that large-scale species distribution modeling is a particularly difficult instance of SPML owing to the extremely large label space and the numerous sources of potential biases in the data. Due to the large output label space size, recent SPML methods such as Cole et al. (2021) are not computationally viable for the large-scale range estimation problem.

Our task is related to the growing number of works that explore the use of coordinate neural networks for different implicit prediction tasks. These include methods for encoding images in neural network weights Tancik et al. (2020), 3D scenes Sitzmann et al. (2019), and methods for performing novel-view image synthesis Mildenhall et al. (2020). A subset if these methods explore the impact of different network activation functions Sitzmann et al. (2019); Ramasinghe & Lucey (2022) and input encodingsTancik et al. (2020); Zheng et al. (2022). In the majority of the cases above, the choice of training objective is clear, e.g. mean squared error between the observed and predicted data. However, in the case of species range estimation from presence-only data, it is not clear what training loss is most effective. We attempt to answer this question in later experiments.

Quantitatively benchmarking the performance of species distribution models is notoriously difficult due to the lack of large-scale evaluation data with confirmed presence and absence information Beery et al. (2021). As a result, one common approach is to evaluate performance on a small set of species from a limited geographical region where it is feasible to collect complete data, e.g. as done in Potts & Elith (2006). In this work, we instead compare the performance of our models to gold standard expert designed models that are trained with additional absence data and environmental metadata. An alternative evaluation approach is to measure the performance on a related "proxy" task. For example, there have been a number of works that combine models trained for species range estimation with deep image classifiers to aid image classification Berg et al. (2014); Mac Aodha et al. (2019); Chu et al. (2019); Mai et al. (2020); Terry et al. (2020); Skreta et al. (2020); Yang et al. (2022;?). Broadly, these works can be divided into those that jointly train a deep image classifier with a geographical prior, e.g. Tang et al. (2015); Chu et al. (2019), versus those, like ours, that train a *separate* prior that can be combined with any image classifier, e.g. Berg et al. (2014); Mac Aodha et al. (2019). Later we evaluate different range estimation methods on the task of aiding fine-grained image classification. We also evaluate the spatial representations learned by our models via transfer learning, using them as inputs for a suite of geospatial regression tasks.

3 METHODS

3.1 PRELIMINARIES

Problem statement. Let $\mathbf{x} = [lon, lat]$ denote a geographical location (i.e. longitude and latitude). Let $\mathbf{y} \in \{0,1\}^S$ denote the true presence (1) or absence (0) of S different species at $\mathbf{x}$. We introduce $\mathbf{z} \in \{0,1,\varnothing\}^S$ to represent our observed data at $\mathbf{x}$, where $z_j = 1$ if species j is present, $z_j = 0$ if species j is absent, and $z_j = \varnothing$ if we do not know whether species j is present or absent. Our goal is to develop a model that produces an estimate of $\mathbf{y}$ at any location $\mathbf{x}$ over some spatial domain $\mathcal{X}$, given observed data $\{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$. We parameterize this model as $\hat{\boldsymbol{y}} = h_\phi(f_\theta(\boldsymbol{x}))$, where $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$ is a location encoder and $h_\phi : \mathbb{R}^k \rightarrow [0,1]^S$ is a multi-label classifier. The prediction $\hat{\boldsymbol{y}} \in [0,1]^S$ is our estimate of the presence or absence of each species at $\mathbf{x}$. Intuitively, the location encoder f_θ learns a representation of space that is used by the multi-label classifier h_ϕ to predict species presence at a given location.

Input encoding. Each species observation is associated with a point $\mathbf{x} = [lon, lat]$. We rescale these values so that $lon, lat \in [-1, 1]$. Following Mac Aodha et al. (2019) we guard against boundary effects by using a sinusoidal encoding for the input. The final result is an input vector

$$\mathbf{x} = [\sin(lon), \cos(lon), \sin(lat), \cos(lat)] \,. \tag{1}$$

Alternative input encodings for related coordinate networks have been explored in existing literature Mai et al. (2020); Tancik et al. (2020); Mai et al. (2022); Zheng et al. (2022). Their choice is orthogonal to the losses we explore, thus we leave their evaluation for future work.

Presence-absence vs. presence-only data. Species observation datasets come in two varieties: (i) *Presence-absence* data consists of locations where a species has been observed to be present and locations where it has been confirmed to be absent. That is, we say we have presence-absence data for species j if $|\{\mathbf{z}_i : z_{ij} = 0\}| > 0$ and $|\{\mathbf{z}_i : z_{ij} = 1\}| > 0$. Unfortunately, presence-absence data is costly to obtain at scale because confirming absence requires skilled observers to exhaustively search an area. (ii) *Presence-only* data is easier to acquire and thus more abundant because absences are not collected, i.e. $z_{ij} \in \{1, \varnothing\}$ for $i \in [N]$ and $j \in [S]$.

3.2 LEARNING FROM LARGE-SCALE PRESENCE-ONLY DATA

In the context of SPML *image* classification, a simple but effective approach is to assume that unobserved labels are negative Cole et al. (2021). This approach is based on a probabilistic argument: since natural images tend to have a small number of categories compared to the size of the label set, the vast majority of the labels will be negative. This is also true for species distribution modeling. Given an arbitrary location and a large set of candidate species, nearly all of them will be absent. In this section we describe several simple and scalable loss functions based on this idea.

"Assume negative" loss (same species, different location). This loss pairs each observation of a species with a pseudo-negative for that species at another location chosen uniformly at random:

$$\mathcal{L}_{\text{AN}-\text{SSDL}}(\hat{\mathbf{y}}, \mathbf{z}) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]} [\log(\hat{y}_j) + \log(1 - \hat{y}'_j)] \tag{2}$$

where $\hat{\mathbf{y}}' = h_\phi(f_\theta(\mathbf{r}))$ with $\mathbf{r} \sim \text{Unif}(\mathcal{X})$ and $n_{\text{pos}} = \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]}$. This approach generates pseudo-negatives across the globe, but many of them are likely to be "easy" because they are in the ocean or far from the true species range.

"Assume negative" loss (same location, different species). This loss pairs each observation of a species with a pseudo-negative at the same location for a different species:

$$\mathcal{L}_{\text{AN}-\text{SLDS}}(\hat{\mathbf{y}}, \mathbf{z}) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]} [\log(\hat{y}_j) + \log(1 - \hat{y}_{j'})] \tag{3}$$

where $j' \sim \text{Unif}(\{j : z_j \neq 1\})$. Intuitively, this approach generates pseudo-negatives that are aligned with the spatial distribution of the observed data.

