

On the value of images to learn ecology-aware geographic embeddings

Adam Menoud⁴, Maximilien Servajean^{2,3}, Pablo Ubilla Pavez¹, Théo Larcher^{1,3}, and Diego Marcos¹

¹ Inria, Univ Montpellier, France

² AMIS, Université Paul Valéry Montpellier, France

³ LIRMM, Univ Montpellier, France

⁴ EPFL, Lausanne, Switzerland

Abstract. Citizen science data provide both species labels and ground-level images, but their relative ecological value remains unclear. We compare these modalities by pretraining separate contrastive models that align either images or species labels with geographic location, producing location embeddings. We use several taxonomic groups for pretraining. We then evaluate with species classification as a downstream task. Results show that species-based embeddings generally perform better when training and evaluation distributions are closely related, while image-based embeddings capture complementary environmental information and perform better when such correlations are weaker. Overall, the images provide more general ecological context and are a strong alternative especially when labels are unavailable. ⁵

1 Introduction

We are experiencing a rapid and global decline in biodiversity [14]. In order to tackle it in an informed way, we need to map, in an accurate and timely manner, and at large scale, the state of ecosystems. Citizen science observations have emerged as a formidable source of data about the presence of species at a global scale, which can in turn be used to understand ecosystems and their dynamics [7]. In many platforms for citizen science observations, such as iNaturalist and Pl@ntNet, users mainly contribute photographs of the specimens, upon which a taxonomic label is assigned, often by other users, although not all contributions end up having an agreed upon taxon [5,2].

This taxonomic label is what is typically used to inform geospatial models of biodiversity [16,13,22], such as species distribution models (SDMs). This results in discarding information contained in the photographs beyond the taxon of the specimen (e.g. phenology, behaviour, substrate, background, etc.), secondary data that can be seen as “ecologically relevant features that extend beyond the primary information” [26,21,18,9], highlighting the need to better leverage this information [4]. In addition, beyond this type of citizen science observations,

⁵ Code is available at: <https://github.com/Adammenoud/species-and-images-CL>

images can originate from a variety of other sources, such as camera traps [1,19], street view imagery or landscape photography.

Contrastive learning is a promising approach to extract such relevant information from images. It enables different modalities to be embedded into a shared vector space, effectively preserving their mutual information as a form of semantic representation. This approach has been widely applied to remote sensing imagery [15,3]. Its effectiveness has also been demonstrated for species identification from ground-level images [12]. Additionally, several contrastive learning approaches have been developed to extract more than species-level information from ground images, often by incorporating textual data or location as a second modality [8,25,24].

While these methods can theoretically achieve strong performance in extracting information from images, it remains unclear how much information is actually contained in images compared to species occurrence data alone. In [25], the authors use contrastive learning with both text and ground level images, but the use of species labels is not considered. On the other hand, works that use both species and images combine them into a single representation space [12,24], which prevents a direct comparison. Given the widespread use of image data, assessing its added value relative to more basic occurrence information is therefore essential.

The question that we want to answer with this work is the following: **Does learning representations using images from citizen science observations encode more ecological information than learning to predict species occurrences?**

Our approach consists of using contrastive models using either citizen science photos or their corresponding species label, along with their location, allowing us to construct identical and independent pipelines for both image and species label information. We then evaluate these representations using species classification *on a different taxonomic group* as a downstream task, which allows us to assess the value of the learned location embeddings on an ecological task beyond the one used for training.

2 Methodology

2.1 General approach

The main idea to answer our research question is to pretrain two location embeddings in a contrastive manner, one using the images and another one using the species labels. The other modality in both cases consists of the geographic location as latitude-longitude (*latlon*) coordinates. By associating the information to a specific place, the model is effectively asked to build a map of features, that is an encoder from the *latlon* coordinates to the embedding space.

Once these models are pretrained, we evaluate the quality of the location embeddings produced by training lightweight Species Distribution Models (SDMs),

with the task of predicting *different set of species* than the one used for pre-training. This is meant to allow studying the generalization of the learned representation to other, related tasks.

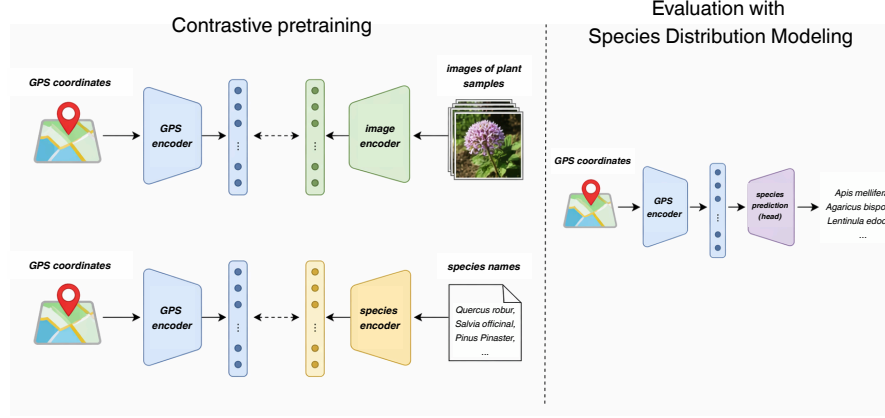


Fig. 1: Summary of the pipeline. We train a different contrastive model for plant images and species labels. We use location as the other modality. We then evaluate and compare the location embeddings by training Species Distribution Models.

