ImageNet Misrepresents Biodiversity, But Does It Matter For Pretraining?

Gaurav Iyer [*†] David Rolnick [*] Esther Rolf [‡] Sasha Luccioni [§]

Abstract

Recent research has shown that the ImageNet-1K dataset is biased and inaccurate in a variety of ways in how it represents wildlife. Even so, pretraining on ImageNet consistently improves performance for downstream applications such as species classification. To investigate why this occurs, we consider the effect that pretraining on different kinds of data has on identification of different types of animals. We first evaluate whether pretraining on ImageNet helps more for species that are well-represented in the dataset than for those that poorly represented or misrepresented. Surprisingly, we find no appreciable difference between performance on these two groups, suggesting that the effectiveness of ImageNet pretraining is largely a result of domain-agnostic features and is relatively unaffected by the identity of individual species or even entire taxonomic groups. To further explore this conclusion, we consider the effect of pretraining on different subsets of ImageNet, finding that while the animal-related images have the most benefit for downstream animal recognition tasks, the effect is not sensitive to taxonomic groups, and non-animal images also provide significant benefit. Our work helps explain why ImageNet, despite its problems, remains useful for pretraining while also signaling the value of large-scale pretraining datasets that are specialized to the target downstream task.

1. Introduction

Across a wide variety of computer vision tasks and domains, it is extremely common for practitioners to pretrain models on the ImageNet-1K dataset [8], or to initialize their models with easily obtainable weights that have already been trained on ImageNet. For instance, Van Horn et al. [33] introduce iNaturalist2021 and iNaturalistMini and find that using pretrained ImageNet weights significantly improves model performance, while Orsic et al. [22] use ImageNet pretraining to create a lightweight neural network architecture for real-time semantic segmentation of road images for autonomous vehicles. Using ImageNet as a starting point is often tacitly assumed, as doing so frequently improves performance on a downstream task by at least a few percentage points. Indeed, even for tasks that are seemingly unrelated to the images present in ImageNet, pretraining on ImageNet can improve results. For example, in the context of remote sensing, ImageNet pretraining has been shown to improve classification results by convolutional networks [20] and leveraged for Vision Transformers [2]. In healthcare contexts, Xie and Richmond [35] show that pretraining on a grayscale version of ImageNet can improve performance for disease classification using X-ray images.

Notably, ImageNet pretraining has been widely used in the domain in biodiversity [6, 28, 33], where computer vision algorithms for recognition of animals have found widespread adoption in monitoring threatened species [7], invasive species [12], understudied taxonomic groups [27], and impacts of human activities on wildlife [11]. Pretraining data can be especially important in this domain due to the limited and long-tailed nature of biodiversity datasets [32], which can arise from the wide variety of possible species, the inherent rarity and difficulty in sampling certain species, and the expert knowledge required for annotation.

However, it has been observed that ImageNet in fact contains a plethora of mistakes and biases in how it represents wildlife [19, 31], arising from its annotations being provided by members of the public with limited domain expertise. In this work, we investigate why, despite such significant problems, ImageNet pretraining helps improve performance on downstream animal classification tasks. To isolate the effects of animal representation during pretraining, we pretrain ResNet50 networks [13] on a variety of ImageNet subsets and evaluate the impact on subsets of the iNaturalist dataset [32], which unlike ImageNet was curated by expert naturalists. We consider the differential impacts on species well-represented by ImageNet, species poorly represented by ImageNet, and groups of species as a whole.

Our main contributions may be summarized as follows:

- We find that the value to pretraining on ImageNet for classification of iNaturalist images does not appear linked to how well- or misrepresented a species or taxonomic group is within the ImageNet dataset. This suggests that the effect of ImageNet pretraining is largely independent

[*]McGill University & Mila - Quebec AI Institute
[†]Correspondence to: *gaurav.iyer@mila.quebec*
[‡]University of Colorado Boulder
[§]Hugging Face

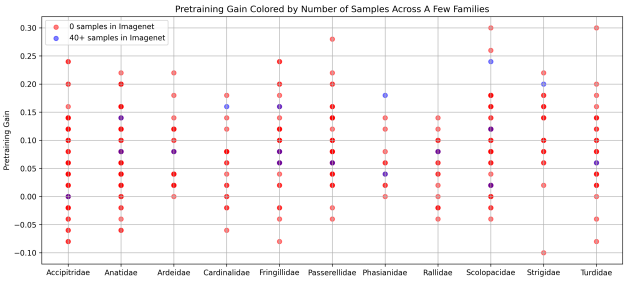


Figure 1. Figure illustrating "pretraining gain" – the difference in model performance between a model pretrained on ImageNet and finetuned on iNat-Mini-A, and a model that is trained on iNat-Mini-A from random initialization. Red points and blue points indicate bird species that have 0 samples and 40+ samples in the corrected version of the ImageNet validation set respectively (serving as a proxy for abundant data in the larger training set, for which corrected labels were not available).