Full "assume negative" loss. The previous two losses are inefficient in the sense that they do not use all of the entries in $\hat{\mathbf{y}}$. We can combine the pseudo-negative sampling strategies of $\mathcal{L}_{\text{AN}-\text{SSDL}}$ and $\mathcal{L}_{\text{AN}-\text{SLDS}}$ and use all available predictions as follows:

$$\mathcal{L}_{\text{AN}-\text{full}}(\hat{\mathbf{y}}, \mathbf{z}) = \frac{1}{S} \sum_{j=1}^{S} \left[\mathbb{1}_{[z_j=1]} \lambda \log(\hat{y}_j) + \mathbb{1}_{[z_j \neq 1]} \log(1 - \hat{y}_j) + \log(1 - \hat{y}'_j) \right] \tag{4}$$

where $\hat{\mathbf{y}}' = h_\phi(f_\theta(\mathbf{r}))$ with $\mathbf{r} \sim \text{Unif}(\mathcal{X})$. The hyperparameter $\lambda > 0$ can be used to prevent the negative labels from dominating the loss. This is equivalent to the loss from Mac Aodha et al. (2019) without their user modeling terms.

Maximum entropy loss. The current state-of-the-art in SPML image classification is a technique proposed by Zhou et al. (2022). The idea is to encourage predictions for unobserved labels to maximize entropy instead of forcing them to zero. Zhou et al. (2022) also includes an additional pseudo-labeling component, but we omit this because the resulting empirical improvement from it is small. We can apply this idea to $\mathcal{L}_{\text{AN}-\text{SSDL}}$, $\mathcal{L}_{\text{AN}-\text{SLDS}}$, and $\mathcal{L}_{\text{AN}-\text{full}}$ by replacing all terms of the form "$-\log(1-p)$" with terms of the form "$H(p)$" where $H(p) = -(p\log(p) + (1-p)\log(1-p))$. We name these "maximum entropy" (ME) variants $\mathcal{L}_{\text{ME}-\text{SSDL}}$, $\mathcal{L}_{\text{ME}-\text{SLDS}}$, and $\mathcal{L}_{\text{ME}-\text{full}}$.

4 Experiments

In this section we evaluate the performance of the different losses outlined previously across a range of species and environmental prediction tasks.

4.1 Models

As described in Section 3.1, our model consists of a location encoder f_θ and a multi-label classifier h_ϕ which produce a vector of predictions $\hat{y} = h_\phi(f_\theta(\mathbf{x}))$ for a location $\mathbf{x}$. The location encoder f_θ is implemented as the fully connected neural network shown in Figure A1. We implement the multi-label classifier h_ϕ as a single linear layer with sigmoid activations. Full implementation details can be found in Section A in the appendix.

Linear model variant. For the rows in Table 1 tagged with "(single layer)", the location encoder f_θ is replaced with the identity function and h_ϕ is unchanged. In this case, we are mapping the encoded location $\mathbf{x}$ to species predictions $\hat{\mathbf{y}}$ using a single linear layer with sigmoid activations. This ablation highlights the importance of the deeper location encoder in terms of learning an effective representation of space.

4.2 Training Data

We train our models on presence-only species observation data obtained from the community science platform iNaturalist iNa. The training set consists of 35.5 million observations covering 47,375

species observed prior to 2022. For each species, we have the geographical coordinate where the species was observed. To be included in the training set, each species needed at least 50 observations. Some species are more common than others, and thus the dataset is heavily imbalanced. Later we use this data in its entirety ("All"), with different maximum observations per class ("X / Class"), or different subsets of classes. See Section B in the appendix for more details on the dataset.

4.3 Evaluation Tasks

Here we briefly describe the four benchmark evaluation tasks we propose for evaluating large-scale species range map estimation models. Further details are provided in Section C in the appendix.

S&T: eBird Stats and Trends. The aim of this task is to quantify how well our presence-only trained models can predict high quality range maps for a set of 536 commonly observed bird species. The target data consists of bird range estimates from the *eBird Status & Trends* dataset Fink et al. (2020). For this data, the ranges are estimated from models that are trained on different higher quality presence-absence data than us sourced from eBird, use environmental features as input, and make use of additional expert knowledge for data filtering. Performance is measured in terms of mean average precision (MAP) by comparing our predicted range maps to the S&T targets.

IUCN: Range Maps. The previous task was restricted to data from only birds species. In this task, we use expert generated ranges from a wider set of 3,938 species (e.g. mammals) from all around the world using range maps from the International Union for Conservation of Nature (IUCN) Red List IUC. Again, performance is measured in terms of MAP across all species.

Geo Prior: Geographical Priors for Computer Vision. In this task, similar to Berg et al. (2014); Mac Aodha et al. (2019), we combine the probabilistic occupancy predictions from our range estimation models with the predictions of a deep image classifier. Specifically, we use a set of 282,974 images with associated location information from iNaturalist, coming from 39,444 species that overlap with our training set. We compute the improvement in top-1 classification accuracy when the output of the image classifier is combined with a range estimation model.

Geo Feature: Environmental Representation Learning. In contrast to the previous tasks which evaluate a model's species range estimation performance, either directly or indirectly, in this task we evaluate the underlying spatial features that are learned. We first remove the final classifier h_ϕ from a given model, and then extract features for a dense set of spatial locations from the location encoder f_θ. Finally we train a linear regressor on the extracted frozen features, where the goal is to predict one of nine different spatially varying data layers, e.g. 'ELE: Elevation'. Performance is evaluated using the coefficient of determination R^2 on a spatially disjoint set of test data.

4.4 Results

We present results for each of our evaluation tasks ("S&T", "IUCN", "Geo Prior", and "Geo Feature") in Table 1. In particular, we train $\mathcal{L}_{\text{AN}-\text{SSDL}}$, $\mathcal{L}_{\text{AN}-\text{SLDS}}$, and $\mathcal{L}_{\text{AN}-\text{full}}$ with an increasing number of examples per category. To understand the importance of the representational power of the neural network, we also train a single layer network using the $\mathcal{L}_{\text{AN}-\text{full}}$ loss with increasing amounts of training data. We also provide a simple but effective baseline that makes predictions by aggregating the 35.5M training examples into a discretized grid ("Best Discretized Grid", described in Section A.3). Lastly we include the training loss from Mac Aodha et al. (2019) (which is essentially $\mathcal{L}_{\text{AN}-\text{full}}$ with additional user modeling) and the "maximum entropy" loss variants $\mathcal{L}_{\text{ME}-\text{SSDL}}$, $\mathcal{L}_{\text{SLDS}}$ and $\text{L}_{\text{ME}-\text{full}}$ inspired by Zhou et al. (2022). Note that the "IUCN" results are quite high in absolute terms - we discuss this in Section C.2.

In Figure 2 we dig deeper into the relationship between "S&T" task performance and training set size. We analyze the effect of the number of training examples per category and the number of categories in the training set. The "Geo Feature" results in Table 1 are actually averaged across nine sub-tasks, which are described in detail in Table A2. In Figure 3 we show per-task performance for different loss functions.

