Context-Selected Negatives: Triplet-Loss Self-Supervised Learning for Wildflower Classification

Elena Sierra
esierra@stanford.edu

Chase Nwamu
cnwamu@stanford.edu

Sean Zhang
szhanggg@stanford.edu

Eugene Hong
eugeneh@stanford.edu

Abstract

We address the problem of self-supervised classification from unlabeled image collections. Unlike existing approaches that learn useful features by maximizing similarity between context-aware positive samples and by randomly picking negative samples, we also make use of context data from image collection to select negative samples. To achieve this, we utilize geolocation data to distinguish between input images and select positive samples that are close to the chosen image and negative samples that are far from the chosen image. By identifying context-selected positive pairs and negative pairs, we increase accuracy in classification compared to methods that do not utilize context-aware negative pair selection. We present results on a variety of wildflower images from across the state of California across a supervised, classical triplet-loss, and a pair-selection triplet loss learning methods. Leveraging the negative image selection in a contrastive method improves accuracy and effectiveness of wildflower classification models, which contributes to more robust ecological monitoring and conservation efforts.

1. Introduction

The process of classifying plant and animal species is essential to monitoring biodiversity as well as understanding ecosystems' health. Despite recent advancements in wildflower classification using deep learning techniques [3,6,9], current methods have not yet achieved optimal accuracy, especially with datasets with limited labeling. Contrastive self-supervised methods have not been deeply explored to date, and findings so far have been mixed in terms of the success of these methods [2, 4, 5]. This lack of success is in part explained by the iNaturalist's dataset being skewed due to geographic imbalance and class imbalance, which makes random sampling highly inefficient for training models, since it is generally assumed that examples are idependent and identically distributed. One way to utilize contrastive methods is to select positive and negative images for a triplet-loss based model, and researchers have found that utilizing image context (geolocation, timestamps, etc.) for selecting positive pairs increases the accuracy of classification [11]. The integration of geolocation data for negative pair selection, which has proven valuable in geography-aware self-supervised learning [1], has been largely overlooked in plant classification studies. As a result, there exists an opportunity to investigate novel approaches that leverage the negative image selection in a contrastive method to improve accuracy and effectiveness of wildflower classification models, which could have important applications for other machine learning tasks that are conducive to having imbalanced data, contribute to more robust ecological monitoring and conservation efforts.

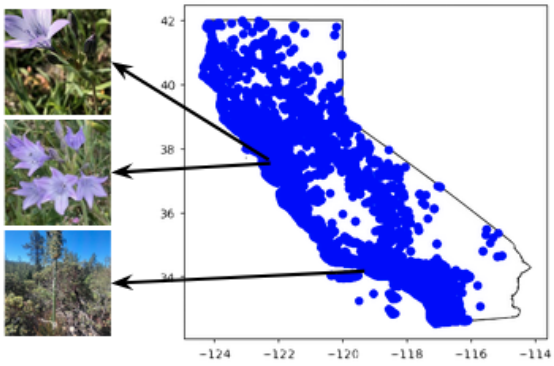


Figure 1. Geolocation provides useful context for species identification. Wildflower images taken nearby are more likely to be the same species while images taken further away are more likely to be different.

In this paper, we introduce a novel approach to wildflower classification that specifically addresses the limitations of previous methods by integrating triplet loss-based contrastive learning with geolocation data. The key innova-

tion in our approach, termed "Context-Aware Negatives," is the utilization of geolocation data for the selection of negative pairs in the contrastive learning process. By incorporating this geospatial information, we are able to effectively harness the discriminative power of triplet loss-based learning, which aims to minimize the distance between positive pairs while maximizing the distance between negative pairs. This unique combination allows our model to capture intricate patterns within wildflower species distributions, ultimately improving classification accuracy and setting our method apart from existing approaches. In doing so, we aim to demonstrate the untapped potential of leveraging both geolocation data and contrastive learning techniques for more accurate and robust wildflower classification.

