Combining Present-Only and Present-Absent Data with Pseudo-Label Generation for Species Distribution Modeling

Notebook for the LifeCLEF Lab at CLEF 2024

Yi-Chia Chen[1], Tai Peng[1], Wei-Hua Li[1] and Chu-Song Chen[1]

[1] *National Taiwan University, Taipei, Taiwan.*

Abstract

Predicting the composition of plant species at specific times and locations is crucial for biodiversity management and conservation. In this report, we leverage data from the GeoLifeCLEF 2024 challenge, which includes approximately 5 million plant occurrence records from Europe, a training set of about 90,000 plots, and a test set with 5,000 plots. These data encompass various modalities, including satellite images, climatic time series, land cover, human footprint, bioclimatic, and soil variables. Our approach combines a pseudo-label training framework based on large-scale data and multimodal pretrained deep learning models to address challenges such as multi-label learning from single positive labels, strong class imbalance, and large-scale data processing. On the private test set, our method achieved a score of 0.36837, securing second place on the leaderboard, just 0.04 points behind first place. We discuss the design of our approach and reflect on the results. Our code is available on GitHub.

Keywords

Species distribution modeling, Presence-Only data, Pseudo labels, LifeCLEF, multimodal deep learning

1. Introduction

Species distribution modeling (SDM) is a field of research focused on predicting species that are most likely to be observed at a given location and time. In recent years, the research community has collected a vast amount of species observations from various regions, providing us with the opportunity to train deep learning model to predict species distribution.

In GeoLifeCLEF 2024 [1, 2], a large-scale training set is provided, but most of these samples have only single or partial positive labels (about 5 million of Presence-Only (PO) data and only 90,000 of Presence-Absence (PA) data with exhaustive labels). Therefore, how to effectively integrate these PO data with PA data is one of the challenging problems.

In this report, we propose a hybrid model that combines different CNN-based architectures for SDM. Furthermore, we will introduce the framework we employed during the competition, which effectively utilized the abundant PO data provided by the organizers to generate pseudo-labels. These pseudo-labels were sequentially integrated with PA data to finetune our models.

The rest of this report is structured as follows. Section 2 reviews related work. Section 3 provides a detailed description of the dataset and the evaluation metric for the competition. Section 4 introduces the proposed method. Section 5 presents the experimental results and ablation study. Finally, section 6 concludes the report.

2. Background and Related works

This section provides a brief overview of relevant works in the field of Single-Positive Multi-Label Learning and SPMLL for Species Distribution Modeling.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

r11922a04@csie.ntu.edu.tw (Y. Chen); d11944016@csie.ntu.edu.tw (T. Peng); d12922009@csie.ntu.edu.tw (W. Li); chusong@csie.ntu.edu.tw (C. Chen)

2.1. Single-Positive Multi-Label Learning, SPMLL

Multi-label learning (MLL) [3] has had many practical applications in the past, aiming to enable models to classify multiple labels. However, collecting large amount of the training data with complete multi-label annotation is quite difficult and time-consuming, therefore, SPMLL have been proposed to alleviate the burden of multi-label annotation.

Different from MLL, the goal of SPMLL is to achieve multi-label learning through the samples are annotated with only single positive label. In the field of computer vision, some works proposed to utilize pseudo-label generation. For example, Zhou *et al.* [4] proposes entropy-maximization (EM) loss and asymmetric pseudo-labeling. Xie *et al.* [5] proposed Label-Aware global Consistency (LAC) regularization. Liu *et al.* [6] provides a theoretical guarantee for learning from pseudo-label on SPMLL and proposes MIME, which can simulataneously train the model and update the pseudo-labels. Although these methods are quite effective, they are not specifically designed for Species Distribution Modeling research, and the number of categories they need to predict is smaller.

2.2. SPMLL for Species Distribution Modeling

In GeoLifeCLEF 2022, several CNN-based SDM models [7, 8] have been proposed for species distribution modeling. However, training CNN-based models for multi-label prediction tasks using samples with single positive labels is challenging, therefore, in GeoLifeCLEF 2023, Ung *et al.* [9] proposed the three-steps training strategy. The three-step training process involves using PA data with BCELoss for pre-training the model, then using PO data with Cross Entropy Loss for extensive training, and finally fine-tuning with PA data. Inspired by this work, we have also designed a three-steps process, aiming to make good use of single-label PO data to assist us in performing species distribution modeling.