We also provide some qualitative results to give more insight into what the models are learning. Figure 4 compares the spatial representations learned from 10 and 1000 examples per class using

$\mathcal{L}_{\text{AN}-\text{full}}$. We also visualize model predictions for different training configurations in Figure 5 and 6. Unless otherwise specified, we mask out any predictions on oceans in these visualizations.

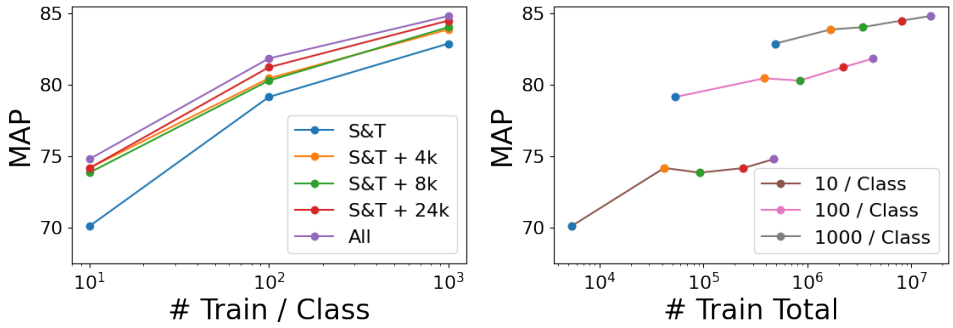


Figure 2: S&T range prediction performance as a function of training set size. We consider two notions of size: number of training examples per class and number of classes (i.e. species). The two plots visualize the same data in different ways. *(Left)* Each curve corresponds to a model trained on a different set of species. For the "S&T" line, we only train on the species in the S&T task. For the "S&T + X" lines, we add in X species chosen uniformly at random. For the "All" line we train on all 47k species. The x-axis represents the number of training examples per class. *(Right)* Each curve corresponds to a different amount of training data per class. The points along each curve correspond to the different training sets examined in the left panel (i.e. "S&T", "S&T + 4k", ..., "All"). The x-axis represents the total number of training examples. Unsurprisingly, increasing the number of training examples per category improves performance. More interestingly, adding training data for additional species (which are not evaluated at test time) improves performance as well.

Table 1: Results for four geospatial tasks: **S&T** (eBird Status & Trends species mapping), **IUCN** (IUCN species mapping) , **Geo Prior** (fine-grained image classification with a geographical prior), and **Geo Feature** (geographical feature regression). We assess performance as a function of the loss function and the amount of training data ("# / Class"). The "single layer" and "Best Discretized Grid" baselines do not have an entry for the Geo Feature task as they do not learn a location encoder.

Loss	# / Class	**S&T** (MAP)	**IUCN** (MAP)	**Geo Prior** (Δ Top-1)	**Geo Feature** (Mean R^2)
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	10	65.65	95.94	+3.6	0.530
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	100	73.54	97.54	+4.7	0.589
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	1000	76.52	98.02	+4.8	0.624
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	All	77.35	98.11	+4.9	0.650
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	10	72.09	92.74	+4.8	0.559
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	100	79.66	95.06	+6.2	0.577
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	1000	82.68	94.45	+6.2	0.619
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	All	82.98	95.36	+5.9	0.707
$\mathcal{L}_{\text{AN}-\text{full}}$	10	74.78	96.70	+4.5	0.611
$\mathcal{L}_{\text{AN}-\text{full}}$	100	81.83	97.78	+6.6	0.552
$\mathcal{L}_{\text{AN}-\text{full}}$	1000	84.80	98.27	+6.2	0.589
$\mathcal{L}_{\text{AN}-\text{full}}$	All	85.15	98.19	+5.0	0.543
Baselines:					
$\mathcal{L}_{\text{AN}-\text{full}}$ (single layer)	10	31.46	49.25	-0.4	-
$\mathcal{L}_{\text{AN}-\text{full}}$ (single layer)	100	30.05	44.95	-0.3	-
$\mathcal{L}_{\text{AN}-\text{full}}$ (single layer)	1000	32.14	39.49	-0.6	-
Best Discretized Grid	All	72.87	94.42	+4.1	-
Loss Variants:					
$\mathcal{L}_{\text{GP}}$ [Mac Aodha et al. (2019)]	1000	81.97	97.73	+5.7	0.535
$\mathcal{L}_{\text{ME}-\text{SSDL}}$	1000	72.48	96.56	+1.6	0.524
$\mathcal{L}_{\text{ME}-\text{SLDS}}$	1000	81.86	93.62	+2.1	0.579
$\mathcal{L}_{\text{ME}-\text{full}}$	1000	82.52	94.71	+1.5	0.703

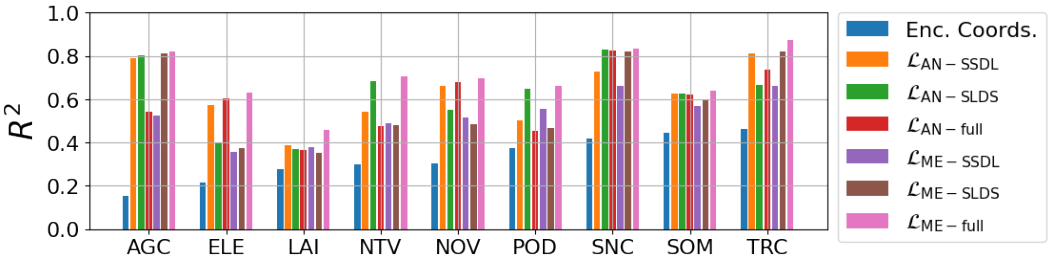


Figure 3: "Geo Feature" task performance broken down by subtask. All methods are trained with 1000 examples per category. The regression tasks are sorted by the performance of "Enc. Coords." which is a naive baseline method that only uses the raw encoded coordinates from Equation 1 as features. All losses outperform this naive baseline, but most losses perform much better on some tasks than others. Only $\mathcal{L}_{\text{ME-full}}$ consistently performs well.

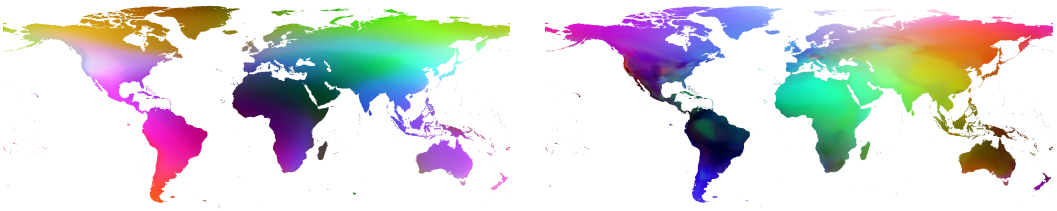


Figure 4: Visualization of the 256-dimensional learned location encoder features projected to three dimensions using Independent Component Analysis. *(Left)* This corresponds to the $\mathcal{L}_{\text{AN-full}}$ model with a maximum of *10* examples per class. The features are smooth and do not appear to encode much high frequency spatial information. *(Right)* In contrast, the $\mathcal{L}_{\text{AN-full}}$ model with a maximum of *1000* examples per class, contains more high frequency information. The increase in training data appears to enable this model to better encode spatially varying environmental properties.