2.2 Contrastive learning

Contrastive learning [17] is a learning paradigm that aims to learn meaningful data representations by pulling together paired samples and pushing apart unpaired ones in a latent space. Instead of relying on explicit labels, the model can learn from naturally paired samples and a large number of implicitly negative examples. Positive pairs are typically formed by different views or modalities of the same data instance (e.g., an image and its caption, or two augmented versions of an image), while negative pairs consist of other examples in the mini-batch or dataset.

In our case, the data used for pretraining the model consists of a set of triplets of modalities $D = \{(I_n, G_n, S_n)\}_{n \leq N}$ of N images I_n , geographic latlon coordinates G_n , and species label S_n of the occurrence. We also have the three encoders $\mathcal{I}, \mathcal{G}, \mathcal{S}$ that allow the embedding of each modality into the same vector space $\mathbb{R}^d$. We denote the encoded data as follow:

$$(\mathbf{I}_n, \mathbf{G}_n, \mathbf{S}_n) := (\mathcal{I}(I_n), \mathcal{G}(G_n), \mathcal{S}(S_n))$$

We train two separate instantiations of the location encoder, $\mathcal{G}_{\text{images}}$ and $\mathcal{G}_{\text{species}}$. Each encoder is obtained via contrastive pretraining using an auxiliary modality

X , either images or species labels. For a mini-batch of B location- X pairs, the following symmetric Cross-Entropy loss (or infoNCE loss) is computed [23]:

$$\mathcal{L}_{\text{infoNCE}} := \frac{1}{2} (\mathcal{L}_{\text{location} \rightarrow X} + \mathcal{L}_{X \rightarrow \text{location}})$$

with, for any pair of modalities u, v ,

$$\mathcal{L}_{u \rightarrow v} := -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp\left(\frac{\mathbf{u}_i^\top \mathbf{v}_i}{\tau}\right)}{\sum_{j=1}^B \exp\left(\frac{\mathbf{u}_i^\top \mathbf{v}_j}{\tau}\right)}$$

Here τ is a temperature parameter, that can be fixed or learnt. By pretraining twice, with $X = I$ and $X = S$, we obtain two location encoders $\mathcal{G}_{\text{images}}$, $\mathcal{G}_{\text{species}}$ that we can use on any new location. This allows testing on downstream task involving location information. Note that the new points need to be located on the same geographical domain than the data used for pretraining, since there is no reason for the contrastive model to generalize outside of it.

2.3 Evaluation

The contrastive pretraining provides a pretext task aimed to learn “useful” embeddings, meaning that we need a downstream task to assess this usefulness. We use the embeddings obtained from our model as covariate/input data for the evaluation task. In order to evaluate with respect to a meaningful ecological task, we fit a linear model with the task of predicting species occurrences of a different set of species. Our final metric is then the ROC AUC obtained by the SDM on the test set; AUC is a standard metric in these cases [22].

2.4 Combinations of embeddings

Our method allow us to compare the information contained in images and in the species through the encoders $\mathcal{I}$ and $\mathcal{S}$ that is shared with the location. Therefore, we go further and try different ways to combine these two kinds of embeddings $\mathbf{I}_n$ and $\mathbf{S}_n$ during training. We use BioCLIP 2 as an encoder for images (see Section 3.3 for more details), which has been shown to encode ecological information beyond mere species identity [12]. Therefore, we try taking the difference between the image embedding and the species embedding, to have the model focus only on the orthogonal information. We input this difference information either directly or concatenated to the image embedding information, since we hypothesize that it might encode for independent information. We also try to take the sum (or equivalently, the average) of both embeddings, as some kind of joint embedding. Finally, we also try a simple concatenation, to allow the model to use to both embeddings freely. In all cases, we add (non-linear) layers after the operation.