We have developed a comprehensive framework for our Context-Aware Negatives model that combines the strengths of triplet loss-based contrastive learning with the incorporation of geolocation data. The key to this framework lies in the careful selection of positive and negative pairs based on both visual features and geographical proximity. By doing so, we are able to create a more informative and discriminative representation of wildflower species, which in turn leads to improved classification accuracy. Furthermore, our method is designed to be robust and adaptable to different wildflower data sets, allowing for its potential application in a wide range of classification tasks. By implementing the Context-Aware Negatives model, we demonstrate the effectiveness of integrating geolocation data into contrastive learning and provide a strong foundation for future research on similar natural data sets, potentially transforming the way we approach classification problems.

To rigorously evaluate the effectiveness of our Context-Aware Negatives model, we plan to compare its performance to existing wildflower classification methods, focusing on accuracy as our primary metric. This comparison will not only highlight the improvements brought about by our approach, but also provide insights into its strengths and potential areas for refinement. Additionally, we intend to conduct experiments on various wildflower data sets, examining the generalizability and robustness of Context-Aware Negatives in real-world scenarios. By performing these evaluations, we aim to demonstrate that our model is not only theoretically sound but also practically applicable, paving the way for broader adoption of our methodology in the domain of wildflower classification and beyond.

Wildflower classification is just one example of a broader challenge in the field of computer vision, where a lack of quality labeled training data and difficulties in distinguishing between visually similar classes hinder progress. Our proposed approach, Context-Aware Negatives, not only addresses these issues in the specific domain of wildflower classification but also lays a foundation for future research in contrastive learning and geolocation data integration in other domains of natural data sets. By using triplet loss-based contrastive learning with geolocation data, we show that it is possible to improve classification accuracy in a challenging real-world scenario with limited labeled data. Furthermore, our approach demonstrates the potential of positive and negative pairing in contrastive learning for effective feature extraction in complex image recognition tasks. In conclusion, our proposed Context-Aware Negatives model has the potential to transform the way we approach classification tasks in natural data sets, opening doors for future research and applications in other fields such as biodiversity monitoring, environmental conservation, and medical imaging.

2. Related Works

2.1. Wildflower Classification

Wildflower classification studies have historically relied heavily on fully supervised approaches, constraining their adaptability to diverse, complex datasets and novel species [**?**]. Recently, however, there has been a growing interest in leveraging machine learning, especially deep learning, for species identification. For instance, the iNaturalist platform employs a computer vision model for automatic species identification, capable of classifying thousands of species with remarkable accuracy [7]. This progress is supported by large, rich datasets like the iNaturalist dataset and the WILDS2.0 dataset, which incorporate a broad range of species, including numerous plant species [8, 16]. Despite these advancements, the complex nature of species identification poses considerable challenges. In response, some researchers have explored contrastive learning for species identification, demonstrating promising results in image-based representation learning for natural world collections [15]. Nevertheless, there is still a significant gap in the literature on the specific application of these techniques to wildflower classification, especially in the context of using geographical information for sample selection. This study aims to address this gap and contribute novel insights into the efficacy of contrastive learning in wildflower classification.

2.2. Contrastive Learning and Augmentation

Contrastive self-supervised learning has recently emerged as a compelling approach for object classification tasks. Its central premise is to use augmented versions of the original image to form positive pairs, encouraging the model to map them close together in the feature space. Meanwhile, a negative pair is created by contrasting the original image with a dissimilar image, guiding the model to create a separation in the feature space [4]. These techniques are highly data-dependent, utilizing the inherent structure of the dataset to inform learning, rather than

relying solely on explicit labels [13].

The use of data augmentation techniques, which include spatial transformations (cropping, resizing, rotation, cutout) and appearance transformations (color distortion, blur, filtering), is central to the success of contrastive learning. Through associating positive pairs, the model learns the intrinsic characteristics of objects and the acceptable degree of noise or distortion, thereby enhancing its object recognition capabilities.