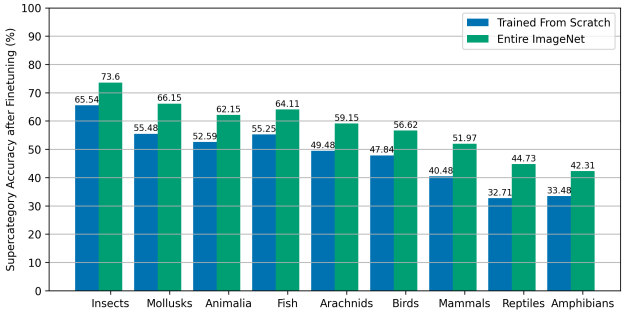


Figure 2. Figure comparing the supercategory level performance between an ImageNet pretrained model finetuned on iNat-Mini-A, and a model trained from scratch on iNat-Mini-A. Note that the superclass Animalia consists of all animal species that do not cleanly fall into one of the other superclasses (e.g. corals).

of the exact species used in pretraining.

- We find that the biodiversity-related images within ImageNet confer especially high value during pretraining as measured by downstream performance on wildlife recognition, but that other images are also valuable, confirming that the value of pretraining is not closely tied to domain.
- Our results suggest that the benefits of pretraining are not "maxed out" by ImageNet-1k – additional data on wildlife-related images could likely further benefit downstream performance on wildlife recognition tasks.

Our experiments and analysis highlight why pretraining on ImageNet, despite its issues, is consistently beneficial for finetuned model performance. In most practical applications where models are trained on ImageNet, it is likely that the finetuning data is dissimilar enough to ImageNet in how it represents relevant data (i.e., animals) that the benefit of pretraining largely comes from the learning of high-level image features that are not specific to the downstream task.

2. Performance on Individual Taxa

We know that ImageNet is incomplete and often erroneous in how it represents biodiversity [19, 31] and yet, pretraining on ImageNet significantly benefits performance on downstream tasks like animal species classification [33]. We consider the possibility that ImageNet pretraining leads to smaller improvements in performance compared to training from scratch when the animal species present in the finetuning dataset is poorly represented in ImageNet.

To answer this question, we pretrain a ResNet50 model on ImageNet and finetune it on iNat-Mini-A, and compare its performance to that of a ResNet50 that was trained from scratch on iNat-Mini-A. In particular, we consider the difference in classwise performance between the two settings on various bird species present in iNat-Mini-A. We restrict ourselves to bird species, since a thorough analysis of the data corresponding to birds in the ImageNet validation set performed by Luccioni and Rolnick [19] allows us to better understand the impact of existing errors in the pretraining data. With the help of experts in bird classification, Luccioni and Rolnick [19] re-annotated all 2850 images from the 57 classes of birds in the ImageNet validation set, highlighting several of its representational issues. To address the aforementioned question, we obtain these annotations from the authors of Luccioni and Rolnick [19], using it as a proxy for the *true* representation of different bird species in the (larger) ImageNet training set. Given this true representation, we check whether bird species that are well-represented in ImageNet show better performance improvements than average after finetuning.

We filter out bird species that are well represented in ImageNet (40+ samples in the ImageNet validation set after correction) and are present in iNat-Mini-A through one or more class labels.[1] We also filter out species that are not *at all* represented in ImageNet, i.e. they (ideally) have 0 samples in ImageNet.

