Contrastive Representation Learning for Natural World Imagery: Habitat prediction for 30,000 species

Sachith Seneviratne[1]

[1]*Transport, Health and Urban Design Research Lab, Melbourne School of Design, The University of Melbourne, Parkville VIC 3010, Australia*

Abstract

Recent work in contrastive representation learning has pushed the boundaries of classification tasks in computer vision, achieving state of the art results on many established benchmarks. However, their performance on natural imagery tasks which fall into the category of fine-grained image classification can be further improved. In this paper, I present a methodology that explores this issue and achieves state of the art results on species distribution modelling from remote sensing imagery as part of the GeoLifeCLEF2021 challenge. My method is able to beat the current state of the art on this challenge (trained on 4 types of imagery) using only base RGB imagery. Initial experiments indicate that modifying the architecture to include additional image modalities leads to further improvements in performance on the task of location-based species recommendation. Additionally, I introduce a consistency function, which relies on the strategy of withholding data from the model and is useful in checking for model generality without relying on a validation split.

Keywords

Fine Grained Visual Categorization, Representation Learning, Self Supervision, Transfer Learning, Domain adaptation

1. Introduction

Species Distribution Modelling (SDM) is the study of computational techniques to predict species distribution across both geographical locations and time using different forms of environmental data. Computer vision techniques have garnered attention in this area due to the ability to effectively incorporate contextual and geographic information to improve the modelling of species distribution[1]. Advances in this area have many implications in ecological analysis including the ability to more effectively engage with citizens regarding wildlife preservation and education[2]. Methods based in computer vision that allow large datasets of habitat imagery to be processed in order to generate a prediction of the most likely species inhabiting that area allow for significant theoretical and applied improvements in this area. However, the key challenges on this problem from a classification-based computer vision perspective are two fold: unbalanced data and having classes with only minute differences to distinguish one from another.

Imagery in the environment can be broadly divided into two categories: **built** and **natural**

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
sachith.seneviratne@unimelb.edu.au (S. Seneviratne)
/0000-0001-9094-2736 (S. Seneviratne)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-2936/paper-140.pdf

world. Remote sensing datasets will generally contain imagery pertaining to both these types. Many challenging tasks in computer vision arise in the natural world imagery domain[3]. Such tasks usually fall under the domain of "fine-grained visual categorization" - an active area of research in computer vision.

Imagery based classification problems with a fine distinction between classes can be difficult for computer vision techniques to perform robustly on, especially when combined with a large number of classes featuring unbalanced data and certain classes being heavily under-represented. This is termed the "long-tailed class distribution" problem. These difficulties are present in the classification problem explored in this paper, where using a satellite image of a habitat location, the species that inhabits that location must be predicted from a list of over 30,000 candidate species. In contrast to standard classification problems, the target candidate for classification is absent within the image in this particular task.

Contrastive representation learning techniques have been extensively explored for classification problems. However, their performance on representation learning across different data domains is less well understood[4]. This work contributes to the body of existing literature exploring self-supervised representation learning methods on remote sensing imagery and related data sources. These include methods exploring the performance of existing self-supervised methods on remote sensing data[5], self-supervision techniques which exploit location and time invariance of remote sensing data to perform representation learning[6] and methods which exploit the spatiotemporal structure of remote sensing data to perform self-supervision[7].

In this paper, I detail my workflow for the winning submission to GeoLifeCLEF2021 and summarize my performance representing the University of Melbourne at this challenge. This competition[1] was organized as GeoLifeCLEF 2021[8], as part of LifeCLEF 2021[9] and in conjunction with FGVC8[2] workshop at CVPR[3] 2021. Comparisons of results are made primarily with existing benchmarks which include the state of the art for this problem. A comparison with other competitors is not included. Additionally, I explore the details around the transformations pipeline used for improving the feature representation learned by the model and also introduce a consistency-based model selection function. This function was useful for the purpose of model selection for evaluation on the public leaderboard. This work is derivative of a larger computer vision framework connecting aspects of the environment (built and natural). This framework draws high-level inspiration from [10] and the insights gained from both projects allowed a winning solution to be crafted for this problem. Further discussion of such insights is beyond the scope of this paper.