Despite its effectiveness, certain limitations exist with these augmentations, especially when applied to plant species identification. Specific augmentations might distort important features crucial to species recognition, which warrants careful consideration in choosing and applying augmentations in this context. In addition, key factors in identification, such as color, are not removed by most augmentations. This can prevent a model from learning to account for these factors, thus leading to poor generalization.

2.3. Feature Utilization for Under-Labeled Data Sets

Contrastive learning is a subset of a self-supervised learning approach that offers a robust strategy for feature learning from under-labeled datasets. Rather than exhaustively annotating data, these methods learn by predicting missing parts of the data, which can include tasks like image inpainting, jigsaw puzzle solving, and rotation prediction [10, 12, 19]. Despite the success of these non-contrastive self-supervised learning methods, they have limitations when used for object detection and classification, especially in scenarios requiring large-scale label-rich datasets.

Emerging research has begun to address this by exploring various image classification models like ViT or MLP mixer. However, their application in specific domains like species identification has often faced challenges, underscoring the need for novel approaches like contrastive learning. Importantly, previous work has not sufficiently examined the role of negative sampling alongside positive pair selection in contrastive learning. This represents a significant opportunity for advancing the state of the art in object detection and classification tasks.

2.4. Positive Pair Selection in Constrastive Learning

In the case of wildflower identification, geographical location is known to be highly predictive of plant species. This has been proven by existing positive-pair selection models, and as such, our approach aims to improve those efforts even more by factoring geographic location into negative pair selection to ensure different negative images. To improve accuracy, researchers have also explored going a layer deeper and using augmented positive pairs to find other positive pairs in the data set to obtain more diverse positive pairs, therefore increasing the accuracy of the model [5]. Researchers with data sets with context, such as biodiversity monitoring images, geographical image data, and ICU patient data have found utilizing time and location context improves positive pairing [1, 11, 18]. Data sets with ICU patients groups patients with others based on contiguous time segments [18]. Data sets with geographic data select positive pairs from nearby regions [1]. Data sets with biodiversity monitoring have utilized pairing images taken at similar times, locations, date, and time of day [11].

2.5. Triplet Loss-Based Contrastive Learning and Its Relevance for Wildflower Classification

Other applications of contrastive learning models include time series representation and FaceNet. Time series representation learning has been improved with contrastive triplet selection based on variance across time intervals [17]. In FaceNet, a system for face recognition and clustering, triplet loss has also enhanced performance [14]. Furthermore, a triplet loss-based approach has contributed to creating accurate and discriminative distance metrics for classification tasks [2].

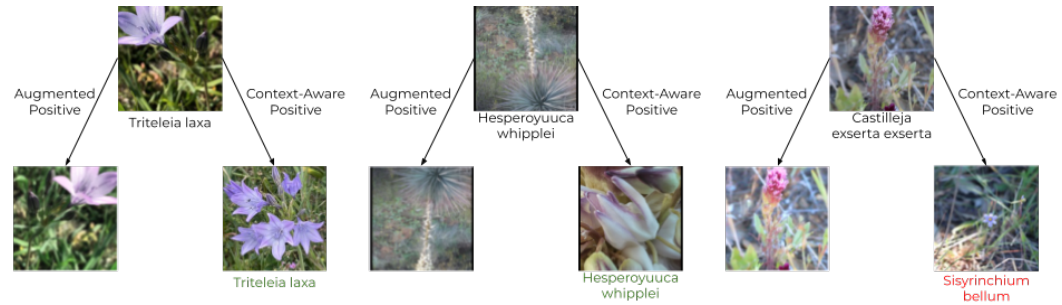
While this method has significantly impacted domains like time series analysis and face recognition, its application in wildflower classification is yet unexplored. This research gap presents an opportunity to innovate on existing methodologies for wildflower identification, especially so given the geographic imbalance present in iNaturalist wildflower datasets. Past work has not yet considered picking context-aware negatives to avoid the risk the imbalance poses of randomly choosing negative pairs that are similar to the anchor. As such, we posit that integrating triplet loss-based contrastive learning, particularly with a focus on selecting informative negative samples along with positive ones, will enhance the accuracy of wildflower classification.