The results of this experiment are illustrated in Figure 1. Each point on the plot is a bird species that is very well-represented (blue) or is not at all represented (red) in the corrected ImageNet validation set. We divide species based on the family it belongs to, and choose families of birds where at least one species from the family is present in ImageNet. This allows us to observe if the presence of a particular species helps with the performance of only that species or other birds in the family as well, as well as to control for family-specific differences in difficulty of identification. We find that pretraining improves performance on a majority of bird species, with no strong correlation between an

[1] As noted by Luccioni and Rolnick [19], an ImageNet label can often correspond to multiple iNat-Mini-A labels (e.g. "kite" refers to multiple species of bird and indeed the class includes multiple species).

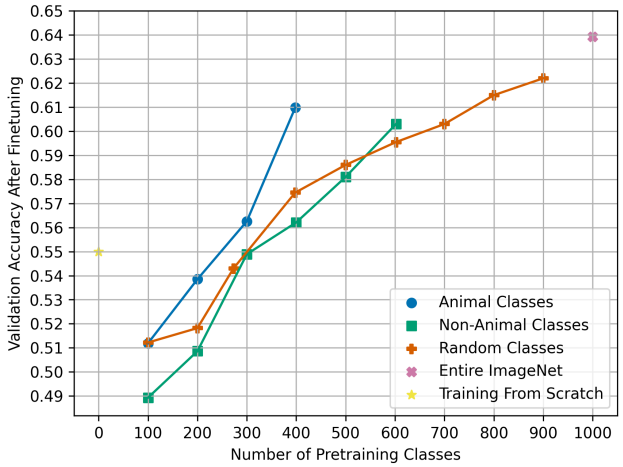


Figure 3. Validation accuracy after finetuning on iNat-Mini-A vs. the number of ImageNet classes used for pretraining, for different ImageNet subsets.

increase or decrease in finetuned model performance and how well represented a species is in ImageNet.

We next consider broader taxonomic groups, and ask the same question with respect to animal supercategories present in iNat-Mini-A. For instance, 127 classes in ImageNet are dedicated to domestic animals, most of which are dogs and cats. Does ImageNet pretraining yield greater performance improvement for the supercategory of mammals in iNat-Mini-A, compared to other types of animals? We find that this is in fact not the case. In Figure 2, we plot the validation performance of an ImageNet pretrained model after being finetuned on iNat-Mini-A, along with the performance of a model trained from scratch on iNat-Mini-A. We find no bias toward specific supercategories in how much performance improves due to pretraining – instead, performance improves uniformly across supercategories.

3. The Effect of Specialized Pretraining Data

To understand the effect of the pretraining data composition on downstream performance, we consider subsets of ImageNet that have vastly different compositions. We consider ImageNet-A, ImageNet-R, and ImageNet-NA, which correspond to animal-only, random, and non-animal classes in ImageNet – more details are provided in Appendix D.1.

3.1. Finetuning on iNat-Mini-A

We pretrain ResNet50 networks on different subsets of ImageNet and finetune the resulting networks on INatMini-A. The finetuned networks are then evaluated on INatMini-A-val, the subset of the validation set which corresponds to animals, the results of which are illustrated in Figure 3. We find that pretraining on the entirety of ImageNet is, by far, the most effective option for maximizing accuracy, while pretraining on a sufficiently large set of "irrelevant" or random data is still better than no pretraining at all. However, pretraining on ImageNet-A is the most sample-efficient option, and performs comparably to pretraining on ImageNet-NA. Furthermore, pretraining on ImageNet-A-k consistently outperforms pretraining on ImageNet-R-k or ImageNet-NA-k across different values of k. We also consider how much pretraining on different subsets of ImageNet improves downstream performance, when they are similarly sized. Pretraining on ImageNet-R-398 performs significantly worse than pretraining on ImageNet-A, suggesting that ImageNet-A might be relying on features that are more specialized toward classifying animal species. On the other hand, pretraining on ImageNet-R-k performs very similarly to pretraining on ImageNet-NA-k. Note that this is despite ImageNet-R-k, on average, containing $\approx 0.4 \times k$ classes that correspond to animals. This implies that the models resulting from pretraining on ImageNet-R-k and ImageNet-NA-k are utilizing similar, low-level features despite the presence of a significant number of classes that correspond to animals in ImageNet-R-k.