2. Data and Evaluation Metrics

In this section, I explore the datasets and evaluation metrics that are used for the purposes of training and evaluating my models. An overall description of all datasets used in this work is also presented, as my workflow only uses either one or two of the available datasets for training and evaluation. Top-30 error is used for comparing different methods. Detailed discussion of the

[1] https://www.kaggle.com/c/geolifeclef-2021

[2] https://sites.google.com/view/fgvc8

[3] IEEE/CVF Conference on Computer Vision and Pattern Recognition - http://cvpr2021.thecvf.com/

metrics used in the competition can be found in [11] with a detailed discussion of the datasets present in [12].

2.1. Dataset

This work builds upon the following types of imagery:

- RGB remote sensing imagery
- Altitude imagery

These imagery types have a pixel-wise correspondence in terms of geographical overlap at each location and are 256x256 in size and have a spatial resolution of 1 meter per pixel. Therefore each image covers an area of 256x256 square meters. Altitude imagery was derived using elevation data from the NASA Shuttle Radar Topography Mission[4]. RGB remote sensing imagery was from 2 sources: **in the US** - from the 2009-2011 cycle of the National Agriculture Imagery Program[5] and **in France** - imagery from BD ORTHOR 2.0 and ORTHO HRR 1.0 databases from the French National Institute of Geographic and Forest Information[6].

2.2. Class Distribution

One of the main difficulties in this problem arises due to the unbalanced class distribution. Interestingly, over 60% of all classes have fewer than 10 training images and nearly 8,000 classes have a single image to train on (about 25% of all classes). The data distribution shown by Figure 1, which shows the number of training records on the x-axis as a closed interval on a discretized logarithimic scale and the number of classes that belong to that range on the y-axis.

2.3. Consistency-based Model Selection Metric

In this section, I introduce a metric which was used in lieu of a validation split on this problem. I use the proxy task of "country prediction" in order to derive an additional validation metric building on the "city prediction" task introduced in [13]. Given that many of the species were endemic to each country (US or France but not both), it is reasonable that a model with higher accuracy in terms of species prediction would be also be able to perform better on the pseudo-task of predicting which country. This is derived from the model's understanding of which species can belong to a particular country. An error rate is calculated for each model corresponding to how many times the model makes an impossible prediction by assigning a species to a country that does not host that species (based on the training data). Note that this consistency only makes sense with the "variable-withholding" strategy described in Section 3, since if the model has access to any geographical information(GPS co-ordinates or country label), it would simply learn this information and not make such mistakes. By intentionally withholding such information from the model I gain two advantages:

[4] https://lpdaac.usgs.gov/products/srtmgl1v003/
[5] https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/
[6] https://geoservices.ign.fr

