Benchmarking Representation Learning for Natural World Image Collections

Grant Van Horn[1] Elijah Cole[2] Sara Beery[2] Kimberly Wilber[3] Serge Belongie[1,3] Oisin Mac Aodha[4]
[1]Cornell University [2]Caltech [3]Google [4]University of Edinburgh
www.github.com/visipedia/newt

Abstract

Recent progress in self-supervised learning has resulted in models that are capable of extracting rich representations from image collections without requiring any explicit label supervision. However, to date the vast majority of these approaches have restricted themselves to training on standard benchmark datasets such as ImageNet. We argue that fine-grained visual categorization problems, such as plant and animal species classification, provide an informative testbed for self-supervised learning. In order to facilitate progress in this area we present two new natural world visual classification datasets, iNat2021 and NeWT. The former consists of 2.7M images from 10k different species uploaded by users of the citizen science application iNaturalist. We designed the latter, NeWT, in collaboration with domain experts with the aim of benchmarking the performance of representation learning algorithms on a suite of challenging natural world binary classification tasks that go beyond standard species classification. These two new datasets allow us to explore questions related to large-scale representation and transfer learning in the context of fine-grained categories. We provide a comprehensive analysis of feature extractors trained with and without supervision on ImageNet and iNat2021, shedding light on the strengths and weaknesses of different learned features across a diverse set of tasks. We find that features produced by standard supervised methods still outperform those produced by self-supervised approaches such as SimCLR. However, improved self-supervised learning methods are constantly being released and the iNat2021 and NeWT datasets are a valuable resource for tracking their progress.

1. Introduction

Learning representations of images through self-supervision alone has seen impressive advancement over the last few years. There are tantalizing results that show self-supervised methods, fine-tuned with 1% of the training labels, reaching the performance of their fully supervised counterparts [9]. In many domains, aggregating large amounts of data is typically not the bottleneck. Rather, it is the subsequent labeling of that data that consumes vast amounts of money and time. This is further compounded in fine-grained domains, e.g. medicine or the natural world, where sufficiently well trained annotators are few or their time is expensive. If the benefits of self-supervised learning come to full fruition, then the applicability and impact of computer vision models across many domains will see a rapid increase.

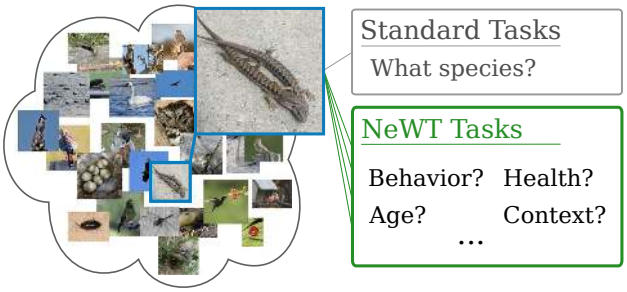

Figure 1. Existing fine-grained image datasets are typically focused on a single task e.g. species identification. As natural world media collections grow, we have the opportunity to extract information beyond species labels to answer important ecological questions. For example, with the help of community scientists, researchers from the NHMLA were able to curate over 500 images of alligator lizards mating, a phenomenon seldomly recorded in the existing scientific literature [18]. We analyze if trained feature extractors can answer similar novel image understanding questions with minimal additional training and present **NeWT**, a diverse benchmark of natural world visual understanding tasks such as animal health, life-stage, behavior, among others.

One particular domain that is well suited for this type of advancement is the study of the natural world through photographs collected by communities of enthusiasts. Websites such as iNaturalist [1] and eBird [54] amass large collections of media annually. To date, there are 60M images in iNaturalist spanning the tree of life and 25M images of birds from around the world in eBird, both representing point-in-time records of wildlife. Identifying the species in an image has been well studied by the computer vision com-

munity [60, 32, 3, 58], however this is only the tip of the iceberg in terms of questions one may wish to answer using these vast collections. These datasets contain evidence of the health and state of the individuals depicted, along with their behavior. Having an automated system mine this data for these types of properties could help scientists fill in missing pieces of basic natural history information that are crucial for our understanding of global biodiversity and help measure the loss of biodiversity due to human impact [6].