3. Data and Evaluation Metric

In this section, we introduce the multimodal dataset provided by the GeoLifeCLEF 2024 competition and the evaluation metrics used for the competition.

3.1. Data

The GeoLifeCLEF 2024 Challenge aims to predict plant species presence at specific locations using various related features based on the GeoLifeCLEF 2023 multimodal dataset [10]. The dataset encompasses 38 European countries, covering eight biogeographic regions: Alpine, Atlantic, Black Sea, Boreal, Continental, Mediterranean, Pannonian, and Steppic. The data were collected between 2017 and 2021, ensuring a comprehensive temporal and spatial coverage across Europe. The GeoLifeCLEF 2024 includes approximately 10,000 plant species observed through both Presence-Only (PO) and Presence-Absence (PA) surveys. The PO data consists of 5 million records extracted from trusted sources, while the PA data comprises 90 thousand surveys conducted by botanical experts.

The dataset incorporates several modalities of data, each providing unique insights into the environmental conditions affecting plant species distribution:

- **Satellite Raster Images**:
  - *Sentinel-2 Images*: These include RGB and Near-Infra-Red (NIR) bands, capturing data over a 1280 meter × 1280 meter area at a 10-meter resolution, formatted into 128 × 128 pixel patches.
  - *Landsat Time Series*: This data spans from 2000 to 2020, offering quarterly median composites of six spectral bands (blue, green, red, NIR, SWIR1, and SWIR2) with a 30-meter resolution.
- **Climatic Data**:
  - *Bioclimatic Rasters*: Nineteen low-resolution rasters describing various climatic variables, such as mean annual air temperature and precipitation, provided as GeoTIFF files with a 30-arcsecond resolution (1 km).

- **Soil Variables**:
  - *Soil-Grids*: Nine low-resolution rasters detailing soil properties like pH, clay content, organic carbon, nitrogen, bulk density, sand, silt, and cation exchange capacity, measured at a depth range of 5 to 15 centimeters.
- **Human Footprint**:
  - Sixteen rasters representing human activities and their pressures on the environment, including population density, road networks, and night-time lights. These data are provided for two time periods (1993 and 2009), allowing for the assessment of changes over time.
- **Elevation and Land Cover**:
  - *Elevation Data*: High-resolution elevation data provided as a single GeoTIFF file with a 1-arcsecond resolution (30 m).
  - *Land Cover*: Multi-band raster files describing land cover classes using classifications like IGBP and LCCS, provided with a resolution of 500 meters.

The dataset matches species observations with different environmental factors commonly used in species distribution modeling, such as climate conditions, soil characteristics, land cover, and human impact. All data are provided at suitable spatial resolutions to support accurate modeling.

3.2. Evaluation Metric

The evaluation metric for the GeoLifeCLEF 2024 competition is the samples-averaged F_1-score, calculated on the test set composed of species Presence-Absence (PA) samples. This metric addresses a multi-label classification problem, providing an average measure of the overlap between the predicted and actual sets of species present at specific locations and times.

The micro F_1-score is computed using the following formula:

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + (\text{FP}_i + \text{FN}_i)/2}$$

In this formula: N is the total number of test PA samples. TP_i (True Positives) is the number of species correctly predicted to be present. FP_i (False Positives) is the number of species incorrectly predicted to be present. FN_i (False Negatives) is the number of species that are present but not predicted.

4. Proposed Method

This section introduces our proposed multimodal deep learning model and pseudo-label training framework.

4.1. Model architectures

To address the multi-label plant species prediction problem, we designed and experimented with a multimodal ensemble neural network model. This model integrates various data sources, including preprocessed tabular data (comprising metadata, Human Footprint, Landcover, and Soil), Landsat and Sentinel satellite imagery, and Bioclimatic Rasters. Each data type is processed by specialized neural networks before being fused for classification. The model architecture is illustrated in Figure 1.