3 Experiments

3.1 Datasets

We use two different and non-overlapping datasets that span over the same geographic domain, for the pretraining and the evaluation with SDMs respectively. For the **pretraining**, we use the **iNaturalist** dataset from GBIF, that we curate to keep only the entries in metropolitan France. The full occurrence dataset can be downloaded here: [11]. We then divide it into three pretraining subsets:

- Plant ('Plantae' kingdom), with 1.5M occurrences and 6540 unique species
- Arthropods ('Arthropoda' phylum), with 2M occurrences and 13382 unique species
- Mushrooms ('Fungi' kingdom), with 130k occurrences and 2594 unique species

For the dataset used for **evaluation**, we use the **GeoPlant** dataset [22] and also consider only the entries in metropolitan France. We then consider the entries for which all relevant species are present, and keep only the species for which the number of occurrences is above 20 for both Presence-Only (PO) and Presence-Absence (PA) data. There is 84 such plant species, and the number of observation is 1262 for PA (for each species) and 26850 for PO (with one species recorded for one observation). We use the PO occurrences to train the SDM, and the PA occurrences to test it, similarly to [10]. Note that this setup is different from some other SDM experiments on the Geoplant dataset [22], and thus the AUC scores obtained with these methods should not be directly compared to those obtained from our experiments.

3.2 Analysis of the pretraining subsets

If the species used during pretraining are the same as those used during evaluation, then memorizing the species information contained in the training set is sufficient, which could bias the analysis and favor using the species encoder only. To this end, we train the main model on different phyla (or kingdoms). Here, we compare the similarity of these subsets with the original 'Plantae' kingdom, to better interpret the performances of the model.

Mutual Information between datasets The relevant quantity to know whether knowledge about one distribution helps predicting the other is the mutual information. Since the number of species is very high, determining the actual mutual information between two subsets is a hard problem. For this reason, we first rasterize the data and then use a clustering algorithm (k-means) to assign a label to each site. Determining the (adjusted) mutual information between two sets of clusters is then a much easier problem. This gives us quantitative information about how the subsets relate to the plant subset (and also to each other).

We rasterize the territory with cells of $10k$ meters. To each cell we associate the vector that counts the number of occurrences for each species. Then we run

the k-means algorithm on these vectors for $3 \leq k \leq 50$, compute the adjusted mutual information (AMI), and report the results (see Figure 4).

Qualitative analysis of samples In order to have an idea of whether the images contain more information than the species labels alone, we also took a look at a sample of images of each of the subsets and identified the shared information in the surroundings. We plot the clusters obtained by running the clustering algorithm on plant species (see Figure 5), and pick three locations that we can clearly identify as being assigned a different label: the mountains (Alps), the Mediterranean basin and a city (Paris). We download 50 images from these locations and report some of the similarities in the background. This gives information on whether the image modality has access to some shared information that is not present in the species modality.

3.3 Encoders

Since we train a contrastive model, the two important components are the encoders for both modalities, that is, species/images on one hand and location on the other.

In both the cases of species and images, we use respectively the text encoder and the image encoder of BioCLIP 2 [12]. BioCLIP 2 is itself a contrastive model pretrained on a dataset of 200 million images of living organisms. It embeds both the image and the full scientific name of the species observed into the same embedding space, effectively giving a way to compare both modalities. Note that it would be problematic for our analysis if the contrastive pretraining from BioCLIP 2 had collapsed all information from the image to only the species information. The authors of BioCLIP 2 show that it is not the case, e.g. the image encoder performs well when applied to other biological visual tasks such as habitat classification and trait prediction. However, we also use another more general image encoder in order to be able to compare, namely DINOv2 [20]. We keep these encoders frozen and add two final linear layers, with ReLU activations, in order to adapt the model to the specific task, as in [6].

For the location encoder, we take the location encoder from [6] and train it with weights initialized at random. Although more details can be found in the original article, we describe it here briefly. The first step is to take the Equal Earth Projection (EEP) in order to counterbalance the distortions of the usual coordinate system. The coordinates are then encoded from two dimensions into higher dimension using Random Fourier Features (RFF) at several frequencies $\{\sigma_k\}_k$. For each frequency, we compute the component of the vector to encode along many random directions, and then encode these values using sine and cosine. Drawing the random directions from a normal distribution once before any training, the RFF encoding for a given frequency σ_k is the following:

$$\gamma(x) = \begin{bmatrix} \cos(2\pi \mathbf{R}x) \\ \sin(2\pi \mathbf{R}x) \end{bmatrix}^T \quad \mathbf{R}_{i,j} \sim \mathcal{N}(0, \sigma_k) \quad \forall i \leq d/4, j \leq 2$$

