BioBench: A Blueprint to Move Beyond ImageNet for Scientific ML Benchmarks

Samuel Stevens
The Ohio State University
stevens.994@osu.edu

Abstract

ImageNet-1K linear-probe transfer accuracy remains the default proxy for visual representation quality, yet it no longer predicts performance on scientific imagery. Across 46 modern vision transformer checkpoints, ImageNet top-1 accuracy explains only 34% of variance on ecology tasks and mis-ranks 30% of models above 75% accuracy. We present BioBench, an open ecology vision benchmark that captures what ImageNet misses. BioBench unifies 9 publicly released, application-driven tasks, 4 taxonomic kingdoms, and 6 acquisition modalities (drone RGB, web video, micrographs, in-situ and specimen photos, camera-trap frames), totaling 3.1M images. A single Python API downloads data, fits lightweight classifiers to frozen backbones, and reports class-balanced macro-F1 (plus domain metrics for FishNet and FungiCLEF); ViT-L models evaluate in 6 hours on an A6000 GPU. BioBench provides new signal for computer vision in ecology and a template recipe for building reliable AI-for-science benchmarks in any domain. Code and predictions are available at github.com/samuelstevens/biobench and a results website is at samuelstevens.me/biobench .

1 Introduction

Machine learning now drives everything from protein structure prediction to planetary-scale biodiversity surveys, yet progress depends on benchmarks that tell us which models to trust. Vision research still orients around ImageNet-1K, MS COCO, and ADE20K [8, 18, 35], and state-of-the-art claims like vision transformers [9], self-supervised pre-training [19] or image-text pre-training [20] are routinely justified by gains on those leaderboards.

Scientific images, however, are *not* web photographs. Radiographs and histopathology slides emphasize internal or cellular structure [32]; microbiology depends on high-magnification micrographs of microorganisms [21]; and ecology relies on camera-trap or specimen imagery in uncontrolled environments [25, 28]. These sources differ in content, scale, and acquisition method from the datasets that govern general computer vision progress.

The mismatch is not merely cosmetic. Across three publicly released ecology tasks (long-tail species ID [12], drone-video behaviour recognition [16], and specimen trait inference [14]) we measure Spearman's rank correlation coefficient ρ between ImageNet-1K top-1 accuracy and task accuracy for 46 modern computer vision checkpoints spanning supervised [29], self-supervised [19], and image–text [11, 20, 34] pre-training objectives. Once models surpass the now-common 75% ImageNet threshold, Spearman's rank correlation ρ falls below 0.25 (see Fig. 1). Generic benchmark accuracy, long used as a barometer of visual understanding, stops predicting performance on the scientific tasks we measure once models clear the 75% ImageNet top-1 threshold. Other work hints that the same "ranking cliff" afflicts other real-world tasks such as [10, 27]. Because ecological
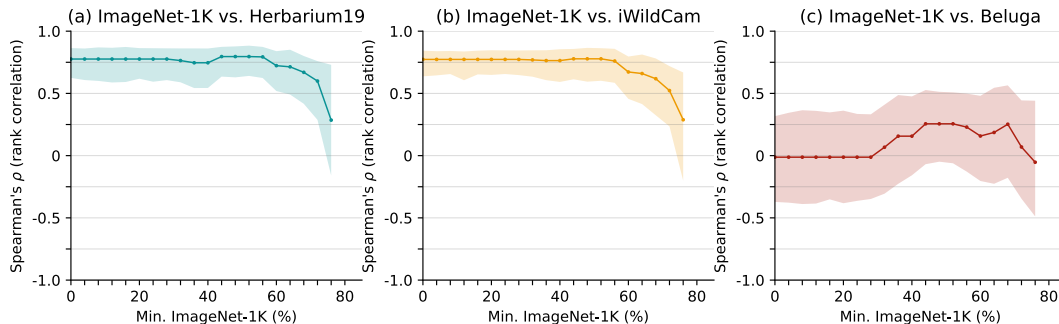
Figure 1: Predictive validity of ImageNet-1K accuracy across (a) species classification of plants [Herbarium19, 23], (b) species classification of animals in camera trap images [iWildcam, 3, 17] and (c) individual identification of beluga whales [Beluga, 1, 5] measured with Spearman's rank correlation coefficient ρ between ImageNet-1K and task rankings, computed across all checkpoints with ImageNet Top-1 accuracy $\geq T\%$ (x-axis). Shaded region shows 95% bootstrapped confidence intervals. **ImageNet-1K fails to predict model rankings on specific tasks as models improve.**

domains offer both scientific diversity and abundant open data, they provide an ideal testbed to systematically investigate how benchmark predictivity fails under realistic distributional shift.