Put concisely, when the pretraining data is largely heterogeneous, finetuned models depend on general, low-level features even if a subset of the pretraining data is highly relevant to finetuning. On the other hand, when the pretraining data is homogeneous and relevant to finetuning, finetuned models leverage features specialized to the downstream task. Even when the pretraining data is completely "irrelevant", general features can be learned from it and is better than no pretraining and training from scratch. In fact, Figure 3 suggests that the performance gain due to pretraining on the entirety of ImageNet is due to these general features, and that the animal data present in ImageNet is almost irrelevant to its usefulness. The caveat, however, is that a sufficiently large amount of pretraining data must be used to observe a meaningful improvement in downstream performance through pretraining.

3.2. Finetuning on ImageNet-A

In Figure 5, we illustrate the differences in how ImageNet and iNaturalist represent animals. How much do such representational differences affect the benefits of pretraining? To understand this, we seek to minimize the representational difference between the pretraining and finetuning dataset. We finetune networks pretrained on different subsets of ImageNet on ImageNet-A, and repeat our analysis from Section 3.1. The results are illustrated in Figure 4.

Some observations from 3.1 still hold – pretraining on the entirety of ImageNet remains the most effective choice for maximizing accuracy, while pretraining on ImageNet-A is still the most sample-efficient option. However, there are also some significant differences. Pretraining on ImageNet-

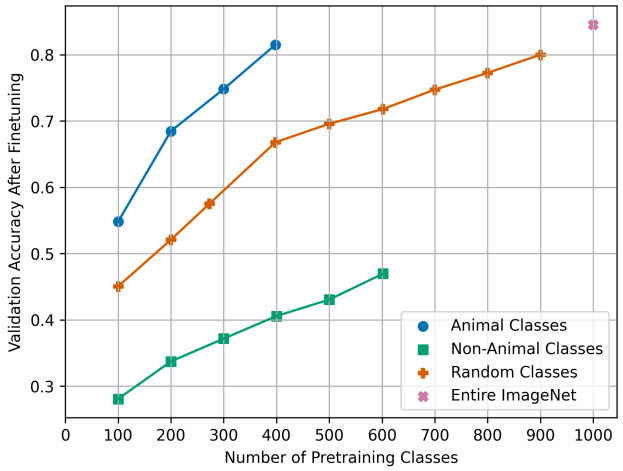


Figure 4. Accuracy on the animal-only subset of ImagenNet's validation set after finetuning on ImageNet-A vs. the number of ImageNet classes used for pretraining, for different ImageNet subsets.

NA results performs significantly worse than pretraining on an equivalent number of ImageNet-R classes – unlike with iNat-Mini-A, the presence of relevant data during pretraining results yields far greater benefits.

As further evidence of this difference in learned features, in Figure 6, we recreate the same plots as in Figure 3 and Figure 4, except the x-axis for the plot corresponding to ImageNet-R is scaled by 0.398 to reflect the expected number of animal-only classes it includes. For iNatMini-A, we see that the line corresponding to $0.398\times$ Random Classes is significantly higher on the y-axis than the line corresponding to ImageNet-A, indicating that the animal-only classes alone cannot account for the improvement in performance due to pretraining. Together with the result in Fig. 3, this suggests that the presence of animal-only classes in ImageNet-R is perhaps irrelevant to the features being learned. For downstream performance on ImageNet-A on the other hand, the line corresponding to $0.398\times$ Random Classes almost exactly matches that of training only on ImageNet-A, which suggests that the animal-only data present in ImageNet-R accounts for all pretraining benefit.

4. Conclusion

This paper considers how pretraining on ImageNet affects the downstream performance of models for wildlife identification, given the biases and inaccuracies present in ImageNet wildlife data. We first examine the performance of models pretrained on ImageNet and finetuned on iNat-Mini-A, considering individual species classes and higher-level taxa, finding that pretraining uniformly improves performance across different species and supercategories. Interestingly, we find that this holds similarly for species that are well and poorly represented in ImageNet, raising questions about the mechanism by which ImageNet pretraining helps improve performance on downstream tasks. To better understand this, we pretrained ResNet50 networks on subsets of ImageNet with different sizes and levels of relevance to the downstream task of animal species classification. These pretrained networks were then finetuned on the subset of animals in iNaturalist-Mini. We found that pretraining on the entirety of ImageNet was the optimal choice to maximize performance, while pretraining on ImageNet-A was the most sample-efficient. Our experimental results also suggest that models pretrained on ImageNet-A learn features that are specialized for animal species classification, while other settings (including pretraining on all of ImageNet) lead to more general, high-level features being learned. In other words, even when animal data makes up for a significant portion of the pretraining data (e.g. $\approx 40\%$ in the case of ImageNet-R), its inclusion in the dataset may not be contributing to the performance any more than an equal amount of non-animal data.