We use Multi-Layer Perceptron (MLP) to extract features for preprocessed high-value tabular data. The input tabular data feature vector first passes through a fully connected layer, transforming the number of features into 1,000 neurons. This is followed by batch normalization [11] to stabilize data distribution and accelerate the training process. The activation function used is ReLU [12], which introduces non-linearity to enhance the model's expressive capability. This structure is repeated three times, resulting in three hidden layers, each containing 1,000 neurons. Finally, a fully connected layer reduces the feature dimension to 512, outputting a 512-dimensional feature vector.

The Landsat data processing module adopts a convolutional neural network structure based on ResNet18 [13]. To fully exploit the rich temporal series of remote sensing data, we modified the ResNet18 architecture to suit the characteristics of Landsat data. Initially, we apply layer normalization [14] to the input Landsat data to stabilize the input data distribution. We then use the ResNet18 model for feature extraction but modify the first convolutional layer to increase the kernel size from the original 3 channels to 6 channels, accommodating the multi-spectral nature of Landsat data. This modification enables the model to capture the rich information within Landsat data better. To simplify the model structure and focus on feature extraction, we removed the max-pooling layer and the fully connected layer from the ResNet18 model, retaining only the convolutional layers for feature extraction. This design ensures the model can efficiently process Landsat data and output high-quality feature vectors.

To effectively utilize the Bioclimatic Rasters data, we designed a deep convolutional neural network structure based on ResNet18 [13], which was modified to suit the characteristics of Bioclimatic Rasters data. Initially, layer normalization is applied to the input data to stabilize its distribution, accommodating the diversity of Bioclim data. We also modified the first convolutional layer of the ResNet18 model, changing it from the default 3 channels to 4 channels. Additionally, we removed the max-pooling layer and the fully connected layer to ensure efficient extraction of features from the Bioclimatic Rasters data.

Handling Sentinel satellite imagery data is crucial for species prediction based on geographic location. Sentinel-2 provides multispectral images, including red, green, blue, and near-infrared (NIR) bands. To leverage these high-resolution multispectral images, we employed a self-supervised pretrained ResNet18 model on the SSL4EO-S12 Earth observation dataset [15]. This approach takes advantage of the off-the-shelf model's learning capability on large-scale datasets, enhancing feature extraction performance. We modified the first convolutional layer of ResNet18 from the default 3 channels to 4 channels to accommodate the four spectral bands of Sentinel data. Specifically, the first convolutional layer was set with a kernel size of 7×7, a stride of 2, and padding of 3, enabling the extraction of more local features while maintaining spatial resolution. To adapt to this modification, we concatenated the convolution kernels without altering the original weight distribution. Through this design, the Sentinel data processing module effectively extracts spatial and spectral features from high-resolution multispectral images, providing rich feature representations for subsequent multimodal feature fusion.

4.2. Pseudo-label training framework

The GeoLifeCLEF 2024 competition provides two types of data, PA (Presence-Absence) and PO (Presence-Only), which exhibit significant differences in scale and quality. Although PO data suffers from biases due to the lack of standardized sampling, its vast volume (approximately five million records) is of immense value for model training, enriching the training data and enhancing prediction accuracy. The effective utilization of PO data is undoubtedly crucial for improving performance in this competition. To this end, we have designed a pseudo-label training framework based on both PO and PA data. This framework comprises three steps, as shown in Figure 2.

In the first step, we train our model with PA data to equip it with the initial ability to classify multiple species. We can assume that each input x from X_{pa} corresponds to a label vector y from $Y = \{y_1, y_2, ..., y_L\} \in \{0, 1\}^L$, where L denotes the total number of classes, $y_i = 1$ represents the species i is present at the given location and $y_i = 0$ otherwise. The primary goal is to find a model (M) that can accurately predict y for each x. To achieve this, we use the common binary cross-entropy (BCE) loss to train the model.