ImageNet fails for two intertwined reasons. First, *distribution mismatch*: its RGB web photos share neither spectrum nor noise with camera-trap infrared, multi-spectral drone passes, or microscope slides, so models optimized for ImageNet seldom work for scientific imagery. Second, scientific tasks are fine-grained and long-tailed: ecologists distinguish thousands of insect species, pathologists dozens of rare tumor sub-types; ImageNet's 1,000 classes contain few such subtle distinctions and are heavily skewed toward frequent objects. Together these gaps explain why increasing ImageNet accuracy ceases to improve performance once models venture into application-driven tasks [22].

The obvious remedy is to benchmark models on the applications themselves. When tasks are drawn directly from practice, their image distributions align by construction, their labels inherit the domain's natural granularity, and their objectives mirror the questions scientists actually ask. Many fields still lack shared datasets of this sort, but ecology is an exception: years of CV4Ecology challenges have produced public tasks for species identification, behavior recognition, and trait inference. By consolidating these efforts into a single suite we can test whether application-driven benchmarks restore predictive power and provide a template for other scientific domains.

Why, then, has no unified benchmark appeared? Because three hurdles discouraged even the most committed researchers. First, fragmentation: every ecology dataset shipped in its own repository with idiosyncratic file trees, splits, and metric scripts. Second, perceived sufficiency: most vision researchers assumed that strong ImageNet accuracy, averaged over scattered per-task leaderboards, already served as an adequate proxy, so consolidating tasks seemed low-yield. Finally, non-overlapping waves of progress: benchmarks surfaced one at a time; every release compared against the "best" backbone of that moment and the authors' favorite tricks. Because checkpoints, hyper-parameters, and evaluation scripts kept changing, nobody could tell whether any single model genuinely excelled across camera traps, drone footage, and specimen photographs simultaneously.

We therefore introduce BioBench, a domain-grounded vision benchmark of 9 application-driven tasks that span 4 taxonomic kingdoms (animals, plants, fungi, and protozoa) captured from 6 distinct image distributions: drone footage, curated web video, microscope micrographs, in-situ RGB photos, RGB specimen images, and camera-trap frames (see Section 2 and Table 1 for more details). The corpus contains 3.1M images (337 GB). Evaluation reports macro-F1 for every task, with two tasks (FungiCLEF and FishNet) scored by their domain-standard metrics. Each dataset downloads via a single-file Python script (fully documented). Evaluation parallelizes across SLURM clusters or runs on a single GPU; ViT-B/16 and ViT-L/14 checkpoints finish in about one hour on an NVIDIA A6000, with larger models scaling predictably.

A linear fit over 46 pre-trained vision checkpoints underscores the need for new benchmarks. Across all checkpoints, ImageNet explains only one-third of BioBench variance ($R^2 = 0.34$) and agrees in rank just $\rho = 0.55$, meaning the ImageNet-preferred model is actually worse on BioBench roughly
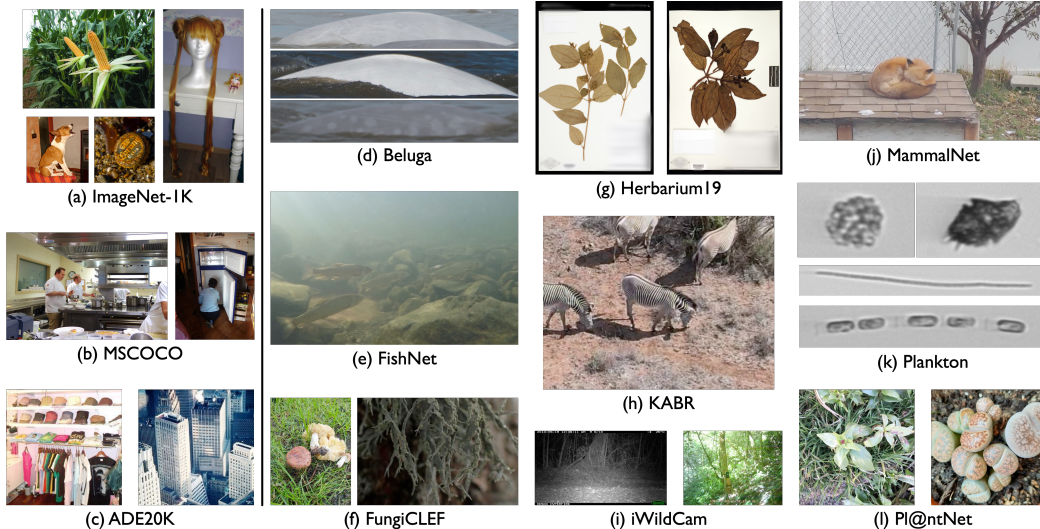
Figure 2: **Left (a-c):** Random example images from ImageNet-1K, MSCOCO and ADE20K, three popular general-domain vision benchmarks [8, 18, 35]. **Right (d-l):** Random example images from each of the nine tasks in BioBench. **Tasks in BioBench have radically different image distributions compared to general-domain vision benchmarks.**