3. Methods

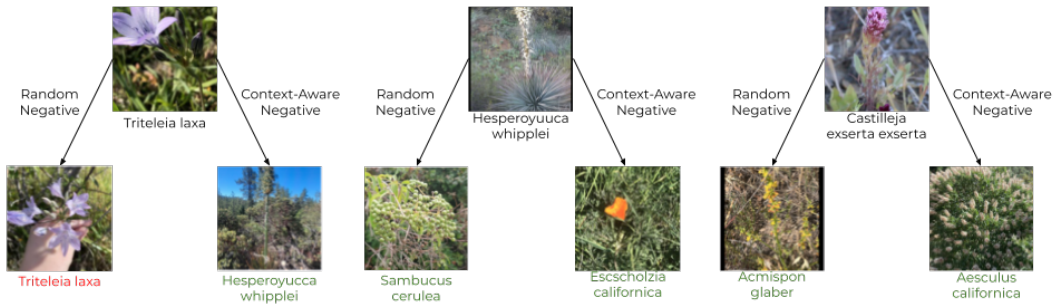
In this project, we run and compare the results of 4 different models: A baseline supervised, a baseline contrastive model, a contrastive model focused on positive pair context, and a contrastive model focused on positive and negative context. The positive focused contrastive model uses geospatial location to determine positive pairs, and randomly selects negative pairs. The positive and negative focused contrastive model uses geospatial location to select both the positive and negative pairs.

3.1. Dataset

To train our models, we used a set of over 100,000 images uploaded to the iNaturalist database. These images were taken from all over the state of California in 2022 (2021 for the testing data set), and each labeled with their species name, the longitude and latitude that the photos



(a) Positive pair selection for augmentation vs. geolocation



(b) Negative pair selection for random vs geolocation

Figure 2. Here we show examples of augmented vs context aware positives (a) and random vs context aware negatives (b). We see that the context aware positives can provide a better perspective than an augmented version of the original image, but there is a risk that the context-aware positive will not be the same species and skew results. We also see that there is a smaller risk of selecting the same species for the negative pair with the context-aware negatives, especially with more common species in the data-set (like the Triteleia laxa).

were captured, the date they were taken, and the quality of the photo. For the purposes of this experiment, we used only the species names as well as the longitude and latitude at which the wildflowers were found. We use only photos of quality research grade or higher. Furthermore, we used only photos with a location uncertainty of less than 120 meters. The iNaturalist dataset contains wildflower images for 64 species.

3.2. Implementation Details

In order to get predictions from the models, we ran batches of 64 150x150 3 channel images through a 5 layer Resnet 18 Network. Each of the 64 images and their corresponding labels were selected at random from the iNaturalist dataset, converted to 3x150x150 tensors, then run through the 5 layers of the Resnet 18 model. The 5th layer of the Resnet outputted a tensor with 64 parameters, so after being run through the Resnet, the batch tensor was of size 64x64 (1 parameter to represent the value for each species). For training of both the supervised and contrastive model, we ran 100 epochs, where one epoch was trained on a batch of 64 images about 1200 times.

3.3. Baseline Triplet-Loss

Regarding the beseline contrastive learning model, we use three images from our dataset, X, $f(X)$, and $\tilde{X}$ where $f(X)$ is the first image with some transformation applied. Specifically, we perform a random horizontal flip, a random vertical flip, and a random rotation between -180 and 180 degrees. $\tilde{X}$ is a different image sampled randomly from the dataset. The positive and negative image selection occurs in a pre-train dataset class. We then freeze the weights of the trained neural network and add an additional last layer with labels, which is 10% of the training data, for fine-tuning.