In the second step, we utilize the pretrained model to derive pseudo-labels for each sample in the PO data. Given a sample (x_{po}, y), our model M predicts a label vector $\tilde{y} = \{\tilde{y_1}, \tilde{y_2}, ..., \tilde{y_L}\}$ and the corresponding probability $s = \{s_1, s_2, ..., s_L\}$ for each class based on x_{po}. To enhance the reliability of positive pseudo-labels, we introduce an ignore label ($\varnothing$) and filter positive labels based on their confidence scores. We define two confidence thresholds, T^- and T^+, to aid in generating the final pseudo-labels $y^p = \{y_1^p, y_2^p, ..., y_L^p\} \in \{0, 1, \varnothing\}$, where y_i^p can be expressed as follows:

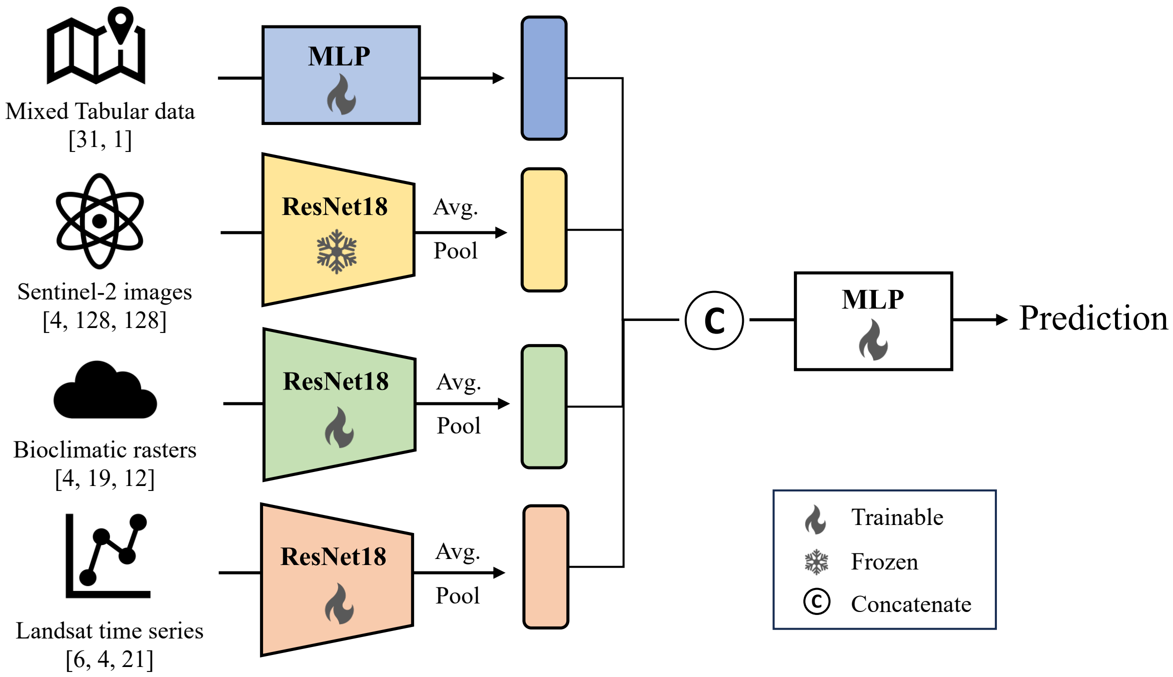

Figure 1: The architecture of the proposed model.

$$y_i^p = \begin{cases} 1, & \text{if } T^+ < s_i \\ \varnothing, & \text{if } T^- < s_i < T^+ \\ 0, & \text{if } s_i < T^- \end{cases} \tag{1}$$

With this filtering mechanism, we can obtain more reliable positive labels, as those with high confidence scores are retained. For those positive labels with uncertain confidence levels (i.e., s_i between T^+ and T^-), we do not include them in the loss calculation. Labels with very low confidence levels are considered negative labels. Additionally, we also retain the original positive samples from the PO data. Therefore, the final pseudo-label can be represented as:

$$y^p = y^p \cup y \tag{2}$$

Finally, we train our model using the PO data with pseudo-labels and the original PA data with multi-labels. This enables the model to undergo training with a larger volume of data. This three-steps process can significantly improve our performance. For related ablation experiments, please refer to section 5.1

5. Experimental Results

In this section, we present the details of the experiments.

Dataset Split: We randomly split the original training set into a training set and validation set in an 8:2 ratio. The training set is used for model training, while the validation set is used for model performance evaluation and hyperparameter tuning. The final model is trained on the complete training set and evaluated on the officially provided test set.

Data Proprocessing: We normalize the tabular data to have a mean of 0 and a standard deviation of 1. We do not use any data augmentation.