To give one example, science is ignorant to the nesting requirements of thousands of bird species, including the vulnerable Pink-throated Brilliant (Heliodoxa gularis) [67]. Knowing how and where this species builds its nest is a crucial piece of information needed when discussing conservation based interventions, particularly as it pertains to the ability of this species to exist in degraded and fragmented habitats [67]. While nothing can replace the capabilities of a biologist in the field, citizen science projects like eBird and iNaturalist are collecting raw images that could help answer some of these questions. However, herein lies the problem. It is currently a daunting task to label training datasets for these specialized questions that would satisfy the data appetite of an off-the-shelf deep network.

Self-supervised learning is one potential solution that could alleviate the labeling burden by taking advantage of large media collections. While most research on self-supervised learning focuses on ImageNet [52], in this work we expand these techniques to the natural world domain and fine-grained classification. Following Goyal *et al*. [20], we maintain that a good representation should generalize to many different tasks, with limited supervision or finetuning. We do not investigate self-supervised learning as an initialization scheme for a model that is further optimized and finetuned, but rather as a way to learn feature representations themselves. Importantly, [20] point out that self-supervised feature learning and subsequent feature evaluation on the same dataset does not test the generalization of the features. Inspired by this, we present a new large-scale pretraining dataset and new benchmark tasks specifically designed to enable us to ask questions about the generalization of self-supervised learning on natural world image collections.

We make the following three contributions:

- **iNat2021** - A new large-scale image dataset collected and annotated by community scientists that contains over 2.7M images from 10k different species.
- **NeWT** - A new suite of 164 challenging natural world visual benchmark tasks that are motivated by real world image understanding use cases.
- A detailed evaluation of self-supervised learning in the context of natural world image collections. We show that despite recent progress, self-supervised features still lag behind supervised variants.

2. Related Work

2.1. Learning Visual Representations

Transfer learning using features extracted from deep networks that have been trained via supervision on large datasets results in powerful features that can be applied to many downstream tasks [14, 63]. However, there is evidence to suggest that pretraining on datasets such as ImageNet [52] is less effective on fine-grained categories when the labels are not well represented in the source dataset [34]. Self-supervised learning, i.e. learning visual representations without requiring explicit label supervision, is an exciting research area that, if successful, could provide a much more scalable way to learn representations for a wide variety of tasks – including fine-grained ones.

Earlier work in self-supervised learning in vision involved framing the learning problem via proxy tasks e.g. predicting context from image patches [13, 49], image colorization [65], or predicting image rotation [19], to name a few. The most effective recent approaches have focused on contrastive learning based training objectives [25, 24], where the aim is to learn features from images such that augmented versions of the same image are nearby in the feature space, and other images are further away. This can require a large batch size during training to ensure that there are a sufficient number of useful negatives [8] – which necessitates large compute resources during training. Recent advances include memory banks to address the need for large batches [61, 26, 10], additional embedding layers [9], and more advanced augmentations [7], among others.

In our experiments, we compare the performance of several leading self-supervised learning algorithms [8, 10, 7, 9] to conventional supervised learning in the context of fine-grained pretraining to try to understand what gap, if any, exists between the features learned by these very different paradigms on natural world image classification tasks.

2.2. Benchmarking Representation Learning

Like Cui *et al*.[12], we are also interested in understanding how well models trained on large-scale natural world datasets can transfer to downstream fine-grained tasks. However, [12] only explored transfer learning using fully supervised, as opposed to self-supervised, training. [53] combined self-supervised and meta learning and showed improved few-shot classification accuracy for fine-grained categories. Instead of jointly training our models, we decouple feature learning from classification so that we can better understand generalization performance.

Our work can be seen as a continuation of recent attempts to benchmark the performance of self-supervised learning e.g. [20, 33, 64]. We swap out their pretext tasks for more recent approaches and utilize natural world evaluation datasets containing a mix of fine and coarse-grained

visual concepts to test the generalization of the learned features. This is in contrast to standard computer vision datasets or synthetic tasks [46] that are commonly used for evaluation.