Overall, our work highlights the importance of considering how well-aligned the pretraining and finetuning datasets are in the manner in which they represent data. The downstream performance gain from pretraining on ImageNet is mostly the product of leveraging general features that can be learned from non-animal data as well. However, while pretraining on ImageNet usually results in better downstream performance compared to training from scratch, our experiments also show that we can likely do even better if we have access to a sufficient amount of animal-only data. Moreover, this "sufficient amount" is probably much smaller than the amount of random or non-relevant data required to achieve similar performance. Therefore, it is worth considering whether ImageNet can be replaced by a similarly sized or even smaller dataset which is better aligned with the finetuning task, not just in terms of the task itself but also in how similar the datasets are in the image domain.

To close, we point out some limitations and directions for future work. First, our experiments are restricted to ImageNet and iNaturalist-Mini, and we only examine ResNet-50 networks in an image classification setting. Extending the experimental setting further in these directions could verify the generalizability of these findings. It is also possible that some architectures require a smaller amount of task relevant data required to achieve a certain accuracy, or that for certain tasks, more specialized features significantly outperform the high-level features from ImageNet pretraining. Identifying such scenarios could provide insights that make pretraining more efficient and accessible. Understanding how different pretraining strategies learn different kinds of high- and low-level features and how this translates into performance in downstream tasks would also be a valuable direction for future work.

References

[1] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1790–1802, 2015. 8

[2] Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3): 516, 2021. 1

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 7

[4] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 6

[5] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021. 6

[6] Kim Bjerge, Quentin Geissmann, Jamie Alison, Hjalte MR Mann, Toke T Høye, Mads Dyrmann, and Henrik Karstoft. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77:102278, 2023. 1, 7

[7] Carl Chalmers, Paul Fergus, Serge Wich, and Aday Curbelo Montanez. Conservation ai: Live stream analysis for the detection of endangered species using convolutional neural networks and drone technology. *arXiv preprint arXiv:1910.07360*, 2019. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 8

[10] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? In *Advances in Neural Information Processing Systems*, pages 25050–25080. Curran Associates, Inc., 2023. 8

[11] Mitchell Fennell, Christopher Beirne, and A Cole Burton. Use of object detection in camera trap image identification: Assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology. *Global Ecology and Conservation*, 35: e02104, 2022. 1

[12] Sapphire Hampshire. Invasive species monitoring and predator control via innovative ai technology. *Technical Report*, 2021. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[15] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning?, 2016. 8

[16] Aditya Jain, Fagner Cunha, Michael James Bunsen, Juan Sebastián Cañas, Léonard Pasi, Nathan Pinoy, Flemming Helsing, JoAnne Russo, Marc Botham, Michael Sabourin, et al. Insect identification in the wild: The ami dataset. In *European Conference on Computer Vision*, pages 55–73. Springer, 2024. 7

[17] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[18] Dimitri Korsch, Paul Bodesheim, and Joachim Denzler. Deep learning pipeline for automated visual moth monitoring: insect localization and species classification. *arXiv preprint arXiv:2307.15427*, 2023. 7

[19] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14382–14390, 2023. 1, 2, 6

[20] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2015. 1

[21] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. 6

[22] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[23] Omiros Pantazis, Peggy Bevan, Holly Pringle, Guilherme Braga Ferreira, Daniel J Ingram, Emily Madsen, Liam Thomas, Dol Raj Thanet, Thakur Silwal, Santosh Rayamajhi, et al. Deep learning-based ecological analysis of camera trap images is impacted by training data quality and size. *arXiv preprint arXiv:2408.14348*, 2, 2024. 7

[24] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 8

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 8

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[27] DB Roy, J Alison, TA August, M Bélisle, K Bjerge, JJ Bowden, MJ Bunsen, F Cunha, Q Geissmann, K Goldmann, et al.