22% of the time.[1] The mismatch widens at the frontier: among models above 75% on ImageNet, rank concordance drops to $\rho = 0.42$, so the supposed "best" model is mis-ranked 30% of the time.[2] These numbers make one conclusion unavoidable: web-photo leaderboards have ceased to be a trustworthy proxy for progress in scientific AI. BioBench stands as a proof-of-concept, showing how domain workflows, long-tail metrics, and modality stress tests can be distilled into a single, open benchmark, and points the way toward equally realistic suites for medicine, manufacturing, and every other data-rich science.

2 Benchmark Suite & Protocol

An effective ecological vision benchmark must address fundamental limitations in existing evaluation frameworks. First, it requires diversity across multiple dimensions: taxonomic breadth spanning microorganisms to mammals; varied image regimes from microscopy to camera traps; task diversity beyond simple classification; and natural class imbalances reflecting real-world species distributions. Second, it must balance proxy-driven tasks (measuring general capability) with mission-driven tasks (assessing operational utility for conservation applications). Third, it must provide rigorous statistical tools (confidence intervals, significance testing, and rank stability analysis) to distinguish genuine performance differences from benchmark lottery effects.

Neither ImageNet-1K [8] nor iNat2021 [26] satisfies these requirements.

ImageNet lacks ecological diversity, while iNat2021 offers taxonomic breadth but limited task variety and no mission-driven evaluation. Most critically, our analysis reveals that once models exceed 75% accuracy on ImageNet, the benchmark loses predictive power for ecological performance (ρ drops from 0.82 to 0.55), rendering it insufficient as a proxy for ecological vision capability. BioBench addresses these limitations through a minimal embedding interface that dramatically reduces integration overhead while providing comprehensive coverage across the ecological axes that matter most.

Tasks. BioBench consolidates 9 public, application-driven tasks spanning 4 kingdoms (animals, plants, fungi, protists) and 6 image regimes (camera-trap RGB/IR, drone video frames, museum

---

[1] R^2 has a 95% confidence interval of $[0.20, 0.58]$; ρ has a 95% confidence interval of $[0.45, 0.64]$; both are significant with $p < 0.0005$ via 5,000-perm randomization. Mis-ranking probability is $\frac{1}{2}(1 - \rho)$.

[2] ρ for $> 75\%$ has a 95% confidence interval of $[0.15, 0.65]$ and is significant with $p < 0.01$ via 5,000-perm randomization.

Table 1: Datasets across key dimensions that distinguish general computer vision benchmarks from ecological vision tasks. *Mission* tasks serve a specific ecological application (✓) rather than a general benchmark purpose (✗). †*Context* indicates whether images show organisms in their natural environment (in-situ) or as preserved specimens. "Target" indicates the classification target. **Takeaway:** ImageNet-1K fundamentally differs from other ecological tasks because it is taxonomically unrestricted and web-scraped rather than scientifically curated.

Name	Mission?*	Taxon	Source	Context†	Target
ImageNet-1K	✗	-	Web-scraped	-	Object
iNat2021	✗	Diverse	Citizen science	In-situ	Species
NeWT	✗	Diverse	Citizen science	In-situ	Varied
BelugaID	✓	*D. leucas*	Citizen science	In-situ	Individuals
FishNet	✗	Fish	Natural collections	Specimen	Functional Traits
FungiCLEF	✗	Fungi	Citizen science	In-situ	Species
Herbarium19	✓	Plants	Natural collections	Specimen	Species
iWildCam21	✓	Mammals	Research studies	In-situ	Species
KABR	✓	Mammals	Research study	In-situ	Behaviors
MammalNet	✗	Mammals	Web-scraped	In-situ	Behaviors
Plankton	✓	Protists	Research study	In-situ	Species
Pl@ntNet	✗	Plants	Citizen science	In-situ	Species

specimens, in-situ macro, web video, micrographs), totaling 3.1M images. Tasks cover species ID, individual re-ID, behavior classification, and functional trait prediction. Example images are in Fig. 2 and task summaries are in Table 1.