The majority of existing self-supervised methods train on ImageNet [52]. There are some exceptions, such as [20] and [21] that also train on alternative datasets such as YFCC100M [55] and Places205 [66], respectively. We present results obtained by learning representations obtained through self-supervision alone on a large-scale natural world dataset – as opposed to just linear evaluation [45, 7, 15] or finetuning in this domain [26].

2.3. Fine-Grained Datasets

The vision community is not lacking in image datasets. The set of existing datasets include those that are large-scale and span broad category groups e.g. [52, 35], through to smaller, but densely annotated, ones e.g. [16, 40, 38, 23]. In addition, there are a number of domain specific (i.e. "fine-grained") datasets covering object categories such as airplanes [44, 59], birds [60, 3, 57, 36], dogs [32, 51, 42], fashion [31], flowers [47, 48], food [4, 28], leaves [39], vehicles [37, 41, 62, 17], and, of course, human faces [29, 50, 22, 5]. Most closely related to our work are the existing iNaturalist species classification datasets [58, 2], which contain a set of coarse and fine-grained species classification problems.

Distinct from these existing datasets, our new NeWT dataset presents a rich set of evaluation tasks that are not solely focused on one type of visual challenge e.g. species classification. Instead, NeWT contains a wide variety of tasks encompassing behavior, health, context, among others. Most importantly, our tasks are informed by natural world domain experts and are thus grounded in real-world use cases. Paired with our new iNat2021 dataset, which contains five times more training images and nearly 20% more categories than the largest previous version [58], they serve as a valuable tool to enable us to better understand and evaluate progress in both transfer and self-supervised learning in challenging visual domains.

3. The iNaturalist 2021 Dataset

3.1. Dataset Overview

While several large-scale natural world datasets already exist, the current largest one, iNat2017 [58], only contains half the number of training images as ImageNet [52]. To better facilitate research in representation learning for this domain, we introduce a new image dataset called iNat2021. iNat2021 consists of 2.7M training images, 100k validation images, and 500k test images, and represents images from 10k species spanning the entire tree of life. In addition to its overall scale, the main distinguishing feature of iNat2021 is that it contains at least 152 images in the training set for each species. We provide a comparison to existing datasets in Table 1 and a breakdown of the image distribution in Table 3. Unlike previous iterations, we have split the training and testing images in iNat2021 by a specific date and have allowed a particular photographer to have images in both the train and test splits. There is an intuitive interpretation to this decision: we are retroactively building a computer vision training dataset, composed of data that was submitted *over a year ago*, to classify the most observed species in the *last year*, which is our test set. While there are many ways we could have decided the train and test split criteria, we believe this is particularly natural and lends itself well to future updates (the date split simply increases by a year). A detailed description of the steps we took to create the dataset are outlined in the supplementary material.

In addition to the full sized dataset, we have also created a smaller version (iNat2021 mini) that contains 50 training images per species, sampled from the full train split. These two different training splits allows researchers to explore the benefits of training algorithms on five times more data. The mini dataset also keeps the training set size reasonable for desktop-scale experiments. In addition to the images themselves, we also include latitude, longitude, and time data for each, facilitating research that incorporates additional meta data to improve fine-grained classification accuracy, e.g. [43, 11].

3.2. Comparisons to iNat2017-2019

In Table 1 we compare the new iNat2021 dataset with previous datasets built from iNaturalist. iNat2017 was the first large-scale species classification dataset [58]. iNat2018 addressed the long tail problem inherent in large-scale media repositories. iNat2019 attempted to focus specifically on genera with large number of species (at least 10), resulting in a smaller dataset consisting of many 10-way fine-grained classification problems. Our iNat2021 dataset is similar to iNat2017 and iNat2018 in terms of its large-scale scope, however we incorporate the iNat2019 style focus on fine-grained challenges with our introduction of the NeWT collection of evaluation datasets, see Section 4. While we have effectively removed the long tail training distribution that was the focus of other iNat datasets, we have included sufficient images per species where this phenomena can still be studied by systematically removing data. More data per species has the effect of decreasing the difficulty of iNat2021 in the purely supervised setting, but we believe that the additional images for each category are essential to enable us to systematically evaluate the effectiveness of self-supervised learning for natural world visual categories.