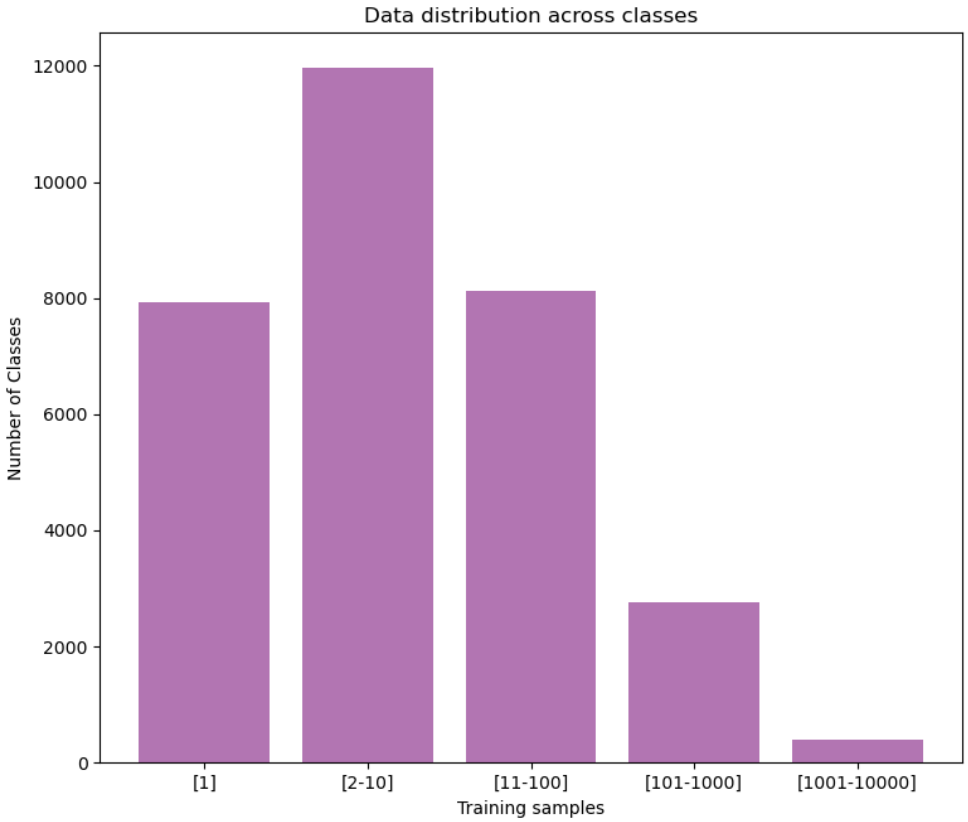

Figure 1: Training dataset distribution. Most classes are heavily under-represented in the dataset.

- I am able to use this consistency error as a pseudo-validation metric.
- It is possible to incorporate withheld-data at a later stage of model training (for example during ensembling of individual models trained on all co-variates) in order to further improve model performance.

The calculation of this function is straightforward:

1. For each species categorize them as "fr", "us" or "both" depending on country of occurrence
2. At validation time, for each predicted label in the top-30 predictions for a particular image do the following:
    - Count the number of "US" species : N_{US}
    - Count the number of "FR" species : N_{FR}
    - Count the number of "US and FR" : N_{BOTH}

3. Count the number of instances where both $N_{US} > 0$ and $N_{FR} > 0$

This count acts as the "confounder" count (or misclassification count) for that model variant where models with fewer confounders are better. This metric was used for model checkpoint selection for submission to the leaderboard, but its effectiveness requires further exploration with respect to performance against an actual validation split of the data.

3. Methodology

The main problem explored in this paper is the overlapping value of the different data sources provided as part of the competition. Since the prediction problem was quite difficult, I focused on approaches that allowed the model to exploit all possible information present in each individual image type, starting with RGB imagery. I explore the following questions in this regard:

- Given that the data consists of base imagery (RGB) augmented by 3 co-variates (NIR, land-use and altitude) at the same location, is it possible to derive most of the information present in all 4 data types using only the base RGB imagery?
- Given the above is achievable, what further information regarding the prediction variable can be extracted from the co-variates?
- What is the best way to combine this information to improve prediction performance?

3.1. Transformations

Image transformations have often been touted as a means of providing more variety to the training process. As the input data used for training neural networks is often fixed, it can lead to the model seeing the same data epoch upon epoch leading to overfitting. This is especially true in fine-grained visual categorization problems with poorly represented classes (<10 images per class) making up the majority. In such cases approaches such as adverserial training and image transformations/augmentations have been shown to provide significant improvements on baseline methods. In this section I explore the image augmentation strategy that was used to combat overfitting. A discussion of modifications for multimodal analysis can be found under Section 3.3.