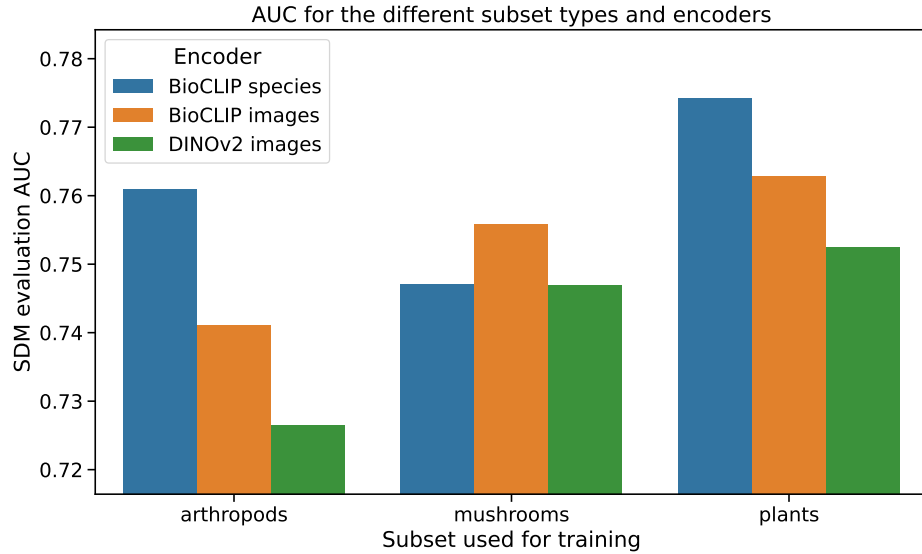


Fig. 2: AUC obtained for all subsets and encoders

with x the $(lat, lon)^T$ coordinate vector after EEP and d the RFF embedding dimension. For each frequency, the vector is then passed through a four-layer MLP, and the results for each frequency are then added to give the final embedding vector.

We noticed that during pretraining the model tends to produce a good quality map after a few epochs, but that the quality seems to drop afterwards, even though the validation loss keeps going down. It seems like the model fits information on the highest frequencies, which produces poor results. We therefore also consider a low-frequency regime as a regularization, and evaluate the models in both settings.

3.4 Implementation details

The contrastive model is trained using 80% of the samples set, and 20% is used for validation. we train for 40 epochs and take the model corresponding to the best performance on the validation set. We train with a batch size of 4096 and a fixed learning rate of 10^{-4} . For the frequencies of the RFF encoding, we use the frequencies $\{1, 2^{-4}, 2^{-8}\}$ for the high frequency regime and $\{1, 2^{-4}\}$ for the low frequency regime. For the evaluation with SDMs, we consider the maximum AUC score over the two regimes, unless specified otherwise. Note that most often, the low regime gives better performances. We repeat the same procedure on each of the three subsets (plants, mushrooms, and arthropods).

The location encoder consists of a four-layer MLP for each frequency, with hidden dimension 1024 and output dimension 512, with ReLU activation layers.

For the image encoder, the embeddings from both BioCLIP and DINOv2 are of dimension 512. We add two learnable layers of dimension 512 with ReLU activation in between. The temperature parameter is learnt as a free parameter, and initialized at 0.07. For evaluation, we use Poisson regression for the linear models, and train one model for each species, and take the average for the final AUC. When combining the embeddings, we train the models only using the low-frequency setting in order to reduce the number of runs. We concatenate the raw outputs of BioCLIP 2 and then apply the two final layers. When concatenation is applied, the input dimension doubles, but we keep the two final layers at the same size. Other than that, the training procedure is identical to that described previously.

4 Results

4.1 Performance of the contrastive models

Figure 2 shows the results on the GeoPlant benchmark subset. We observe that, for both pretraining subsets containing plants and arthropods, the species embedding gives higher AUC than the BioCLIP 2 image embedding. The opposite is true for the mushroom subset, for which the species embeddings gives poorer results than the BioCLIP 2 image embeddings and comparable to DINOv2. We also observe that the DINOv2 encoder consistently performs worse than BioCLIP 2. Note that it is also possible, for any location, to compute its similarity with a specific image. The similarity is just computed as the dot product of the normalized embeddings (i.e. cosine similarity). This gives a similarity map over the geographic domain [25], that we can use to have visual insight about how the model works, and ensure that the pretraining was successful. Using this process with an image of a plant that grows in the mountains, for instance "Epilobum Angustifolium", we get a similarity map of France that recovers the topology of the territory (see Figure 3).

4.2 Additional analysis of pretraining subsets

Mutual information When evaluating the AMI for several values of k on the three pairs of pretraining subsets, we can see clear trends. The mushrooms subset has a low AMI with both other subsets (see Figure 4). On the contrary, the subset of arthropods is relatively similar to the subset of plants. These differences can also partly be seen of the raster plots; the mushrooms do not seem to have a specific cluster for the Mediterranean basin (see Figure 5), for instance.