dataset	# classes	# train	# val	# test	min # ims	max # ims	avg # ims
iNat2017 [58]	5,089	579,184	95,986	182,707	9	3919	114
iNat2018 [2]	8,142	437,513	24,426	149,394	2	1,000	54
iNat2019 [2]	1,010	265,213	3,030	35,350	16	500	263
iNat2021 mini	10,000	500,000	*100,000	*500,000	50	50	50
iNat2021	10,000	2,686,843	*100,000	*500,000	152	300	267

Table 1. Comparison of iNat2021 dataset to previous iterations. iNat2021 is more than five times larger than existing large-scale species classification datasets, making it a valuable tool for benchmarking representation learning. Min, max, and avg refer to the number of images per class in the respective training sets. *Both variants of iNat2021 use the same validation and test sets.

train split	top-1	top-2	top-3	top-4	top-5
iNat2021 mini	0.654	0.759	0.806	0.833	0.851
iNat2021	0.760	0.848	0.882	0.901	0.914
iNat2021 mini *	0.616	0.722	0.769	0.798	0.818
iNat2021 *	0.746	0.836	0.872	0.891	0.904

Table 2. Top-K Accuracy on the iNat2021 test set. Models marked with a * have been initialized with random weights, otherwise ImageNet initialization is used.

	Iconic Group	Species Count	Train Images	Full ACC	Mini ACC
	Insects	2,526	663,682	0.813	0.715
	Fungi	341	90,048	0.786	0.707
	Plants	4,271	1,148,702	0.800	0.692
	Mollusks	169	44,670	0.756	0.670
	Animalia	142	37,042	0.747	0.654
	Fish	183	45,166	0.725	0.640
	Arachnids	153	40,687	0.704	0.582
	Birds	1,486	414,847	0.662	0.537
	Mammals	246	68,917	0.590	0.496
	Reptiles	313	86,830	0.554	0.430
	Amphibians	170	46,252	0.526	0.417

Table 3. Number of species, training images, and mean test accuracy in iNat2021 for each iconic group. 'Animalia' is a catch-all category that contains species that do not fit in the other iconic groups. For the mini train split, each species has 50 train images.

3.3. Baseline Supervised Experiments

We train ResNet50 [27] networks, both with and without ImageNet initialization, to benchmark the performance of iNat2021. Table 2 shows the top-k accuracy achieved when training using the full and mini datasets, and Table 3 shows the top-1 accuracy broken down by iconic groups. The model trained on the mini dataset results in a top-1 accuracy of 65.4%, while the full model achieves 76.0%, showing that an increase from 500k training images to 2.7M results in an $\sim$11 percentage point increase in accuracy. The corresponding top-1 results for the validation set are 65.8% and 76.4%. On average, insects are the best performing iconic group, and amphibians are the worst performing group. While these average statistics are interesting, we do not believe they demonstrate that insects are necessarily "easier" to identify than amphibians. We are most likely seeing a bias in the iNat2021 dataset. Perhaps, on average, it is easier to take a close-up photograph of an insect than it is to photograph an amphibian. Or perhaps the amphibian species have more visual modalities than insects. Finally, we observe that models trained from randomly initialized weights perform slightly worse than those trained from ImageNet initialization, but the gap closes when training on the full dataset.

4. NeWT: Natural World Tasks

Large media repositories, such as Flickr, the Macaulay Library, and iNaturalist have been utilized to create species classification datasets such as CUB [60], BirdSnap [3], NABirds [57], and the collection of iNaturalist competition datasets [58]. These datasets have become standard experimental resources for computer vision researchers and have been used to benchmark the progress of classification models over the last decade. Improvements on these datasets have in turn led to the incorporation of these models into useful applications that assist everyday users in recognizing the wildlife around them, e.g. [39, 30, 56]. However, there are far more questions that biologists and practitioners would like to ask of these large media repositories in addition to "What species is in this photo?" For example, an ornithologist may like to ask, "Does this photo contain a nest?" or "Does this photo show an adult feeding a nestling?" Similarly, a herpetologist may like to ask, "Does this photo show mating behavior for the Southern Alligator Lizard?" Researchers can certainly answer these questions themselves for a few images. The problem is the scale of these archives, and the fact that they are continually growing. Can a computer vision model be used to answer these questions? While we do not have large collections of datasets labeled with nests or eggs or mating behavior, we do have large-scale species classification datasets. This raises the question about the adaptability of a model trained for species classification to these new types of questions. Similarly, with the recent advances in self-supervised learning there is the potential for a self-supervised model to be readily adapted to answer these varied tasks. To help address these questions we have constructed a collection of Natural World Tasks (NeWT) that can be used to benchmark current representation learning methods.