The transformation pipeline is as follows:

- Subtracting the per-channel ImageNet[14] mean and dividing by the per-channel ImageNet standard deviation.
- Random horizontal flip
- Random vertical flip
- RandAugment[15] was used to augment images N, M with hyperparameters N set to 2, and M set to 9. N represents the number of augmentation transformations to be applied, while M controls the magnitude for all the transformations.

3.2. Unimodal Analysis

In order to explore the possibility of extracting more information from the base RGB imagery, the initial experiment focused on creating a workflow that uses only RGB imagery and ignores **all** other information available to the model for training and evaluation purposes. This includes co-variate images, geographic (GPS) location, country tag and environmental feature vectors. Additionally, past work [16] indicates the benefits of using pretrained feature representations for fine-grained visual categorization tasks. MoCo[17] was used as a contrastive representation learning framework to initialize a feature representation for the model to build off of with pretraining carried out for 20 epochs using a single 4 GPU node on Spartan[18] using the hyperparameters in Table 1. The standard protocol for pretraining was followed, but combining all data across the US and France to form a combined representation, which is required for the combined (both countries at the same time) modeling approach followed in this paper. Further training was conducted for 7 epochs in a supervised manner to finetune the feature representation further. This training was performed with end-to-end finetuning of the ResNet50 using the parameters available in Table 2. Checkpoints were generated each epoch and the model with the lowest consistency error (as defined in section 2.3) was used to determine the best performing model.

Table 1
Representation Learning Parameters

Parameter	Value	Comments
Architecture	ResNet50	Smaller backbone for faster training
Batch size	128	
Learning rate	1.5e-2	
Softmax temperature	0.2	

Table 2
Training Parameters

Parameter	Value	Comments
Framework	PyTorch[19]	
Architecture	ResNet50	Same as above
Batch size	128	
Learning rate	1e-3	

3.3. Multimodal Analysis

In this section I explore how multimodal imagery was incorporated into the training workflow. Only the addition of altitude imagery is covered in this section with the other co-variates being left as future work for exploration. This section uses same workflow as in Section 3.2 with a few key differences.

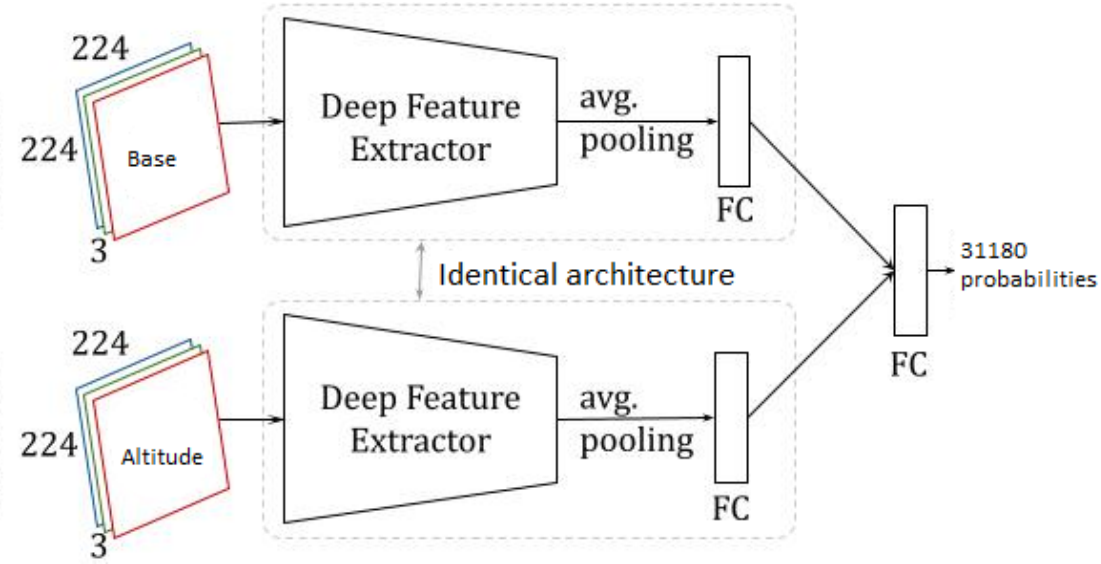

Figure 2: Generic Architecture applicable to this problem. Resnet50 is used as the Deep Feature Extractor and the unimodal workflow only uses the top branch for training and analysis