Qualitative analysis of sample images In addition to the AMI analysis, we also examine the visual background present in the images (see Figure 6). We can see that, for each location, several recurring elements can be observed across datasets, such as meadows in the Alps, limestone in the Mediterranean and manmade elements in Paris. We hypothesize that the image encoders capture

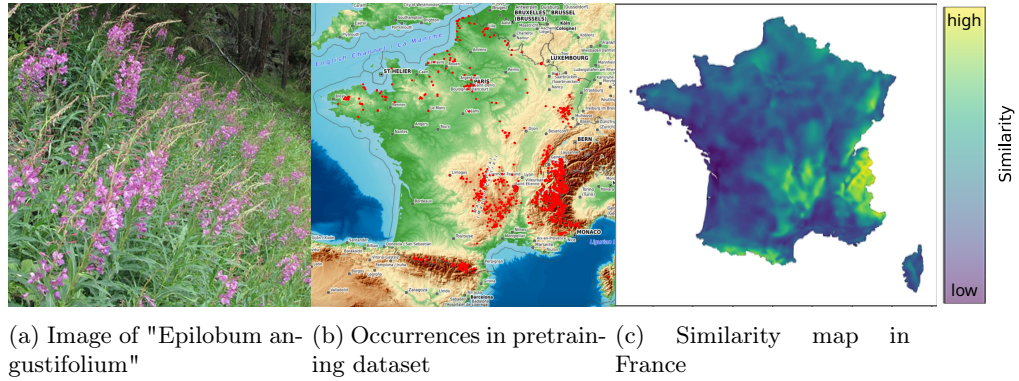
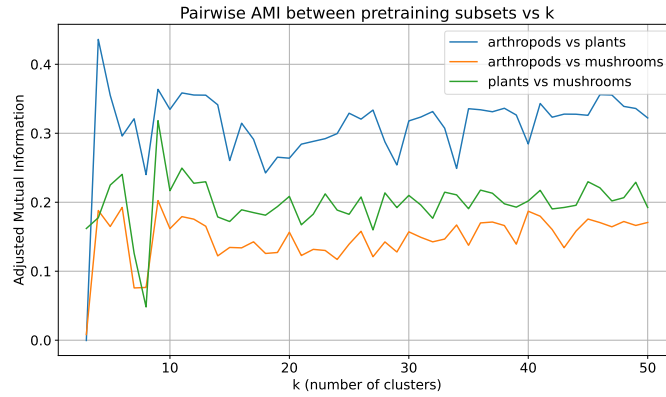


Fig. 3: Occurrences and similarity map comparison

Fig. 4: Ajusted Mutual Information for each number of clusters (with $3 \leq k \leq 50$)

this information. This is discussed by the authors of BioCLIP 2, indicating that the encoder performs well for habitat classification, for instance. Contrastive learning could then incorporate this information in the location embeddings. This could enable the model to perform better with images than with species in the case where species information is correlated with the downstream task.

4.3 Combinations of embeddings

Finally, we explore what happens when we apply operations to combine the different kind of embeddings before pretraining. For both the arthropods and the plant pretraining subsets, in which the species embeddings perform better than the image, we notice that all the different combinations lie in between the two or lower than both 7. This suggests that none of them are a better way to incorporate the two sources than the species or image embeddings alone.

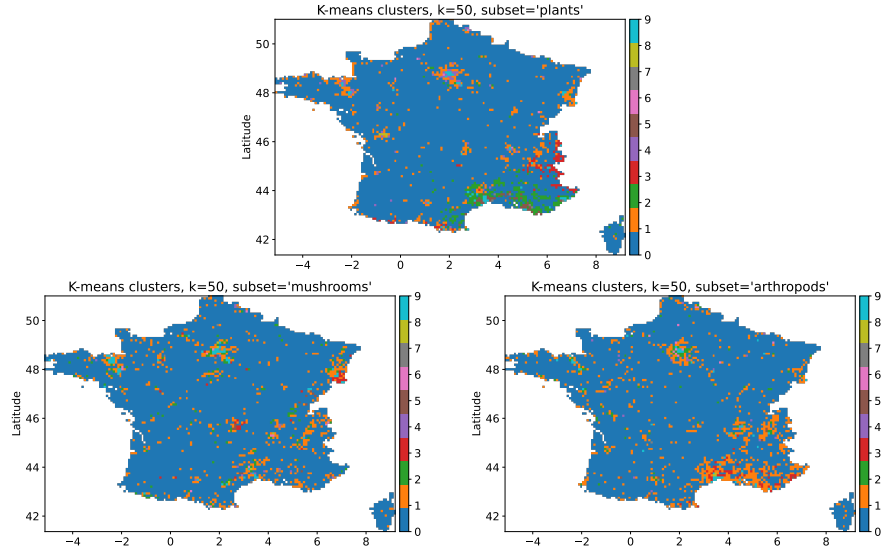


Fig. 5: Cluster assigned to each site for each subset (all 41 clusters with the least cells are aggregated in '9')