4.5 Discussion

Joint learning across categories is beneficial, but more data is better. In Figure 2 we study the effect of the amount of training data on performance for the "S&T" task. We first note that increasing the number of training examples per species reliably and significantly improves performance. One possible mechanism for this is suggested by Figure 4, which shows a more spatially detailed representation emerging with more training data. We observe that adding additional species to the training set (which are not in the test set) also tends to increase performance, though by a smaller margin. In the right panel of Figure 2 we can see that it is better to have more training data for fewer species than less training data spread across more species. Table 1 shows a clear trend towards higher performance with more training data across tasks, with a few exceptions: $\mathcal{L}_{\text{AN-full}}$ performs worse on the "Geo Feature" task as we add data, and both $\mathcal{L}_{\text{AN-SLDS}}$ and $\mathcal{L}_{\text{AN-full}}$ have worse performance on the "Geo Prior" task when using all data compared to using 1000 / class.

The "best" loss depends on the task. In Table 1 we evaluate a number of different losses across four tasks, but no loss is best across all tasks. For instance, Figure 3 shows that most losses do well on some "Geo Feature" subtasks and poorly on others. The exception is $\mathcal{L}_{\text{ME-full}}$, which is a consistent top performer. However, $\mathcal{L}_{\text{ME-full}}$ is by far worst loss for the "Geo Prior" task (excluding the "single layer" variants). We can also observe stark qualitative differences between the losses when we visualize the predictions as in Figure 5 and 6.

Method rankings may not generalize across domains. Formally speaking, the presence-only range estimation problem in this work and the SPML image classification problem in Cole et al. (2021) are both SPML problems. Despite this formal equivalence, it does not seem to be the case that the best current method for SPML image classification is also the best methods for presence-only SDM. Zhou et al. (2022) show that their "maximum entropy" loss performs much better than the "assume negative" loss across a number of image classification datasets. However, all of the "maximum entropy" losses in Table 1 ($\mathcal{L}_{\text{ME-SSDL}}$, $\mathcal{L}_{\text{ME-SLDS}}$, $\mathcal{L}_{\text{ME-full}}$) under perform their "assume negative" counterparts ($\mathcal{L}_{\text{AN-SSDL}}$, $\mathcal{L}_{\text{AN-SLDS}}$, $\mathcal{L}_{\text{AN-full}}$) on the same datasets.

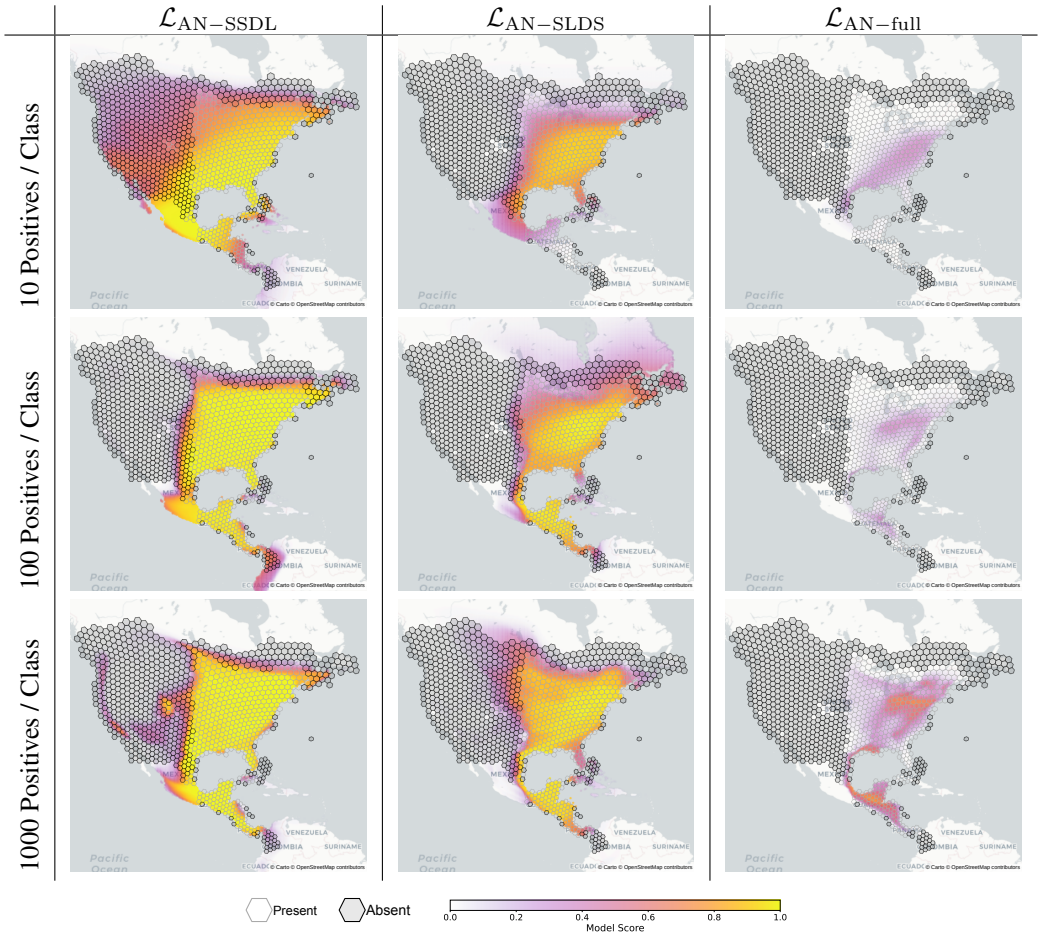


Figure 5: Dense predictions for Wood Thrush when varying the amount of positive examples (rows) under different loss functions (columns). The model score values are generated on a dense grid of $[lon, lat]$ coordinates. Hexagons signify locations where we can evaluate the model outputs using S&T data. Transparent hexagons are locations where the model should predict present, gray hexagons with black border are locations where the model should predict absent, and locations without hexagons are not used for S&T evaluation. All models improve their range maps when given access to more data, as expected. $\mathcal{L}_{\text{AN-SSDL}}$ over estimates the range with few examples, and refines the range with additional data, while $\mathcal{L}_{\text{AN-full}}$ starts off conservative with few training examples and expands as more data is available. The range predicted by $\mathcal{L}_{\text{AN-SLDS}}$ is somewhere in-between. The magnitude of the score values is also interesting, with both $\mathcal{L}_{\text{AN-SSDL}}$ and $\mathcal{L}_{\text{AN-SLDS}}$ predicting more locations confidently than $\mathcal{L}_{\text{AN-full}}$. No calibration process has been applied to the outputs.