Implementation. Models implement one contract $f : \text{image} \to \mathbb{R}^d$ (frozen embeddings). We fit linear or logistic probes per task, report macro-F1 by default (FishNet and FungiCLEF use task-specific metrics), and bootstrap confidence intervals.

Design Goals. Embrace distributional diversity, evaluate long-tail class balance explicitly, and isolate representation quality from task-specific engineering via a uniform probing protocol.

3 Benchmark Results

We evaluate 46 pre-trained vision models across 11 model families on BioBench. We use a single Nvidia A6000 GPUs to evaluate all models; we will release both the individual model predictions and the aggregate statistics upon acceptance. The results for each model family's top checkpoint are in Table 2; results for all checkpoints are publicly available at samstevens.me/biobench. Our analysis throughout this work considers all checkpoints.

Across 46 checkpoints, ImageNet-1K top-1 accuracy explains only $R^2 = 0.34$ of BioBench variance; rank concordance is $\rho = 0.55$ overall and drops further above 75% ImageNet (Fig. 1). Thus, the ImageNet-preferred model is worse on BioBench **roughly** 30% **of the time at the frontier**.

We measure progress over BioBench over time in Fig. 3; despite general performance claims from many released generalist models, only CLIP [20], SigLIP [34] and SigLIP 2 [24] set new state-of-the-art scores on BioBench.

4 Related Work

General-domain benchmarks (ImageNet, COCO, ADE20K) catalyzed vision progress but are fragile under distribution shift and long-tail structure. Transfer suites such as VTAB and Taskonomy assess representation reuse but contain little ecological content. Transfer suites such as VTAB [spans 19 tasks across diverse domains 33] or Taskonomy [26 visual tasks 31] assess representation reuse across domains. However, these benchmarks include minimal ecological content and fail to capture the specific challenges of biodiversity monitoring: fine-grained taxonomic distinctions, extreme environmental variability, and long-tailed species distributions.

Table 2: An overview of each model family's top-performing model on ImageNet-1K, NeWT and all tasks in BioBench. State-of-the-art results for each task, along with their source, are reported at the bottom. "-" indicates no published state-of-the-art result. **Mean** is across all tasks in BioBench (not ImageNet-1K or NeWT). †Micro-accuracy (SOTA), not macro-F1 (ours). ‡Macro-accuracy (SOTA), not macro-F1 (ours).

Family	Architecture	Res. (px)	ImageNet-1K	NeWT	Beluga	FishNet	FungiCLEF	Herbarium19	iWildCam	KABR	MammalNet	Plankton	Pl@ntNet	Mean
CLIP	ViT-L/14	336	83.9	83.6	2.8	64.4	27.7	53.6	23.2	52.2	62.8	3.7	40.4	36.7
SigLIP	SO400M/14	384	87.8	86.0	4.0	69.0	38.6	63.7	25.7	**59.3**	66.3	4.0	47.4	42.0
SigLIP 2	ViT-1B/16	384	**88.9**	86.7	3.6	70.7	39.0	65.2	29.3	58.4	**73.9**	4.0	47.9	**43.5**
DINOv2	ViT-g/14	224	86.7	82.8	4.5	**75.2**	34.2	64.3	**30.5**	53.7	57.1	**4.2**	51.5	41.7
AIMv2	ViT-3B/14	448	86.7	84.0	1.7	59.2	34.4	48.3	20.5	**58.9**	68.8	4.0	36.7	36.9
SAM 2	Hiera Large	1024	33.9	64.2	3.1	45.8	16.1	12.7	5.4	38.5	33.7	3.2	9.5	18.7
V-JEPA	ViT-H/16	224	49.0	68.0	**9.2**	50.7	20.8	13.4	6.0	47.4	38.2	3.2	17.5	22.9
BioCLIP	ViT-B/16	224	58.5	82.7	4.6	62.6	40.6	52.6	17.2	46.1	35.7	3.8	45.4	34.3
BioCLIP 2	ViT-L/14	224	80.0	**89.1**	3.0	71.8	**51.0**	**73.1**	24.7	48.0	46.4	3.9	**53.8**	41.7
BioTrove	ViT-B/16	224	45.3	82.9	3.7	59.7	41.6	47.0	11.1	37.3	30.0	3.8	48.1	31.4
MegaDesc.	Swin-L/4	384	49.9	71.3	8.0	50.2	22.1	14.0	6.9	32.3	31.1	2.1	17.6	20.5
Random Prediction			0.1	50.0	0.1	47.9	13.9	0.1	0.5	12.5	8.3	2.1	0.1	9.5
Task-Specific State-of-the-Art			91.0	80.6	66.5	81.7	-	†89.9	66.7	65.8	‡37.8	-	-	
[Source]			[30]	[26]	[4]	[14]		[23]	[2]	[15]	[6]			