Hyperparameters: We use the AdamW [16] optimizer with an initial learning rate of 0.00025 and a weight decay of 0.01. The batch size is set to 64, and the total number of training epochs is 10. The

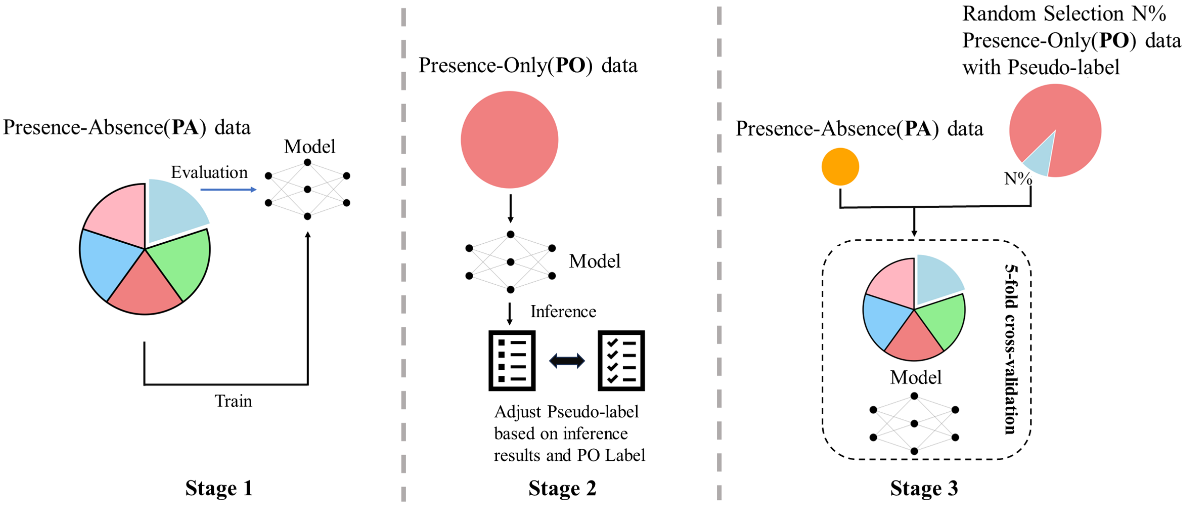
Presence-Absence(PA) data
Model
Evaluation
Train
Stage 1
Presence-Only(PO) data
Model
Inference
Adjust Pseudo-label based on inference results and PO Label
Stage 2
Random Selection N% Presence-Only(PO) data with Pseudo-label
Presence-Absence(PA) data
N%
5-fold cross-validation
Model
Stage 3

Figure 2: The overview of the Pseudo-label framework. In stage 1, we train our model with PA data in 5-fold strategy. In stage 2, with the trained model, we inference on PO data to obtain abundant data with pseudo-label. In the final stage, we train our model with PA and partial PO data (with pseudo label generated in stage 2) to obtain final model.

learning rate is decayed using cosine annealing. The confidence threshold T^- and T^+ are set to 0.05 and 0.4, respectively. The model checkpoint with the highest F1 score on the validation set is selected for testing.

Hardware Environment: Our experiments are conducted on a computer equipped with an NVIDIA RTX 4090 GPU. We use the PyTorch deep learning framework for model training and inference.

Table 1
Kaggle Scores

Private Rank	Private Score	Public Score
1	0.40890	0.41092
2 (Ours)	**0.36837**	**0.37327**
3	0.35292	0.35405
4	0.35220	0.35579
5	0.34898	0.34873

5.1. Ablation Study

To better understand the impact of different components and design choices in our proposed method, we conducted an ablation study. The results are presented in Table 3.

We began with a baseline model submitted by organizer[1] and gradually added various components to observe their effect on the model's performance. The baseline achieved a Kaggle Private Score of 0.31535. By utilizing 5-fold cross-validation, we saw an improvement of 0.02075, reaching a score of 0.33610. Next, we investigated the effect of different threshold values on the model's performance, as shown in Table 2. We found that setting the threshold to 0.2 yielded the best result, with a Kaggle Private Score of 0.34886, an improvement of 0.01276 over the previous step.