Towards a standardized framework for ai-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B*, 379(1904):20230108, 2024. 1

[28] Shoaib Ahmed Siddiqui, Ahmad Salman, Muhammad Imran Malik, Faisal Shafait, Ajmal Mian, Mark R Shortis, and Euan S Harvey. Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science*, 75(1):374–389, 2018. 1, 7

[29] Linda Studer, Michele Alberti, Vinaychandran Pondenkandath, Pinar Goktepe, Thomas Kolonko, Andreas Fischer, Marcus Liwicki, and Rolf Ingold. A comprehensive study of imagenet pre-training for historical document image analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 720–725, 2019. 7

[30] Thanh-Dat Truong, Hoang-Quan Nguyen, Xuan-Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. Insect-foundation: A foundation model and large multimodal dataset for vision-language insect understanding. *International Journal of Computer Vision*, pages 1–26, 2025. 7

[31] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 1, 2, 6

[32] 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*, pages 8769–8778, 2018. 1, 7

[33] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12884–12893, 2021. 1, 2, 7

[34] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *Advances in Neural Information Processing Systems*, 35:6720–6734, 2022. 6

[35] Yiting Xie and David Richmond. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1, 7

[36] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558, 2020. 6

[37] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: From single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021. 6

[38] Bingchen Zhao and Oisin Mac Aodha. Incremental generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19137–19147, 2023. 7

A. Differences in How ImageNet and iNaturalist Represents Birds

(a) Sample of birds in the ImageNet validation set

(b) Sample of birds in the iNaturalist-Mini training set

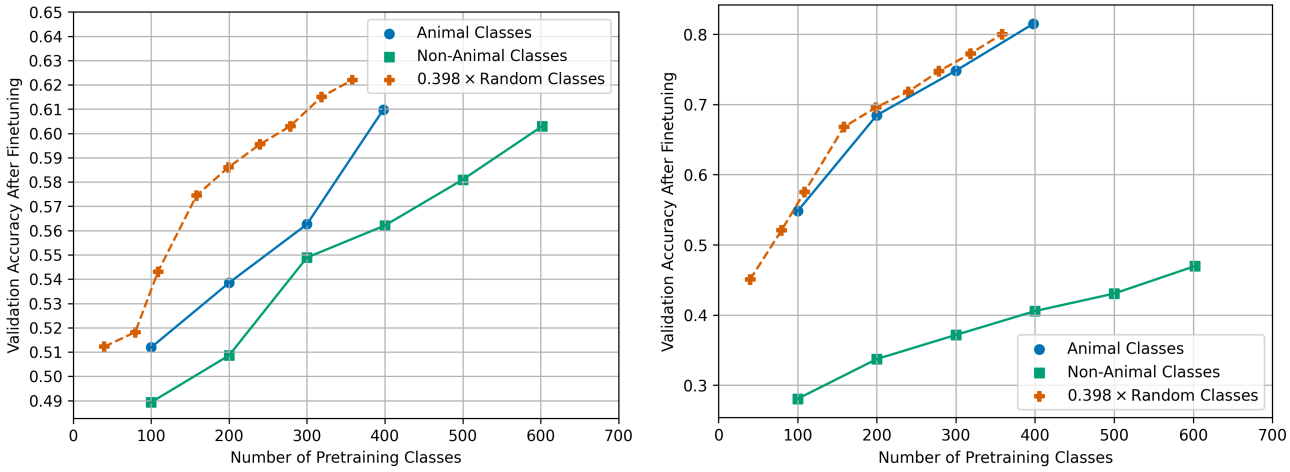
Figure 5. A comparison of the bird images between ImageNet and iNaturalist-Mini. Animal images in ImageNet are generally quite clear and consist of close-up shots of the birds. In contrast, images in iNaturalist-Mini are more likely to be blurry or far away, to have multiple birds or even species of birds in the same image – some images don't even contain birds. We find that these representational differences also extend to the other supercategories of animals in iNaturalist-Mini.

B. Pretraining Benefit When Only Animal Classes Are Counted

C. Related Work

Our work touches on a range of research topics, such as pretraining data selection, biodiversity representation in datasets, and the impact of biases and label errors on performance.