3.4. Geolocation Pair Selection Triplet-Loss

Regarding the triplet-loss model with the positive and negative pair selection based on geolocation, we use three images from our dataset, X, X_p, and X_n. We calculate the euclidean distance between images by using their latitude and longitude values in degrees. X_p is an image at most 0.08 degrees away from X and must be one of the 5 closest images to X, and X_n is an image at least 5 degrees away from X. The positive and negative image selection occurs in a pre-train dataset class. We then freeze the weights of the trained neural network and add an additional last layer with labels, which is 10% of the training data, for fine-tuning.

4. Evaluation

A contrastive triplet based loss method will be effective for the task of wildflower classification. Furthermore, utiliz-

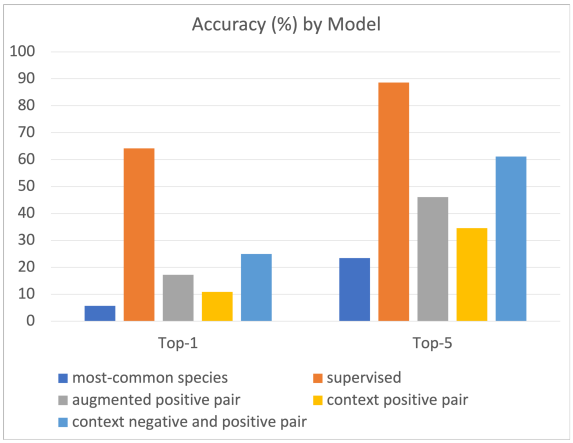


Figure 3. Accuracy Across Models. We observe that a fully supervised model has the highest accuracy for Top-1 and Top-5 classification. However, the context-aware negative and positive pair model outperforms all other self-supervised methods.

ing contextual information such as geospatial location to select both positive and negative pairs in the triplet loss model will be even more effective than both the baseline triplet loss model, and a triplet loss model that only utilizes context information in creating positive pairs. Selecting an informative negative pair using context is just as important as selecting an informative positive pair.

4.1. Results

Baseline Models Performance. We first run our data set some baseline modeling methods before introducing context-aware pair selection. To do this, we implemented a "most-common species" model and a fully supervised model. The "most-common species" model always guesses the most common or five most common species, and has a 5.7% top-1 accuracy and a 23.4% top-5 accuracy. Regarding the fully supervised model, we constructed a custom dataset class that cropped all images to 150x150 pixels and ensured that all images had rgb format. We utilize an index map to represent the 64 species in the dataset. After running the custom dataset and dataloader, we utilize the RESNET 18 infrastructure to train the model. The fully supervised model has a 63% top-1 accuracy and a 88.6% top-5 accuracy.

Impact of Augmented Contrastive Model. Since our baseline models both require access to labels, which may not be applicable in many natural datasets, we run our data set on a more traditional self-supervised contrastive model using augmented positive pairs. The augmented positive pair model has a 17.2% top-1 accuracy and a 46.1% top-5 accuracy. This model performs better than the most-common-species model and worse than the fully supervised model.

Impact of Context-Aware Positives. The context-aware positives model has a 10.9% top-1 accuracy and a 34.6% top-5 accuracy. This performs worse than the augmented contrastive model and better than the most-common species model. This model introduces more variety in the positive pairs, but it may introduce pairs that are not the same species as well.

Impact of Context-Aware Negatives. The context-aware negatives and positives model has a 25.0% top-1 accuracy and a 61.1% top-5 accuracy. This model outperforms all other self-supervised model and only performs worse than the fully supervised model. Clearly, there is a benefit from being able to select negative image pairs further away from the anchor image.

Analysis: When comparing the results from the different models, we see that adding context-aware selection can drastically change the performance of our contrastive models. The context-aware positives model performs worse than the augmented contrastive model. This is likely because the context-aware positive pairs can be of different species since they are selected solely by geolocation. The context-aware negatives and positives model works the best likely because the negative pairs must be from a different geographic region. The only model that outperforms the context-aware negatives approach is the fully supervised model, which has labels 100% the training data, likely causing the increased accuracy.