Pretraining using MoCo was carried out on altitude imagery as well, using an architecture identical to the bottom branch in Figure 3.3. The architecture was modified to include an identical architectural sister network as in the unimodal analysis, which was combined using concatenation at the final bottleneck layer of the ResNet50. The new layer containing 4096 nodes had a 31180 node linear layer with softmax applied in order to infer labels for the task at hand. In this regard, the architecture, which is shown in Figure 3.3, was identical to the unimodal case with the key difference being the number of inputs to the linear layer (multimodal - 4096 vs unimodal - 2048). The single altitude channel was replicated across 3 channels to be compatible with a standard ResNet-50. An advantage of this architecture is its extensibility to different image modalities with the added ability to create seperate filters for the individual image modalities and thereby combine higher level features rather than lower level features (which was the main reason for stacking near the end of the ResNet50 architecture as opposed to near the beginning). My intuition in doing so is that the architecture is able to process more refined knowledge about the different image domains instead of trying to learn an embedding that attempts to unify its representation of all domains combined. This has the marked disadvantage of increasing the GPU memory footprint of the architecture which significantly impacts training time and is perhaps the key weakness of this approach. The batch size was lowered to 64 to accommodate the larger architecture, leading to a roughly 3-fold slowdown on training the model. End-to-end finetuning of the ResNet50 was only conducted for 4 epochs because of these additional computational requirements. A Siamese network based representation learning approach based on the approach from [20] (where weights are shared between the branches, thereby reducing the model footprint on the GPU) was considered but quickly discarded on the basis that the image domains in this problem are too different to each other to benefit from

shared knowledge from each other at the filter levels.

One other key difference was the modification of the transformations pipeline to remove most augmentations during training. This is primarily an artefact of the implementation, which used two seperate PyTorch Dataloaders instead of a single dataloader. Therefore, horizontal and vertical flipping and other transformations would occur independently of each other, impacting the overall correspondence of the image patches due to not having the same orientation. Therefore, all transformations other than normalization (using ImageNet statistics) were removed from the dataloaders.

4. Results

Several methods (including a random-forest based approach) were compared using prior work in this area. More details around low-level implementation details of these benchmarks can be found in [21]. The multimodal approach is able to beat existing supervised techniques by a considerable margin, while the unimodal implementation shows equivalent performance to the existing state of the art. In the results featured in table 3, public leaderboard and private leaderboard performance is indicated, with a 10% vs 90% data split respectively.

Table 3
Results of Top-30 error rate across compared models

Method	Public leaderboard	Private leaderboard
Random Forest	0.78325	0.79711
Supervised CNN(multimodal)	0.75283	0.76680
Mine (unimodal)	0.75726	0.75188
Mine (multimodal)	**0.73679**	**0.74838**

5. Future Work

While initial analysis on this problem is promising, there are many research directions still open to exploration.

The impact of transformations was not fully explored in this work. For multi-modal analysis, a better implementation may be to ensure all transformations are consistently applied across all data sources, so that the image patches propagated through the neural network correspond to the exact same geographic region (which is not the case when the transformations are applied independently across data sources). While the consistency metric introduced in this work was useful for model selection, further comparison with standard validation splits would be useful in further evaluating its utility on this problem. Due to the absence of key ablations, it is unclear where some of the performance gains are being derived, and future work could shed further light on this issue. Additionally, for the consistency function introduced in this work, it is possible that certain species may inhabit nearly identical habitats across both geographies, which may affect the broader usability of this function in different situations.