4.6 LIMITATIONS

There are some limitations associated with our analysis. As noted, the training set is heavily imbalanced, both in terms of the species themselves and where the data was collected. This is partially because some species are more common and thus more likely to be observed than others. We do not explicitly attempted to deal with the species imbalance, other than by showing that the ordering of methods does not significantly vary even when the training data is capped to the same upper limit per species, i.e. 10, 100, or 1,000 examples per class in Table 1.

Reliably evaluating the performance of models for species range estimation is a long standing and challenging question. To address this issue, we present a suite of complementary benchmarks that attempt to evaluate different facets of the environmental spatial prediction problem. However, obtaining oracle ground truth distribution data for tens of thousands of species remains a very difficult problem. Our benchmarks may be lacking in some aspects, e.g. potentially inaccurate range maps for under sampled locations. We flag this here to ensure that readers are aware of these limitations.

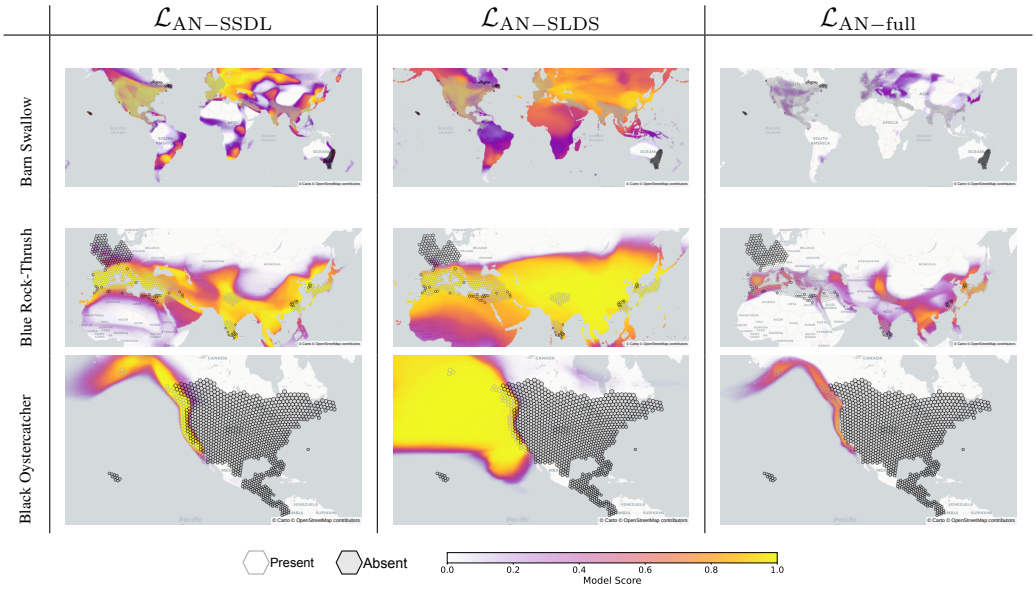


Figure 6: Loss function comparison (columns) for three different species of birds (rows); all models were trained with 1000 positives / class. See Figure 5 for an explanation of the plots. These species were chosen for visualization because they are particularly interesting: Barn Swallow is a species that occurs across the globe; Blue Rock-Thrush is a species whose range goes from the data rich Western Palearctic, through a data sparse area, and back to a more data rich area of eastern Asia; Black Oystercatcher is a coastal species the hugs the west coast of the United States. The three visualizations for Black Oystercatcher are unique in that they do not use an ocean mask to remove predictions from non-land regions. Here, we specifically wanted to see if the models learned to follow the coastline, and we observe that $\mathcal{L}_{\text{AN}-\text{SLDS}}$ incorrectly expands the range into the Pacific.

Finally, extreme care should be taken before attempting to make any down stream conservation decisions based on the outputs of models such as the ones presented here. Our goal in this work is to show that the scaling properties of deep networks, and the power of representation learning, can have a big impact on the task of species range estimation. Our models have not been calibrated or validated beyond the experiments illustrated above.

5 Conclusion

We explored the problem of species range estimation through the lens of representation learning. Across a range of spatial prediction tasks, we showed that adding more and diverse data improves model performance. In doing so, we also made connections to recent literature on learning multi-label classifiers from limited supervision in addition to implicit coordinate networks. We hope that our work will encourage more machine learning researchers to work on this important problem.

While the initial results are encouraging, there are still many avenues for future work. For example, currently our models do not make use of any temporal input information Mac Aodha et al. (2019), we also do not account for the spatial biases present in the data Chen & Gomes (2019), and our neural network models could be adapted to have better inductive biases for encoding spatially varying signals Ramasinghe & Lucey (2022).

REFERENCES

IUCN Red List. https://www.iucnredlist.org, accessed 28 September 2022.

iNaturalist. www.inaturalist.org, accessed 28 September 2022.

Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. In *Conference on Computing and Sustainable Societies*, 2021.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014.

Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A deep learning approach to species distribution modelling. In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. 2018.

Christophe Botella, Alexis Joly, Pierre Bonnet, François Munoz, and Pascal Monestiez. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 2021.

Di Chen and Carla P Gomes. Bias reduction via end-to-end shift learning: Application to citizen science. In *AAAI*, 2019.

Di Chen, Yexiang Xue, Daniel Fink, Shuo Chen, and Carla P Gomes. Deep multi-species embedding. In *IJCAI*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *ICCV Workshops*, 2019.

Elijah Cole, Benjamin Deneu, Titouan Lorieul, Maximilien Servajean, Christophe Botella, Dan Morris, Nebojsa Jojic, Pierre Bonnet, and Alexis Joly. The geolifeclef 2020 dataset. *arXiv:2004.04192*, 2020.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *CVPR*, 2021.

Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 2009.

Jane Elith, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 2006.

Daniel Fink, Wesley M Hochachka, Benjamin Zuckerberg, David W Winkler, Ben Shaby, M Arthur Munson, Giles Hooker, Mirek Riedewald, Daniel Sheldon, and Steve Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 2010.

Daniel Fink, Tom Auer, Alison Johnston, Matt Strimas-Mackey, Orin Robinson, Shawn Ligocki, Wesley Hochachka, Lauren Jaromczyk, Chris Wood, Ian Davies, Marshall Iliff, and Luke Seitz. ebird status and trends, data version: 2020; released: 2021. *Cornell Lab of Ornithology, Ithaca, New York*, 10, 2020.

Trevor Hastie and Will Fithian. Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867, 2013.

Alison Johnston, Nick Moran, Andy Musgrove, Daniel Fink, and Stephen R Baillie. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 2020.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *ICLR*, 2020.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Krzysztof Janowicz, and Ni Lao. Sphere2vec: Multi-scale representation learning over a spherical surface for geospatial predictions. *arXiv:2201.10489*, 2022.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Steven J Phillips, Miroslav Dudík, and Robert E Schapire. A maximum entropy approach to species distribution modeling. In *ICML*, 2004.