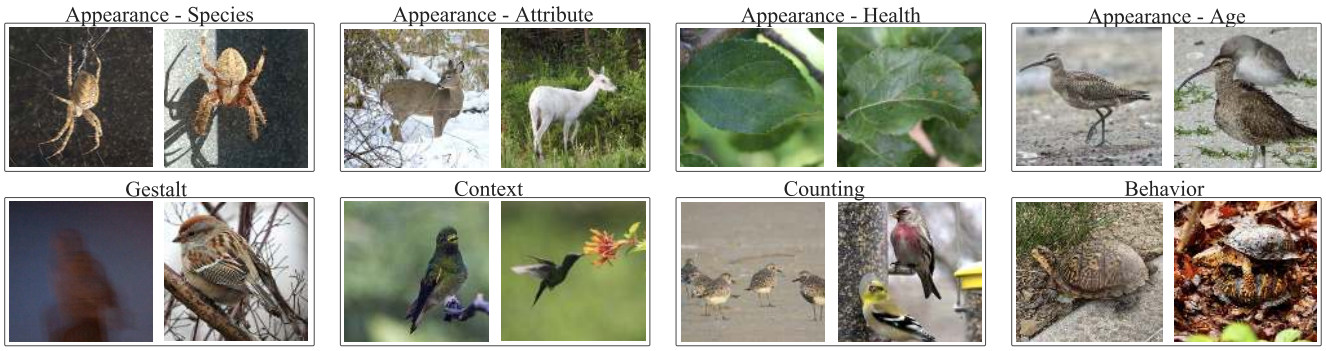

Figure 2. Example image pairs from a binary classification task within each coarse task grouping of the NeWT dataset

NeWT is comprised of 164 highly curated binary classification tasks sourced from iNaturalist, the Macaulay Library, and the NABirds dataset, among others. No images from NeWT occur in the iNat2021 training dataset, and the images in tasks not sourced from iNaturalist are reasonably similar to images found on iNaturalist. This makes the iNat2021 dataset a perfect pretraining dataset for NeWT. Unlike some of the potential data quality issues found in iNat2021 (see supplementary material), each task in NeWT has been vetted for data quality with the assistance of domain experts. While species classification still plays a large role in NeWT (albeit reduced down to difficult fine-grained pairs of species), the addition of other types of tasks makes this dataset uniquely positioned to determine how well different pretrained models can answer various natural world questions. Each task has approximately uniform positive and negative samples, as well as approximately uniform train and test samples. The size of each task is modest, on the order of 50-100 images per class per split (for a total of 200-400 images per task), which makes them very convenient for training and evaluating linear classifiers. We have coarsely categorized the tasks into eight groups (see Figure 2 for visual examples) with the total number of binary tasks per group in parentheses:

- **Appearance - Age (14)** Tasks where the age of the species is the decision criteria, e.g. "Is this a hatch-year Whimbrel?"
- **Appearance - Attribute (7)**: Tasks where a specific attribute of an organism is used to make the decision, e.g. "Is the deer leucistic?"
- **Appearance - Health (9)**: Tasks where the health of the organism is the decision criteria, e.g. "Is the plant diseased?"
- **Appearance - Species (102)**: Tasks where the goal is to distinguish two visually similar species. This can include species from iNat2021, but with new, unseen training data, and tasks from species not included in iNat2021.
- **Behavior (16)** Tasks where the evidence of a behavior is the decision criteria, e.g. "Are the lizards mating?"
- **Context (8)** Tasks where the immediate or surrounding context of the organism is the decision criteria, e.g. "Is the hummingbird feeding at a flower?"
- **Counting (2)** Tasks where the number of specific instances is the decision criteria, e.g. "Are there multiple bird species present?"
- **Gestalt (6)** Tasks where the quality, composition, or type of photo is the decision criteria, e.g. "Is this a high quality or low quality photograph of a bird?"