Problems with ImageNet-1K. Past work has analyzed and identified a variety of problems in ImageNet-1K – for instance, the quality of existing ImageNet labels [4] and label errors [21, 34], the presence of multiple classes in a single image [37], and several kinds of biases [5, 36]. In particular, Van Horn et al. [31] work with bird experts to propose class error rates for the bird classes in ImageNet, finding the error rates to be *at least* 4%. Luccioni and Rolnick [19] also work with wildlife experts to perform an in-depth study and label verification of the wildlife data in the ImageNet validation set and find that wildlife data is misrepresented in a



(a) iNat-Mini-A

(b) ImageNet-A

Figure 6. Validation accuracies after finetuning on (a) iNat-Mini-A and (b) ImageNet-A. The figures here are the same as Figure 3 and Figure 4, except for the orange dashed line, which shows the expected number of classes in an ImageNet-R subset, i.e. the x-axis for the orange lines in Figure 3 and Figure 4 are scaled by 0.398. In (a), observe that the orange line is significantly higher on the y-axis than the blue line, indicating that the pretraining benefits cannot be explained completely by the presence of the animal classes in ImageNet-R. The trends also suggest that this gap closes as more animal-only data is considered for pretraining. On the other hand, in (b), the blue and orange dashed lines almost completely match, suggesting that performance improvements from pretraining can almost completely be explained by the animal-only data present in ImageNet-R.

multitude of ways. The experiments presented in this work analyze if such problems in ImageNet impact its effectiveness as a pretraining dataset, specifically in the context of animal species classification.

ImageNet Pretraining for Species Classification. For the better part of a decade, models pretrained on ImageNet-1K have been used as a starting point for many finetuning applications, across a variety of domains [26, 29, 35]. This holds true in the context of ecology as well. Beery et al. [3] introduce the CCT dataset that contains animal images from collections of camera traps placed at the same location and utilize ImageNet weights as a starting point to finetune models on this dataset. Van Horn et al. [33] introduce the iNaturalist2021 dataset, which contains nearly 2.7M images across 10k species – even for such a large and diverse dataset, using ImageNet for pretraining consistently improves performance over training from scratch. Jain et al. [16] introduce benchmarks for insect recognition, in addition to two datasets for moths. They train and evaluate ImageNet-pretrained models that act as their baselines for further analyses. Siddiqui et al. [28] tackle the problem of limited labeled data availability in fish species classification by using models pretrained on ImageNet. Bjerge et al. [6] use pretrained ImageNet weights to conduct hierarchical classifications according to taxonomic rank. Each taxonomic rank is formulated as a different task to learning, framing the task of hierarchical classification as a multi-task problem. The authors also find that this approach is able to correctly distinguish between insect species that are visually very similar.

Use of ImageNet for Other Tasks. Aside from species classification, ImageNet has also been effectively used as a pretraining dataset for various other tasks in the biodiversity domain. For instance, Pantazis et al. [23] use models initialized with ImageNet weights to compare ecological metrics – namely, species richness, occupancy, and activity patterns – that are generated by experts versus those generated by deep learning models. They find that the ecological metrics obtained from deep learning models closely match those provided by experts, outside of cases where a given species is less common due to which there is a drop in model accuracy. Zhao and Mac Aodha [38] explore the problem of incremental generalized category discovery, starting from the use of ResNet-18 models initialized from ImageNet weights obtained via MoCo [14] as a feature extractor.

Alternatives to ImageNet-1K. In most cases, ImageNet pretraining significantly improves performance, but it has been shown that using appropriate alternatives for pretraining can improve performance even more. Korsch et al. [18] develop a system for automated monitoring of moths based on camera traps. The classification of moths is then performed with the help of finetuning a pretrained convolutional network, where authors find that pretraining on iNaturalist2017 [32] is marginally better than pretraining on ImageNet. Similarly, Truong et al. [30] propose Insect-1M a multi-modal insect dataset for training foundation models

for insects – they find that pretraining on Insect-1M outperforms pretraining on ImageNet significantly on multiple classification tasks. Indeed, in some cases, pretraining on ImageNet fails to yield any significant improvement in performance in comparison to training from scratch with a random initialization [10, 24].

Analyses of ImageNet. The generality and widespread use of ImageNet features has also led to several analyses of ImageNet and models that are pretrained on it – for instance, Recht et al. [25] demonstrate the brittleness of models trained on ImageNet by building a new ImageNet test set and Kornblith et al. [17] find a strong correlation between ImageNet accuracy and downstream task accuracy. Huh et al. [15], in a manner similar to us, address various questions about ImageNet pretraining and its efficacy by pretraining on various subsets of ImageNet. However, their work focuses on assessing finetuning on general datasets like PASCAL-VOC [9], while we focus specifically on the transferability of ImageNet to animal species classification. In a similar vein, Azizpour et al. [1] find that changing the number of images per class versus changing the number of classes when pretraining on ImageNet leads to different transfer performance on downstream tasks.