Incorporating tabular data into the model provided a slight boost in performance, increasing the score by 0.00618 to 0.35504. Using a self-supervised pretrained ResNet (SSL pretrained ResNet) to the model resulted in an improvement, raising the score by 0.00520 to 0.36024. Finally, the introduction of

[1] https://www.kaggle.com/code/picekl/sentinel-landsat-bioclim-baseline-0-31626

our pseudo-labeling technique led to a significant improvement, raising the Kaggle Private Score to 0.36837, an increase of 0.00813 compared to the previous step.

These results demonstrate that each component of our proposed method contributes to the overall performance, with the pseudo-labeling technique being the most influential. The ablation study highlights the effectiveness of our design choices and validates the importance of utilizing both the PA and PO data through our pseudo-labeling framework.

Table 2
Threshold Ablation Study Results

Threshold	Kaggle Private Score
Top 25	0.33610
Top 30	0.32966
Top 20	0.33915
0.2	**0.34886**
0.22	0.34721

Table 3
Ablation Study Results

Method/Feature Added	Kaggle Private Score	Score Improvement
Baseline	0.31535	-
+ 5 folds cross validation	0.33610	+0.02075
+ Positive threshold=0.2	0.34886	+0.01276
+ Tabular Data	0.35504	+0.00618
+ SSL pretrained ResNet	0.36024	+0.00520
+ Pseudo Label	**0.36837**	+0.00813

6. Conclusion

This paper presents our participation in the 2024 GeolifeCLEF competition. For the multi-label plant species prediction task, we propose our multimodal deep learning model, which consists of multiple ResNet-based multimodal feature extractors, and we further use a pre-training model to ensure the effectiveness of feature extraction for satellite image information. To effectively utilize the huge amount of PO data, we propose a pseudo-label training framework to further improve the accuracy and robustness of the model on the task. Our experiments demonstrate that our proposed multimodal deep learning model improves on both public and private test sets, and we also demonstrate the effectiveness of our proposed Pseudo-label training framework through ablation experiments.

7. Acknowledgments

This work was supported in part by the National Science and Technology Council, Taiwan under grants NSTC 112-2634-F-002-005 and NSTC112-2634-F-006-002, and National Taiwan University under grants 113L900902.

References

[1] L. Picek, C. Botella, M. Servajean, B. Deneu, D. Marcos Gonzalez, R. Palard, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2024: Species presence prediction based on occurrence data and high-resolution remote sensing images, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[2] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, et al., Overview of lifeclef 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[3] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering 26 (2013) 1819–1837.

[4] D. Zhou, P. Chen, Q. Wang, G. Chen, P.-A. Heng, Acknowledging the unknown for multi-label learning with single positive labels, in: European Conference on Computer Vision, Springer, 2022, pp. 423–440.

[5] M.-K. Xie, J. Xiao, S.-J. Huang, Label-aware global consistency for multi-label learning with single positive labels, Advances in Neural Information Processing Systems 35 (2022) 18430–18441.

[6] B. Liu, N. Xu, J. Lv, X. Geng, Revisiting pseudo-label for single-positive multi-label learning, in: International Conference on Machine Learning, PMLR, 2023, pp. 22249–22265.

[7] B. Kellenberger, D. Tuia, Block label swap for species distribution modelling., in: CLEF (Working Notes), 2022, pp. 2103–2114.

[8] C. Leblanc, A. Joly, T. Lorieul, M. Servajean, P. Bonnet, Species distribution modeling based on aerial images and environmental features with convolutional neural networks., in: CLEF (Working Notes), 2022, pp. 2123–2150.

[9] H. Q. Ung, R. Kojima, S. Wada, Leverage samples with single positive labels to train cnn-based models for multi-label plant species prediction, Working Notes of CLEF (2023).

[10] C. Botella, B. Deneu, D. Marcos, M. Servajean, J. Estopinan, T. Larcher, C. Leblanc, P. Bonnet, A. Joly, The geolifeclef 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe, arXiv preprint arXiv:2308.05121 (2023).

[11] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, pmlr, 2015, pp. 448–456.

[12] A. F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375 (2018).

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[14] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[15] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, X. X. Zhu, Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation, arXiv preprint arXiv:2211.07044 (2022).

[16] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).