5. Experiments

Here we present an analysis of different learned image representations trained on multiple datasets and evaluate their effectiveness on existing fine-grained datasets and NeWT.

5.1. Implementation Details

Given a specific configuration of {feature extractor, pretraining dataset, training objective}, our feature representation evaluation protocol is the same for all experiments. Every experiment uses the ResNet50 [27] model as the feature extractor, with some experiments modifying the width multiplier parameter of the network to 4. We consider ImageNet, iNat2018, iNat2021, and the iNat2021 mini dataset for the pretraining dataset. The training objective can either be a supervised classification loss (standard cross-entropy) or one of the following self-supervised objectives: SimCLR [8], SimCLR v2 [9], SwAV [7], or MoCo v2 [10].

The supervised experiments using iNat2021 mini and iNat2018 are trained for 65-90 epochs, starting from ImageNet initialization, and we used the model checkpoint that performed the best on the respective validation set. The supervised experiments using iNat2021 were trained for 20 epochs, also starting from ImageNet initialization. For self-supervised techniques pretrained on ImageNet, we make use of model checkpoint files accompanying the official implementation of the method. For models self-supervised on iNat datasets we used default parameters from the respective techniques unless otherwise stated. Our experiments using SimCLR v2 on iNat datasets do not incorporate knowledge distillation from a larger network nor the MoCo

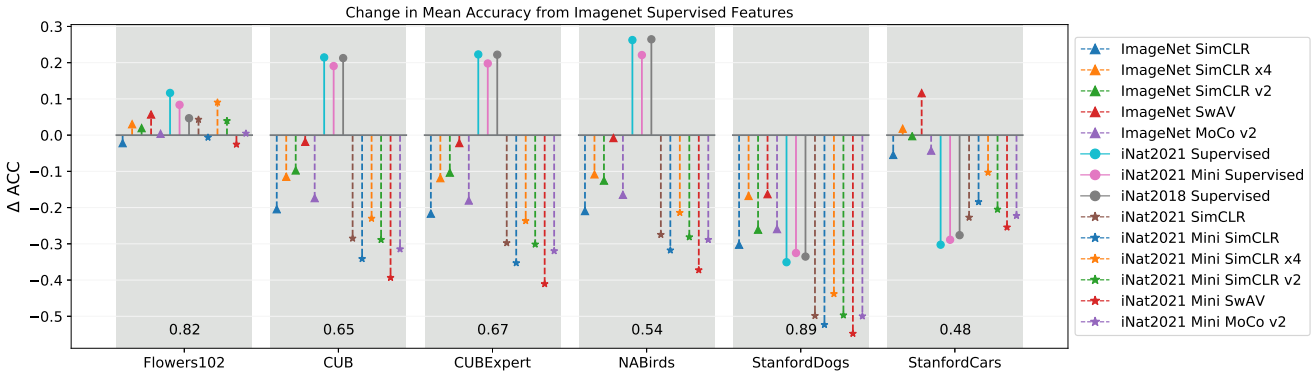

Figure 3. **Fine-grained evaluation.** The mean top-1 accuracy difference between "off-the-shelf" supervised ImageNet features and other pretraining strategies on existing fine-grained datasets. For context, the accuracy of the ImageNet features are printed above the dataset labels along the x-axis. All methods utilize a ResNet50 backbone architecture, and all experiments use features extracted by the last convolution block (dim=2048) to train a linear SVM using SGD (x4 models have dim=8192). Techniques that make use of supervised pretraining have a solid stem line, while techniques that use self-supervision for pretraining have a dashed stem line. Techniques that utilize ImageNet have a triangle marker, techniques that utilize an iNat dataset with supervision have a circle marker, and techniques that utilize an iNat dataset with a self-supervision training objective have a star marker. Several patterns are apparent: (1) Self-supervised methods rarely do better than "off-the-shelf" supervised ImageNet features. (2) Pretraining on iNat datasets with supervision leads to better results on downstream tasks that contain categories similar to those found in iNat datasets (i.e. flowers and birds), but this does not hold for self-supervised objectives. (3) Self-supervised models trained on ImageNet do better than their iNat counterparts. For detailed accuracy numbers see the supplementary material.