6. Conclusion

In this paper, I have presented a workflow for achieving state of the art results on computer vision based SDM. I have introduced a consistency-based model selection function that relies on the strategy of withholding information from the models during the training process in order to improve performance. Additionally, this work pushes the boundaries of using contrastive visual representation learning on remote sensing imagery: an area which is currently under-represented in research literature. This paper makes a significant contribution to the area of finely grained visual categorization. My methods are able to surpass the current state of the art using only a quarter of the data used by the current state of the art supervised work in this area, using only a single data modality whereas the current state of the art uses 4. I have also presented initial work on future research directions and provide a methodology and initial results for including further image modalities to drive increased model performance.

Acknowledgments

This project is supported by National Health and Medical Research Grant GA80134. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This research was undertaken using University of Melbourne Research Computing facilities established by the Petascale Campus Initiative.

References

[1] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, A. Joly, Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment, PLoS computational biology 17 (2021) e1008856.

[2] P. Bonnet, A. Joly, J.-M. Faton, S. Brown, D. Kimiti, B. Deneu, M. Servajean, A. Affouard, J.-C. Lombardo, L. Mary, et al., How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools, Ecological Solutions and Evidence 1 (2020) e12023.

[3] G. V. Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, O. M. Aodha, Benchmarking representation learning for natural world image collections, 2021. arXiv:2103.16483.

[4] E. Cole, X. Yang, K. Wilber, O. M. Aodha, S. Belongie, When does contrastive visual representation learning work?, 2021. arXiv:2105.05837.

[5] V. Stojnić, V. Risojević, Self-supervised learning of remote sensing scene representations using contrastive multiview coding, 2021. arXiv:2104.07070.

[6] O. Mañas, A. Lacoste, X. G. i Nieto, D. Vazquez, P. Rodriguez, Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data, 2021. arXiv:2103.16607.

[7] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, S. Ermon, Geography-aware self-supervised learning, 2020. arXiv:2011.09980.

[8] T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Joly, Overview of geolifeclef 2021: Predicting

species distribution from 2 million remote sensing images, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, 2021.
[9] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, R. Ruiz De Castañeda, I. Bolon, H. Glotin, R. Planqué, W.-P. Vellinga, A. Durso, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of lifeclef 2021: a system-oriented evaluation of automated species identification and species distribution prediction, in: Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), 2021.
[10] S. H. Seneviratne, Automatic Code Generation for Statistical Models with Augmentation and Collapsing, Ph.D. thesis, Monash University, 2020.
[11] B. Deneu, T. Lorieul, E. Cole, M. Servajean, C. Botella, P. Bonnet, A. Joly, Overview of lifeclef location-based species prediction task 2020 (geolifeclef), in: CLEF 2020, 2020.
[12] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, The geolifeclef 2020 dataset, arXiv preprint arXiv:2004.04192 (2020).
[13] M. Stevenson, J. Thompson, T. H. de Sá, R. Ewing, D. Mohan, R. McClure, I. Roberts, G. Tiwari, B. Giles-Corti, X. Sun, et al., Land use, transport, and population health: estimating the health benefits of compact cities, The lancet 388 (2016) 2925–2935.
[14] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
[15] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, 2019. arXiv:1909.13719.
[16] M. G. Krishnan, Impact of pretrained networks for snake species classification (2020).
[17] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020. arXiv:2003.04297.
[18] L. Lafayette, G. Sauter, L. Vu, B. Meade, Spartan performance and flexibility: An hpc-cloud chimera, OpenStack Summit, Barcelona 27 (2016).
[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, arXiv preprint arXiv:1912.01703 (2019).
[20] S. Seneviratne, N. Kasthuriaarachchi, S. Rasnayaka, Multi-dataset benchmarks for masked identification using contrastive representation learning, 2021. arXiv:2106.05596.
[21] B. Deneu, M. Servajean, P. Bonnet, F. Munoz, A. Joly, Participation of lirmm/inria to the geolifeclef 2020 challenge (2020).