Joanne M Potts and Jane Elith. Comparing species abundance models. *Ecological modelling*, 2006.

Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *ECCV*, 2022.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 2021.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 2019.

Marta Skreta, Alexandra Luccioni, and David Rolnick. Spatiotemporal features improve fine-grained butterfly image classification. In *Tackling Climate Change with Machine Learning Workshop at NeurIPS*, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *ICCV*, 2015.

Luming Tang, Yexiang Xue, Di Chen, and Carla Gomes. Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *AAAI*, 2018.

J Christopher D Terry, Helen E Roy, and Tom A August. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, 2020.

Thomas Verelst, Paul K Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. *arXiv:2203.06127*, 2022.

Robert Watson, Ivar Baste, Anne Larigauderie, Paul Leadley, Unai Pascual, Brigitte Baptiste, Sebsebe Demissew, Luthando Dziba, Günay Erpul, Asghar Fazel, et al. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Bonn, Germany: IPBES Secretariat, 2019.

Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiajun Liang, and Jian Yang. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *CVPR*, 2022.

Jianqiao Zheng, Sameera Ramasinghe, Xueqian Li, and Simon Lucey. Trading positional complexity vs. deepness in coordinate networks. In *ECCV*, 2022.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *ECCV*, 2022.

Appendix

A Implementation Details

A.1 Network Architecture

Unless otherwise specified, we use the network in Figure A1 for our location encoder f_θ. The right side of the figure shows the network structure, consisting of one standard linear layer and four residual layers. The left side of the figure shows the structure of a single residual layer. Note that all layers have the same number of nodes. Every layer of the network has 256 nodes and we use $p = 0.5$ for the dropout probability.

A.2 Training Details

Environment. All models were trained on an Amazon AWS `p3.2xlarge` instance with a Tesla V100 GPU and 60 GB RAM. The model training code was written in PyTorch (v1.7.0).

Hyperparameters. All models were trained for 10 epochs using a batch size of 2048. We used the Adam optimizer with an exponential learning rate decay schedule of

$$\texttt{learning_rate} = \texttt{initial_learning_rate} \times \texttt{epoch}^{0.98}$$

where $\texttt{epoch} \in \{0, 1, \ldots, 9\}$. For each model we tuned the learning rate over $\{1e{-}2, 1e{-}3, 1e{-}4\}$ and selected the model with the best validation performance on the "S&T" task. For $\mathcal{L}_{\text{AN-full}}$ and $\mathcal{L}_{\text{GP}}$ we set $\lambda = 2048$ as in Mac Aodha et al. (2019).

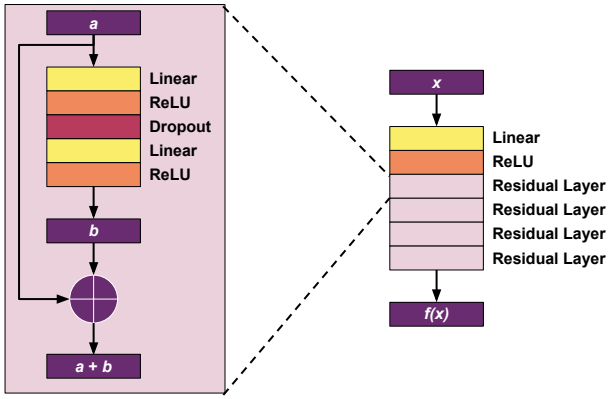


Figure A1: Network diagram for the fully connected network (with residual connections) which we use for our location encoder f_θ.

A.3 Baselines

Our baseline system for the S&T, IUCN, and Geo Prior tasks is a simple binning method. The world is discretized into a set of hexagons using the H3 geospatial indexing library. We then bin the training observations into the hexagon "buckets." For the S&T and IUCN tasks, for a given species, we convert the bucket counts into a score by dividing by the maximum count for that species in any bucket. For a given test observation, we hash the location to the corresponding hexagon, and retrieve the score for the species under question, providing a number between 0 and 1 that can then be used to compute average precision metrics. For the Geo Prior task, we build a binary mask for each hexagon that gives a value of 1 to all species found in the bucket (regardless of count), and a value of 0 for any species that did not hash to the bucket. This binary vector is then used to modulate the predictions of the computer vision system. We sweep a few different hexagon resolutions and provide results in Table A1.

Table A1: Discretized Grid baseline results on test data when using various hexagon resolution for 'training' the model. As this baseline does not learn a location encoder, it is not possible to evaluate it on the Geo Feature task.

		Species Range	IUCN	Geo Prior	Geo Feature
Hex Res	# / Cls.	MAP	MAP	Δ Top-1	r^2
0	All	61.50	94.42	3.5	-
1	All	70.05	90.43	4.1	-
2	All	72.87	81.07	3.1	-
3	All	62.85	69.22	-0.9	-

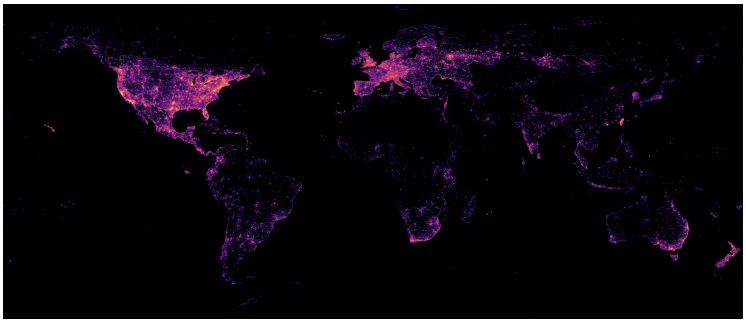


Figure A2: Histogram of the locations of the observations from our iNaturalist training set. Here we bin the data for all 35 million observations across all 47,375 species. Darker colors indicate fewer observations, brighter colors indicate more. The training data is biased towards North America, Europe, and New Zealand.

B Training Dataset

B.1 Dataset Construction

Our training set was collected from the community science platform iNaturalist iNa. iNaturalist enables members of the public, around the world, to collect and share observations pertaining to the presence of different species that span the entire tree of life. Each observation typically takes the form of an image and associated meta data indicating when, where, by whom the observation was made. For our purposes, we only need the observation location information. iNaturalist data only contains presence observations, i.e. we do not have access to any confirmed absences in our training data.

The training data was taken from the iNaturalist AWS Open Dataset[1] in May 2022. The data went through several filtering steps. Specifically, taxa (i.e. a taxonomic group of any rank, such as a species, genus, or family) were filtered so that they contained valid latitude and longitude data, they came from high quality observations i.e. ones with 'research grade' status which indicates that there is a consensus from the iNaturalist community regarding their taxonomic identity, the taxa had to still be active, they were identified to the species-level by the community, and a taxa had to have a minimum of 50 observations to be included. Finally, we only included observations made prior to 2022 to ensure that we could perform a temporal hold-out split to enable us to perform evaluation on other iNaturalist data in the future. This resulted in a total of 35,500,262 valid species observations from 47,375 distinct species. A visualization of the resulting data can be seen in Figure A2. The species in the dataset are heavily imbalanced, reflecting the natural frequency that they are reported to iNaturalist. We use this dataset as our source of training data for the models in the main paper.