style memory mechanism; instead we train the ResNet50 backbone using a 3-layer projection head instead of the 2-layer projection head found in the original SimCLR objective. See the supplementary material for additional details on model training.

After training the ResNet50 model on the selected dataset, it is then used as a feature extractor on "downstream" evaluation datasets. Images are resized so the smaller edge is 256 then we take a center crop of 224x224, which is then passed through the model. No other form of augmentation is used. Features are extracted from the last convolutional block of the ResNet50 model and have a dimension of 2048 unless the width of the network was modified to 4, in which case the dimension is 8192. A linear model is then trained on these features and the associated ground truth class labels. Details of the linear model are provided below. We use top-1 accuracy on the held out test set of the respective "downstream" dataset as the evaluation metric for the linear model. We compare different feature representations by measuring the relative change in accuracy when using supervised ImageNet features as the baseline (Δ ACC in Figure 3 and Figure 4). We chose supervised ImageNet features as the baseline because these features are readily accessible to nearly all practitioners, requiring zero additional training and very little computational resources. To facilitate reproducibility, all pretrained models are accessible from our GitHub project page.

5.2. Experiments on Fine-Grained Datasets

In this section we demonstrate the utility of iNat2021 as a pretraining dataset for existing fine-grained datasets. The extracted features are evaluated on Flowers102 [48], CUB [60], NABirds [57], StanfordDogs [32], and StanfordCars [37]. We also present results on CUBExpert, which is the standard CUB dataset but the class labels have been verified and cleaned by domain experts [57]. For these experiments, the linear model is a SVM trained using SGD for a maximum of 3k epochs with a stopping criteria tolerance of $1e{-}5$. For every experiment, we use 3-fold cross validation to determine the appropriate regularization constant $\alpha \in [1e{-}6, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 0.1, 1, 10]$.

We present the relative accuracy changes in relation to supervised ImageNet features for the various techniques in Figure 3. Please consult the supplementary material for specific accuracy values. Overall we find that supervised techniques produce the best features for all datasets except Stanford Cars, where the SwAV model trained on ImageNet produced the best features. The iNat2021 supervised model is the best performing on Flowers102, CUB, and CUBExpert; the iNat2018 supervised model is the best on NABirds, narrowly eclipsing the iNat2021 supervised model (0.806 vs. 0.804 top-1 accuracy); and the supervised ImageNet model is the best on StanfordDogs. When considering self-supervised methods, the SwAV model trained on ImageNet is consistently the top performer except for the Flowers102 dataset, where the SimCLR x4 model trained on iNat2021 mini achieves better performance (using a 4x larger feature

vector than the SwAV model).

In terms of pretraining datasets for self-supervised techniques, the ImageNet dataset appears better than the iNat2021 dataset: note the lines for self-supervised methods trained on iNat2021 and iNat2021 mini in Figure 3 are uniformly below their ImageNet counterparts for all datasets except Flowers102. While not particular surprising for the Stanford Dogs and Cars datasets that differ fundamentally from the iNaturalist domain, this is a surprising result for the bird datasets: CUB, CUBExpert, and NABirds. The ImageNet dataset has about 60 species of birds with ∼60k training images, while the iNat2021 dataset has 1,486 species with 414,847 and 74,300 training images in the large and mini splits respectively. Even with increased species and training samples, the ImageNet dataset out performs the iNat2021 dataset on downstream bird tasks. Perhaps this is an artifact of the *types* of images within these datasets as opposed to the *domain* of the datasets. The self-supervised techniques considered in this work were designed for ImageNet, therefore their default augmentation strategy appears to be designed for objects that take up a large fraction of the image size. Applying these strategies to datasets where objects do not necessarily take up large fraction of the image size (like iNat2021) appears to be inappropriate. See the supplementary material for an analysis of the sizes of bird bounding boxes across the datasets.