4.2. Discussion

Using geolocation context for selecting positive and negative pairs improved accuracy on wildflower classification for 64 species of wildflowers in California from an iNaturalist dataset, assuming that cropping and rgb correcting does not significantly impact image quality. One limitation is that our training data contains only 64 species of wildflower, so to extrapolate to wildflower classification or species classification as a whole is nontrivial.

Also, the data that our model is trained on is not necessarily perfect, as all of the pictures are human identified, so there may be errors. We also edited some of the images in our data (the black and white ones) to make them run correctly, so this could potentially cause issues with uniformity.

Lastly, our parameters might be overfitted to our specific data. Similar to the first point, its hard to prove that results can be generalized when we are testing on a limited amount of species.

5. Conclusion

All in all, through the comparison of four different models, we found that the model incorporating both context-aware positive and negative pair selection outperformed other self-supervised models in terms of top-1 and top-5 accuracies. Importantly, this model was second only to the fully supervised model, which requires comprehensive label data that might not be always available, thus underlining the efficacy of our proposed approach. The contrastive model that used only context-aware positive pair selection did not perform as expected; while it did manage to outperform the most common species model, it was overshadowed by the augmented contrastive model. This result highlights the importance of incorporating context in both the selection of positive and negative pairs.

However, we acknowledge some limitations in our study. While we have successfully applied our methodology on a specific dataset containing 64 wildflower species from California, it is important to consider that extrapolating these findings to a broader context might pose challenges. The human-identified and modified images in our training data could potentially introduce errors. Additionally, the model might be overfitting to our specific data, calling for caution when generalizing these results. Further research is required to validate these findings in diverse datasets.

Despite these challenges, our work represents a step towards unlocking the potential of contrastive learning models for species identification tasks. We have demonstrated that it is possible to improve model accuracy by integrating geospatial context into the learning process. This technique can be employed in a range of environmental and biodiversity monitoring tasks where labeled data is scarce or not uniformly distributed.

By bridging the gap between self-supervised learning and geospatial context integration, we hope to inspire further exploration and innovation in this field. As we continue to refine and develop these techniques, we draw closer to our broader vision of creating intelligent systems that can drive meaningful insights from under-labeled data in diverse and complex domains. Our work serves as a stepping stone towards realizing this ambitious goal.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 1, 3

[2] Yuan-Chi Chang, Dharmashankar Subramanian, Raju Pavuluri, and Timothy Dinger. Time series representation learning with contrastive triplet selection. In *5th Joint International Conference on Data Science Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD 2022, page 46–53, New York, NY, USA, 2022. Association for Computing Machinery. 1, 3

[3] Jianghao Chen, Yiming Huo, and Junyu Li. Recognition and classification of flower species based on artificial intelligence. *Academic Journal of Computing & Information Science*, 4(8):78–82. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[5] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 1, 3

[6] I Gogul and V Sathiesh Kumar. Flower species recognition system using convolution neural networks and transfer learning. In *2017 fourth international conference on signal processing, communication and networking (ICSCN)*, pages 1–6. IEEE, 2017. 1

[7] The iNaturalist Team. New computer vision model. *iNaturalist*, 2022. 2

[8] Pang Wei Koh, Percy Liang, Arushi Gupta, Benjamin Recht, and Daniel Selsam. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2112.05090*, 2021. 2

[9] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1

[10] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles (2016). *arXiv preprint arXiv:1603.09246*, 2. 3

[11] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10583–10592, 2021. 1, 3

[12] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[13] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020. 3

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 3

[15] Grant Van Horn, Oisin Mac Aodha, Yang Song, Barak Barz, Alex Shepard, Pietro Perona, Serge Belongie, and Yin Cui. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[16] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2

[17] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. 3

[18] Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pages 11964–11974. PMLR, 2021. 3

[19] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 3