5 Discussion

5.1 Interpretation of Results

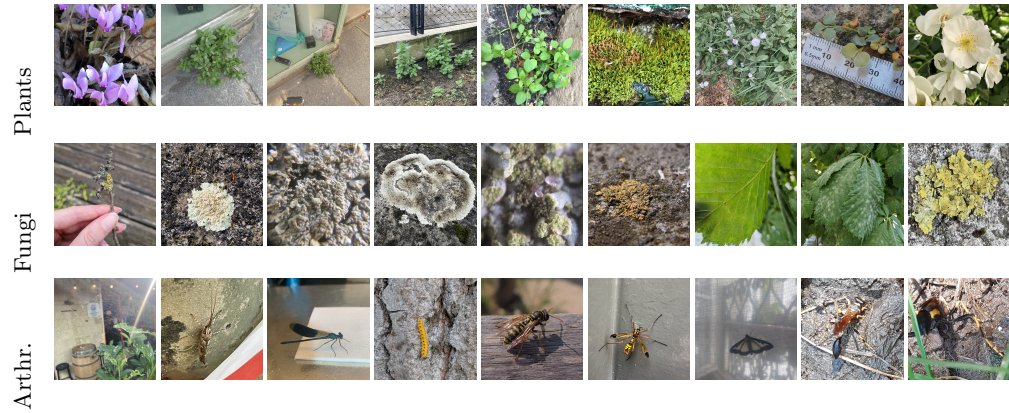
The observed results exhibit strong dataset-dependent differences. For the BioCLIP 2 embeddings, images outperform species-only embeddings only for the mushroom subset. The additional analysis provides a clue about the origin of these differences. The strong performance of species embeddings for plants and arthropods can be explained by their similarity in species distribution. The distribution of species learned during pretraining can be directly exploited by the SDMs, resulting in high AUC values. We also hypothesize that the models capture background information in all three cases. While species information appears sufficient for plants and arthropods, background information seems to play a more important role in the mushroom dataset, where image-based embeddings perform better. Both findings are informative. If the goal is to use contrastive learning to **directly model species distributions**, then using **species occurrences alone** appears to be more effective. If the objective is to **learn general ecological features**, using **images seems preferable**. Additionally, in cases where species information is unavailable [5], images seem sufficient for generating ecological features. Finally, despite these complementary strengths, we do not yet identify an effective strategy to jointly integrate both information sources during pretraining.



(a) Image samples in the Alps.



(b) Image samples in the Mediterranean basin.



(c) Image samples near Paris.

Fig. 6: Randomly selected samples from each dataset stemming from the same spatial cell: one cell in the Alps (a), one in the Mediterranean basin (b) and one within Paris (c).

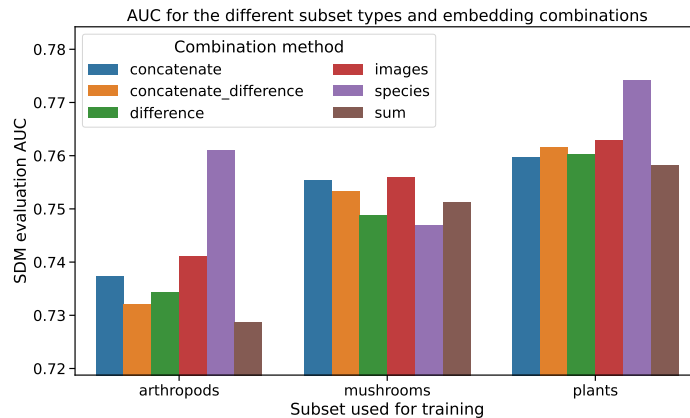


Fig. 7: AUC for all combination of embeddings used for pretraining, per subset

5.2 Limitations

There remains a degree of similarity between the pretraining and downstream evaluation tasks. Although this issue was partially mitigated through the analysis of multiple datasets, our conclusions would be stronger if the pretraining and evaluation tasks were more clearly separated.

We also discuss the information contained in images and species labels at a relatively abstract level. While our framework provides informative comparisons, the specific modalities used to extract ecological information may influence the conclusions. For example, incorporating remote sensing or textual data could alter the relative importance of image-based and species-based representations. Likewise, the results may also depend on the encoders used, here bioCLIP 2. Consequently, the comparisons presented here should be reproduced across a broader range of settings and modalities.

Finally, although the evaluation dataset was carefully curated [22], the pre-training data was obtained directly from iNaturalist and is therefore subject to any bias associated with citizen science data. Figure 8, for instance, highlights substantially stronger sampling in densely populated areas, leading to more occurrences and more species represented.