iNaturalist [26] provides fine-grained species classification but doesn't incorporate temporal behavior or ecological trait prediction. Pl@ntNet [12] focuses exclusively on plant identification. WILDS [17] includes iWildCam [3] for camera trap imagery but treats ecological monitoring as just one of many domains rather than exploring its multi-faceted challenges.

These isolated efforts highlight the critical need for BioBench: conservation practitioners currently lack systematic guidance on which vision architectures best transfer to the complex, interconnected tasks comprising ecological monitoring workflows.

Methodological work [7, 13] highlights the importance of consistent protocols—an ethos we adopt via a single embedding API, class-balanced metrics, and bootstrap uncertainty.

5   Limitations & Future Work

While we argue that BioBench meaningfully improves the state of ecological benchmarking and offers lessons applicable to other scientific domains, we have not explored every aspect. **Limited Scope.** We focus on ecology; medicine and manufacturing may emphasize different tasks (e.g., detection/segmentation, calibration). **Frozen features.** Probing isolates representation quality but underestimates task-specific fine-tuning gains. **Metrics.** Macro-F1 rewards tail performance; some applications prefer operating-point metrics (e.g., precision@recall).

Figure 3: BioBench scores over time. The majority of new models fail to improve on BioBench.

BioBench shows that ImageNet-driven model choice is unreliable for scientific imagery and offers a minimal, reproducible recipe to evaluate models where it matters. We hope BioBench serves both as a practical guide for ecological workflows and as a template for building equally grounded benchmarks in other sciences.
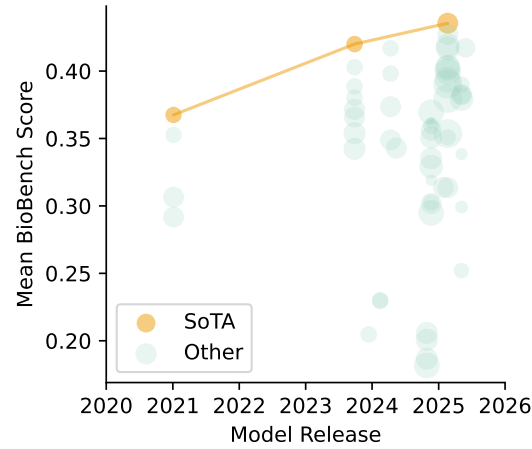
Acknowledgments and Disclosure of Funding

References

[1] Aleksandr Algasov, Ekaterina Nepovinnykh, Tuomas Eerola, Heikki Kälviäinen, Charles V Stewart, Lasha Otarashvili, and Jason A Holmberg. Understanding the impact of training set size on animal re-identification. *arXiv preprint arXiv:2405.15976*, 2024.

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[4] Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, 2024.

[5] Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, 2024.

[6] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13061, 2023.

[7] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[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.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *Advances in Neural Information Processing Systems*, 36:25050–25080, 2023.

[11] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024.

[12] Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS Datasets and Benchmarks 2021*, 2021.

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

[14] Faizan Farooq Khan, Xiang Li, Andrew J. Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20496–20506, October 2023.

[15] Maksim Kholiavchenko, Jenna Kline, Maksim Kukushkin, Otto Brookes, Sam Stevens, Isla Duporge, Alec Sheets, Reshma R Babu, Namrata Banerji, Elizabeth Campolongo, et al. Deep dive into kabr: a dataset for understanding ungulate behavior from in-situ drone video. *Multimedia Tools and Applications*, pages 1–20, 2024.

[16] Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, et al. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 31–40, 2024.

[17] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[21] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[22] David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L. Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, and Adam White. Position: Application-driven innovation in machine learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42707–42718. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/rolnick24a.html.

[23] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

[24] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[25] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[26] 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*, pages 12884–12893, 2021.

[27] Kirill Vishniakov, Zhiqiang Shen, and Zhuang Liu. Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://icml.cc/virtual/2024/poster/34818.

[28] Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3): 533–545, 2018.

[29] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

[30] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[31] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[32] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[33] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019.

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.