B.2 Changing the Number of Training Examples per Category

In the main paper we consider the impact of the number of *observations* per species in the training set by training on different sub-sampled datasets. We construct these datasets by choosing k obser-

[1] https://github.com/inaturalist/inaturalist-open-data

vations per species, uniformly at random. We set a seed to ensure that we are always using the same k observations per category. We also make certain that sampled datasets are nested, so the dataset with k_1 examples per category is a superset of the dataset with $k_2 < k_1$ examples per category. If a category has fewer than k observations, we use them all.

B.3 Changing the Number of Training Categories

In the main paper we also consider the impact of the number of *species* in the training set. In particular, we consider the following nested subsets:

- The set of 536 bird species in the eBird Status & Trends dataset.
- The eBird Status & Trends species plus an additional A randomly selected species, where $A \in \{4000, 8000, 24000\}$.

C Evaluation Tasks

Here we provide additional details on the benchmark tasks used in the main paper. For each task, we outline the dataset properties, how it was collected, and the evaluation metrics used.

C.1 S&T: eBird Stats and Trends

Task: The goal of this task is to evaluate how effective models trained on noisy crowed sourced data from iNaturalist are at predicting species range maps. We use the *eBird Status & Trends* data from Fink et al. (2020) to evaluate our range predictions. This dataset consists of estimated range maps for 536 species of birds predominately found in North America, but also other regions. The range maps are the output of an expert crafted model Fink et al. (2020) that has been trained on tens of millions of presence and absence observations, makes use of additional expert knowledge to perform data filtering, and uses rich environmental covariates as input. While not without its own limitations, we treat this data as 'ground truth' for evaluation purposes as it is developed using much higher quality data compared to what we use to train our models.

Dataset: We perform several pre-processing steps on the raw data. We first spatially bin the data using hexagons at resolution five as defined by the H3 project [2]. $2,016,842$ hexagons cover the world at this resolution, each with an average area of 252.9km^2. Our goal is to predict the presence or absence of a given species in each hexagon using the *eBird Status & Trends* output as 'psuedo' ground truth. The evaluation regions are restricted to those where the *eBird Status & Trends* models have determined that there is sufficient data to make a prediction for a given species. Thus, the set of evaluation regions can vary from species to species. For example, `Melozone aberti` has 23,282 locations with known presence or absence, of which 271 are deemed present. On the other hand, `Columba livia` has 67,251 locations with known presence or absence, of which 39,754 are deemed present.

eBird Status & Trends predictions for a given species are made a multiple different time points throughout the year, i.e. for each week. For the purposes of our evaluation, we collapse all time points to one, and count a hexagon region as being a presence for a given bird if the output of their model is greater than zero for any week in the year for that species.

We create a validation split by choosing 50% of the hexes with known presence/absence status uniformly at random for each species. This is in line we the notion of generalization we care about. That is, we want to know how well we can reproduce expert range maps from raw, potentially noisy, data, *not* how well we can make predictions in previously unseen locations.

Evaluation: We use mean average precision (MAP) as our evaluation metric, only evaluating on valid regions for a given species.

[2] https://h3geo.org

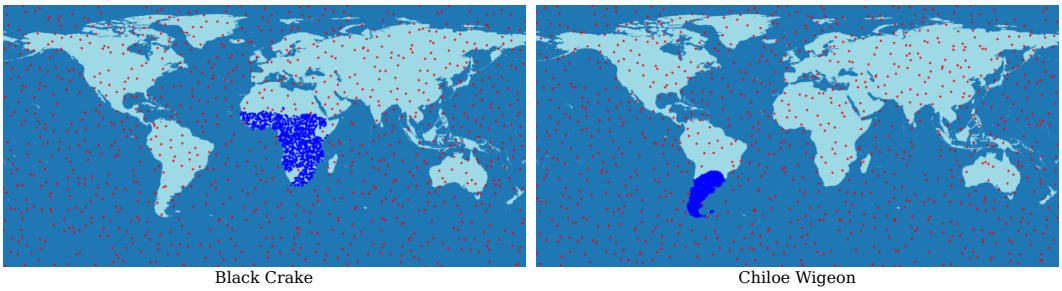


Figure A3: Here we show two example species from the IUNC range map evaluation task. The blue points represent positive observations, i.e. where the expert generated range maps indicate that the species is likely to be found. In contrast, the red points indicate absences.

C.2 ICUN: Range Maps

Task: The goal of this task is similar to the previous one, i.e. to predict the geographical range of a set of species. However, instead of 'ground truth' target maps that are estimated by another model, here we use expert curated range maps from the International Union for Conservation of Nature (IUCN) [3]. This set of data contains a more taxonomically diverse set of species compared to the *Stats and Trends* task, as it contains mammal and fish species, among others, in addition to birds.

Dataset: We select a subset of 3,938 species that overlap with our training species. For each species, we take the expert generated range polygon(s), which indicates the geographical range of a given species, and sample a set of points inside the polygons(s) and denote them as *presence* observations. In addition, we sample another set of points, uniformly at random outside those polygons(s) and denote them as *absence* observations. In total, we sample 2,000 observations per species, half presence and half absence. A visualization of the resulting sampling process is shown in Figure A3.

Note that the absolute performance on this dataset in Table 1 is often well above 90%. We believe this is due to the fact that the absence samples are "easy" because many of them are far away from the presence samples. Alternative sampling processes could have been used, e.g. with more emphasis on the boundary regions between presence and absence areas as denoted by the expert maps. However, as the expert maps cannot necessarily be assumed to be the objective ground truth (i.e. species ranges can shift over time), we instead opted for the strategy outlined above.

Evaluation: Again we use mean average precision (MAP) as the evaluation metric, which results in a single score for a model averaged across all species.

C.3 Geo Prior: Geographical Priors for Computer Vision

Task: The goal of this task is to combine the outputs of the models trained for species range estimation on the iNaturalist dataset with computer vision image classifier predictions. This evaluation protocol has also been explored in other work, e.g. Berg et al. (2014); Mac Aodha et al. (2019). We simply weight the probabilistic image classifier predictions for a given image with the species occupancy predictions from the location where that image was taken. The intuition is that the occupancy prediction down-weights the probability of a given species being predicted by the vision model if the range estimation model predicts that the species is *not* likely to be present at that location.

Dataset: For the vision classifier, we use an image classification model developed by the iNaturalist community science platform [4]. This model is an Xception network Chollet (2017) that has been trained on 55,000 different taxonomic entities (i.e. classes) from over 27 million images. We take the predictions from the final classification layer of the classifier, and do not apply any of their sophisticated taxonomic post-processing. There are a total of 49,333 species in the set of 55,000 classes, the rest are higher order groups, e.g. genera. The images used to train the image classifier come from observations that were added to iNaturalist prior to December 2021.