6 Conclusion

In this work, we compared the ecological information contained in occurrence species labels with that contained the corresponding images. We used contrastive learning with geographic coordinates to extract this information, using three data subsets corresponding to three different taxonomic groups, and evaluated the resulting representations through plant species distribution estimation. We found that the relative performance of images compared to species labels depends

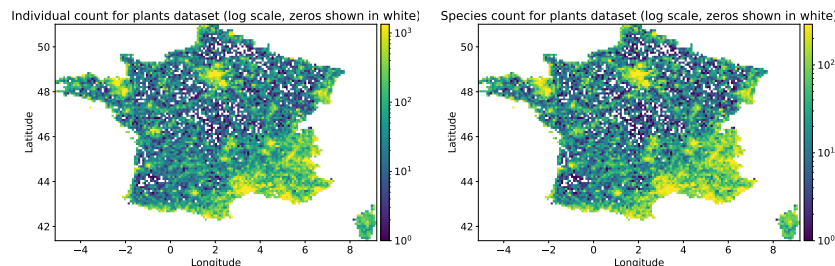


Fig. 8: Occurrence number (left) and number of species (right) for each site

on the similarity between the distributions of the species observations in the pretraining dataset and those in the evaluation task. When the observations are distributed in a dissimilar way, measured as the mutual information between species observation sets, images do perform better than species labels alone. This suggests that making use of images rather than species labels may provide a representation that generalizes better to tasks that are less related to species distribution modelling as captured by the pretraining dataset, although verifying this needs further work on more diverse downstream tasks.

References

1. Born-digital biodiversity data: Millions and billions - Kays - 2020 - Diversity and Distributions - Wiley Online Library, <https://onlinelibrary.wiley.com/doi/10.1111/ddi.12993>
2. Bonnet, P., Champ, J., Goëau, H., Stöter, F.R., Deneu, B., Servajean, M., Af-fouard, A., Lombardo, J.C., Levchenko, O., Gresse, H., Joly, A.: Pl@ntNet Services, a Contribution to the Monitoring and Sharing of Information on the World Flora. *Biodiversity Information Science and Standards* 4 (Sep 2020). <https://doi.org/10.3897/biss.4.58933>
3. Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L.L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data (Jul 2025), <https://arxiv.org/abs/2507.22291v2>
4. Callaghan, C.T., Poore, A.G.B., Mesaglio, T., Moles, A.T., Nakagawa, S., Roberts, C., Rowley, J.J.L., VergÉs, A., Wilshire, J.H., Cornwell, W.K.: Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience* 71(1), 55–63 (Jan 2021). <https://doi.org/10.1093/biosci/biaa131>, <https://doi.org/10.1093/biosci/biaa131>
5. Campbell, C.J., Barve, V., Belitz, M.W., Doby, J.R., White, E., Seltzer, C., Di Cecco, G., Hurlbert, A.H., Guralnick, R.: Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters

- for biodiversity science. *BioScience* **73**(7), 533–541 (Jul 2023). <https://doi.org/10.1093/biosci/biad051>, <https://doi.org/10.1093/biosci/biad051>
6. Cepeda, V.V., Nayak, G.K., Shah, M.: GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization (Nov 2023). <https://doi.org/10.48550/arXiv.2309.16020>, <http://arxiv.org/abs/2309.16020>, arXiv:2309.16020 [cs]
 7. Chandler, M., See, L., Copas, K., Bonde, A.M.Z., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A., Turak, E.: Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* **213**, 280–294 (Sep 2017). <https://doi.org/10.1016/j.biocon.2016.09.004>, <https://www.sciencedirect.com/science/article/pii/S0006320716303639>
 8. Daroya, R., Cole, E., Mac Aodha, O., Van Horn, G., Maji, S.: WildSAT: Learning Satellite Image Representations from Wildlife Observations (Dec 2024), <https://arxiv.org/abs/2412.14428v2>
 9. Davison, A.M., de Koning, K., Taubert, F., Schakel, J.K.: Automated near real-time monitoring in ecology: Status quo and ways forward. *Ecological Informatics* **89**, 103157 (Nov 2025). <https://doi.org/10.1016/j.ecoinf.2025.103157>, <https://www.sciencedirect.com/science/article/pii/S1574954125001669>
 10. Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., Ford, A., Guisan, A., Hijmans, R.J., Huettmann, F., Lohmann, L., Loiselle, B., Moritz, C., Overton, J., Peterson, A.T., Phillips, S., Richardson, K., Williams, S., Wiser, S.K., Wohlgemuth, T., Zimmermann, N.E.: Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics* **15**(2), 69–80 (Jul 2020). <https://doi.org/10.17161/bi.v15i2.13384>, <https://journals.ku.edu/jbi/article/view/13384>
 11. GBIF.org User: Occurrence Download (2026). <https://doi.org/10.15468/DL.PXXRF9>, <https://www.gbif.org/occurrence/download/0003002-260119153935675>
 12. Gu, J., Stevens, S., Campolongo, E.G., Thompson, M.J., Zhang, N., Wu, J., Kopanav, A., Mai, Z., White, A.E., Balhoff, J., Dahdul, W., Rubenstein, D., Lapp, H., Berger-Wolf, T., Chao, W.L., Su, Y.: BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning (May 2025), <https://arxiv.org/abs/2505.23883v2>
 13. Humphreys, J.M., Murrow, J.L., Sullivan, J.D., Prosser, D.J.: Seasonal occurrence and abundance of dabbling ducks across the continental United States: Joint spatio-temporal modelling for the Genus *Anas*. *Diversity and Distributions* **25**(9), 1497–1508 (2019). <https://doi.org/10.1111/ddi.12960>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.12960>, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.12960>
 14. IPBES: Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Bonn, Germany (may 2019). <https://doi.org/10.5281/zenodo.3831673>, <https://doi.org/10.5281/zenodo.3831673>
 15. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rufwurm, M.: SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery (Nov 2023), <https://arxiv.org/abs/2311.17179v3>
 16. Kumar, A., Sinha, A., Kanaujia, A.: Using citizen science in assessing the distribution of Sarus Crane (*Grus antigone antigone*) in Uttar Pradesh, India. *International Journal of Biodiversity and Conservation* **11**(2), 58–68 (Feb 2019). <https://doi.org/10.5897/IJBC2018.1245>, <http://academicjournals.org/journal/IJBC/article-abstract/4DE29E260077>

17. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive Representation Learning: A Framework and Review (Oct 2020). <https://doi.org/10.1109/ACCESS.2020.3031549>, <https://arxiv.org/abs/2010.05113v2>
18. Marques, N., Soares, C.D.d.M., Casali, D.d.M., Guimarães, E.C., Fava, F.G., Abreu, J.M.d.S., Moras, L.M., da Silva, L.G., Matias, R., de Assis, R.L., Fraga, R., Almeida, S.M., Lopes, V.G., Oliveira, V., Missagia, R., Carvalho, E.C., Carneiro, N.J., Alves, R., Souza-Filho, P., Oliveira, G., Miranda, M., Tavares, V.d.C.: Retrieving biodiversity data from multiple sources: making secondary data standardised and accessible. *Biodiversity Data Journal* **12**, e133775 (2024). <https://doi.org/10.3897/BDJ.12.e133775>
19. Oliver, R.Y., Iannarilli, F., Ahumada, J., Fegraus, E., Flores, N., Kays, R., Birch, T., Ranipeta, A., Rogan, M.S., Sica, Y.V., Jetz, W.: Camera trapping expands the view into global biodiversity and its change. *Philosophical Transactions of the Royal Society B: Biological Sciences* **378**(1881), 20220232 (May 2023). <https://doi.org/10.1098/rstb.2022.0232>, <https://doi.org/10.1098/rstb.2022.0232>
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024). <https://doi.org/10.48550/arXiv.2304.07193>, <http://arxiv.org/abs/2304.07193>, arXiv:2304.07193 [cs]
21. Pernat, N., Canavan, S., Golivets, M., Hillaert, J., Itescu, Y., Jarić, I., Mann, H.M.R., Pipek, P., Preda, C., Richardson, D.M., Teixeira, H., Vaz, A.S., Groom, Q.: Overcoming biodiversity blindness: Secondary data in primary citizen science observations. *Ecological Solutions and Evidence* **5**(1), e12295 (2024). <https://doi.org/10.1002/2688-8319.12295>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/2688-8319.12295>,
_eprint: <https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/2688-8319.12295>
22. Picek, L., Botella, C., Servajean, M., Leblanc, C., Palard, R., Larcher, T., Deneu, B., Marcos, D., Bonnet, P., Joly, A.: GeoPlant: Spatial Plant Species Prediction Dataset (Aug 2024), <https://arxiv.org/abs/2408.13928v2>
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (Feb 2021), <https://arxiv.org/abs/2103.00020v1>
24. Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., Jacobs, N.: TaxaBind: A Unified Embedding Space for Ecological Applications (Nov 2024). <https://doi.org/10.48550/arXiv.2411.00683>, <http://arxiv.org/abs/2411.00683>, arXiv:2411.00683 [cs]
25. Sastry, S., Khanal, S., Dhakal, A., Huang, D., Jacobs, N.: BirdSAT: Cross-View Contrastive Masked Autoencoders for Bird Species Classification and Mapping (Oct 2023), <https://arxiv.org/abs/2310.19168v1>
26. Vendrow, E., Chae, J., Kurinchi-Vendhan, R., Eckert, I., Hall, J., Jarzyna, M., Miyajima, R., Oliver, R., Pollock, L., Shrack, L., Yanco, S., Aodha, O.M., Beery, S.: INQUIRE-Search: A Framework for Interactive Discovery in Large-Scale Biodiversity Databases (Dec 2025). <https://doi.org/10.48550/arXiv.2511.15656>, <http://arxiv.org/abs/2511.15656>, arXiv:2511.15656 [cs]