D. Methodology

D.1. Datasets

ImageNet-1K. For pretraining, we use ImageNet-1K (henceforth referred to as ImageNet) and various subsets of ImageNet, which we now describe. We use ImageNet-A to refer to the subset that corresponds to all animals in ImageNet and consists of 398 classes. Out of these 398 classes, 269 classes correspond to wildlife, while 127 classes correspond to domestic animals. ImageNet-NA is the complement of ImageNet-A, comprised of the remaining 602 classes. ImageNet-R is a subset of ImageNet where classes are randomly sampled from ImageNet. ImageNet-X-k is a subset of ImageNet that consists of k classes which are randomly sampled from ImageNet-X (for X $\in$ {A, NA, R}).

iNaturalist-Mini. iNaturalist-Mini is a subset of the iNaturalist2021 dataset. It consists of 10K classes, with 50 examples in each class, for a total of 500K training images. Each class label corresponds to a species of animal, plant, or fungus. For finetuning we use the subset of the training set of iNaturalist-Mini corresponding to animals, resulting in 5388 classes out of 10000 classes, which we refer to as iNat-Mini-A. The test set of iNaturalist-Mini also consists of 50 examples for each class, however the labels are not made publicly available. On the other hand, the validation set has merely 10 examples per class, which we find to be too small a number to gauge performance accurately. Therefore, we construct our own evaluation set by randomly sampling 50 examples per class from iNaturalist2021, ensuring we do not select examples which are present in the train set of iNaturalist-Mini.

We decide to focus on animals for two reasons:

1. We want to isolate and emphasize consequences of ImageNet subset choice in the finetuned model. Finetuning on animals *and* plants instead of *only* animals or plants could make observing these consequences more difficult.
2. The poorer representation of plants compared to that of animals in ImageNet makes plants a non-viable choice for studying the effects of pretraining. There are a total of 398 animal classes, out of which 269 classes correspond to wildlife. In contrast, the representation of "plants" is restricted to $\sim$ 25 classes, with nearly all of them being fruits or vegetables.

Using iNaturalist-Mini instead of the full iNaturalist2021 dataset is a deliberate choice as well – it is a better approximation of the finetuning setting faced in many deployed use cases since iNaturalist2021 contains $\sim$ 2.7M training images, which is more than double ImageNet's $\sim$ 1.3M. At the same time, the diversity of animal species present in iNaturalist-Mini also ensures that the downstream task explored in this work is appropriately challenging.

D.2. Experimental Setup

The goal of our experiments is to observe the effect of ImageNet pretraining on the performance of the model finetuned on a given downstream task. We focus on animal species classification from images, where models are pretrained and then finetuned in a supervised manner.

Pretraining. We conduct all our experiments with ResNet50 networks with batch normalization. We use this architecture since it is a highly standard one across the ecology literature leveraging computer vision for wildlife recognition; its popularity stems from the fact that it is expressive enough for most wildlife recognition datasets, yet small enough that it is accessibly to ecology researchers with limited computational infrastructure. Networks are pretrained with stochastic gradient descent (SGD) and cross-entropy loss for 90 epochs, with a batch size of 256. We employ a step learning rate schedule starting at 0.1, and dropping by $10\times$ every 30 epochs. We also use a momentum factor 0.9 and weight decay factor of 0.0001. The hyperparameters used here are taken from the standard training recipe proposed by He et al. [13].

Finetuning. Our finetuning setup remains largely the same as our setup for pretraining, with a few exceptions to tailor our procedure for each finetuning dataset.

When finetuning on iNat-Mini-A, networks are trained for a total of 60 epochs with the learning rate being dropped by $10\times$ every 20 epochs, and a batch size of 1024. Here, we do not freeze any weights during the finetuning phase, as we empirically find that this severely hurts the final performance of the finetuned model.

When finetuning on ImageNet-A, networks are trained for 15 epochs, with the learning rate being dropped by $10\times$ every 5 epochs. Since we are performing both pretraining and finetuning on subsets of ImageNet, we find that finetuning for a small number of epochs is sufficient. For the same reason, we also freeze all but the final layer of the network during finetuning.