[3] https://www.iucnredlist.org/resources/spatial-data-download

[4] https://www.inaturalist.org/blog/63931-the-latest-computer-vision-model-updates

We then constructed a test set consisting of all research grade observations (i.e. those where there is a consensus from the iNaturalist community as to what species is present in the image). The images in the test set only contain the set of species that were observed at training time, i.e. we do not consider the open-set prediction problem. The observations where selected from between January and May 2022 to ensure that they did not overlap with the training set. We take at most ten observations per species, which results in 282,974 total observations from 39,444 species. In practice, many species do not have 10 observations. In total there are 2,721 species, with 9,808 total images, that are not present in our range estimation training set. For each of the 282,974 observations, we extract the predictions from the deep image classifier across all 39,444 remaining species.

Evaluation: Performance is evaluated in terms of top-1 accuracy, where the ground truth species label is provided by the iNaturalist community. Without using any information about where an image was taken, the computer vision model alone obtains an accuracy of 75.4%, which increases to 90.4% for top-5 accuracy. During evaluation, if a species is not present in a range model, we simply set the output for the range model for that species to 1.0.

C.4 GEO FEATURE: ENVIRONMENTAL REPRESENTATION LEARNING

Task: This task aims to evaluate how well features extracted from deep models trained to perform species range estimation can generalize to *other* dense spatial prediction tasks. Unlike the other benchmark tasks that focus on using the species occupancy outputs directly, here the final classifier of the model h_ϕ is removed and instead we evaluate how effective the trained location encoder f_θ features are for other downstream environmental prediction tasks. The intuition here is that a model that is effective at range estimation will have learned a good representation of the local environment. If that is the case, that representation should transferable, with minimal adaptation, to other tasks.

This task is inspired by the linear evaluation protocol that is commonly used in self-supervised learning, e.g. Chen et al. (2020). In that setting, the features of the backbone model are frozen and a linear classifier is trained on them to evaluate how effective they are on various downstream classification tasks. In our case, instead of classification, we aim to *regress* various continuous environmental properties from the learned feature representations of our range estimation models. A related evaluation protocol was recently used in Rolf et al. (2021) for the case of evaluating models trained on remote sensing data.

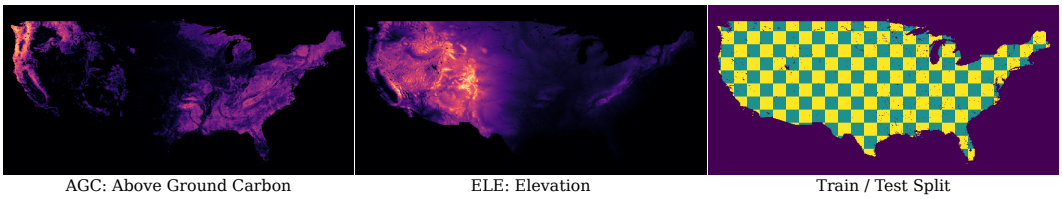
Dataset: The task contains nine different environmental data layers which have been collected using Google Earth Engine [5]. The nine data layers are described in Table A2. For each of the layers, we have rasterized the data so that the entire globe is represented as a 2004×4008 pixel image. Each pixel represents the measured value for a given layer for the geographical region encompassed by the pixel. Example images can be seen in Figure A4.

Evaluation: For evaluation, we crop the region of interest to the contiguous United States and grid it into training and test cells. The spatial resolution of the training and testing cells are illustrated in Figure A4 (right). Note, we simply ignore locations that are not in the training or test sets, e.g. the ocean. This results in 51,422 training points and 50,140 test points. Features are then extracted from the location encoder for the spatial coordinates specified in the training split, and then a linear ridge regressor is trained on the train pixels and evaluated on the held out test pixels. We cross validate the regularization weighting term α of the regressor on the training set, exploring the set $\alpha \in \{0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$. Performance is reported as the coefficient of determination R^2 on the test pixels, averaged across all nine layers.

D REPRODUCIBILITY STATEMENT

The information needed to implement and train the models outlined in this paper are provided in Section A. In addition, the different training losses evaluated are described in Section 3.2. Upon acceptance, we will provide code to enable researchers to build on our results.

[5] https://earthengine.google.com



AGC: Above Ground Carbon | ELE: Elevation | Train / Test Split

Figure A4: Here we illustrate two of the nine evaluation layers used in the Geo Feature prediction task (left and middle). On the right we indicate which regions contain pixels that are in the train or test split. Here, green pixels are train, yellow are test, and purple ones are ignored during both training and test.

Table A2: Description and sources of the nine environmental spatial layers that are part of our 'Geo Feature' prediction task.

Name	**Task**	**Units**	**Range**
AGC	Above ground living biomass carbon stock density of combined woody and herbaceous cover in 2010. `NASA/ORNL/biomass_carbon_density/v1` - `agb`	Mg C/ha	0 to 129
ELE	GMTED2010: Global multi-resolution terrain elevation data 2010. Masked to land only. `USGS/GMTED2010` - `be75`	meters	-457 to 8746
LAI	The sum of the one-sided green leaf area per unit ground area. `JAXA/GCOM-C/L3/LAND/LAI/V2` - `LAI_AVE` - `2020`	(see previous)	0 to 65531
NTV	Percent of a pixel which is covered by non-tree vegetation. `JAXA/GCOM-C/L3/LAND/LAI/V2` - `LAI_AVE` - `20202`	%	0 to 100
NOV	Percent of a pixel which is not vegetated. `MODIS/006/MOD44B` - `Percent_NonVegetated`	%	0 to 100
POD	UN adjusted estimated population density. `CIESIN/GPWv411/GPW_UNWPP-Adjusted_Population_Density` - `unwpp-adjusted_population_density`	# of persons / km^2	0 to 778120
SNC	Normalized difference snow index snow cover. `MODIS/006/MOD10A1` - `NDSI_Snow_Cover` - `2019`	(see previous)	0 to 100
SOM	Soil moisture, derived using a one-dimensional soil water balance model. `IDAHO_EPSCORTERRACLIMATE` - `soil` - 2020	mm	0 to 8882
TRC	The percentage of pixel area covered by trees. `NASA/MEASURES/GFCC/TC/v3` - `tree_canopy_cover` - `2000-2020`	%	0 to 100

E ETHICS

This work makes use of species observation data provided by the iNaturalist community. We explicitly only use data from the public data exports from iNaturalist to ensure that no data, that is not already available publicly, related to sensitive species (e.g. those who are at risk of extinction) is used by our models. As noted in the limitations section in the main paper, extreme care needs to be taken when attempting to interpret any species range predictions from the models presented in this paper. The models have not been evaluated on those types of downstream tasks.