Note that supervised methods can still recover discriminative features from the iNat datasets (see the performance of supervised iNat2021 and iNat2021 mini in Figure 3), so it should be feasible for self-supervised methods to leverage these datasets to learn better representations. Interestingly, the effect of data size is not very apparent in Figure 3 for the experiments that use the large and mini variants of the iNat2021 dataset. While performance on the actual iNat2021 improved by 11 percentage points when switching from the mini to the large (see Table 2), we do not see a similar level of improvement for downstream tasks.

5.3. Experiments on NeWT

In this section we use the collection of binary tasks in NeWT as "downstream" classification tasks to investigate the effect of different pretraining methods. For these experiments the linear model is a SVM trained using liblinear for a maximum of 1k iterations with a stopping criteria tolerance of $1e{-}5$. For every experiment, we use 3-fold cross validation to determine the appropriate regularization constant C $\in [1e{-}4, 1e{-}3, 1e{-}2, 0.1, 1, 10, 1e2, 1e3]$.

The supervised ImageNet model achieved an average accuracy of 0.744 across all 164 NeWT tasks. The supervised iNat2021 model achieved the best average accuracy with a score of 0.806, followed by the supervised iNat2021 mini model at 0.793 and then the supervised iNat2018 model at 0.791. For self-supervised models, the SwAV model trained on ImageNet did the best at 0.733 average accuracy. We show the relative accuracy changes in relation to supervised ImageNet features for the various techniques in Figure 4, see the supplementary material for specific accuracy values.

For the *Appearance* based tasks in NeWT (which focus on a specific individual in the photo), we can see that there is a clear benefit to doing supervised pretraining on data from iNaturalist (using either iNat2018, iNat2021, or iNat2021 mini). *Species* classification, unsurprisingly, and *Age* have the biggest improvement followed by *Attribute* and then *Health*. We do not see the same benefit when using self-supervision for these *Appearance* based tasks. We instead find self-supervised models performing worse on average than ImageNet supervised features, even though they are trained on data from iNaturalist. Similarly, the *Behavior* tasks benefited from supervised pretraining on iNat datasets, but did not benefit from self-supervised pretraining. No method significantly improved performance on the *Context* tasks compared to supervised ImageNet features. All methods did relatively poorly on the two *Counting* tasks (0.59 baseline performance, note that chance is 50%). This could highlight the inappropriateness of using a classifier for detection style tasks, or it could highlight a particularly disappointing generalization behavior of these models. The SimCLR method trained on iNat2021 is a notable outlier in this experiment but the reason is unclear. Interestingly, all self-supervised models appear to provide a benefit over supervised ImageNet features and supervised iNat features for the *Gestalt* tasks, where the whole image needs to be analyzed as opposed to focusing on a particular subject.

Similar to the fine-grained datasets result, we see a reduced improvement between the iNat2021 large and mini datasets on the NeWT tasks as compared to evaluating on the iNat2021 test set. The SimCLR model achieved 0.678 mean accuracy using the iNat2021 mini split, and 0.689 with the full dataset. The supervised model went from 0.793 mean accuracy to 0.806. This result is surprising given the typical expectation of performance improvement when training with more data. Goyal *et al*. [20] perform experiments where they scale the amount of training data by a factor of 10, 50, and 100 and they see a larger performance gain for the ResNet50 model, albeit using Jigsaw [49] and Colorization [65] as pretext tasks, and Pascal VOC07 [16] as the downstream task. So either 5x more data is not a sufficient data increase, or self-supervision objectives like SimCLR behave differently.

While the experiments on existing fine-grained datasets in Figure 3 showed a benefit to using ImageNet over iNat2021 as the pretraining dataset for self-supervision, the NeWT results are much more mixed. For example SimCLR trained using ImageNet achieves better performance on average for the *Appearance - Age* tasks than SimCLR trained using iNat2021 (0.702 vs 0.688), but the results are flipped for the *Appearance - Species* tasks (0.647 vs 0.661).

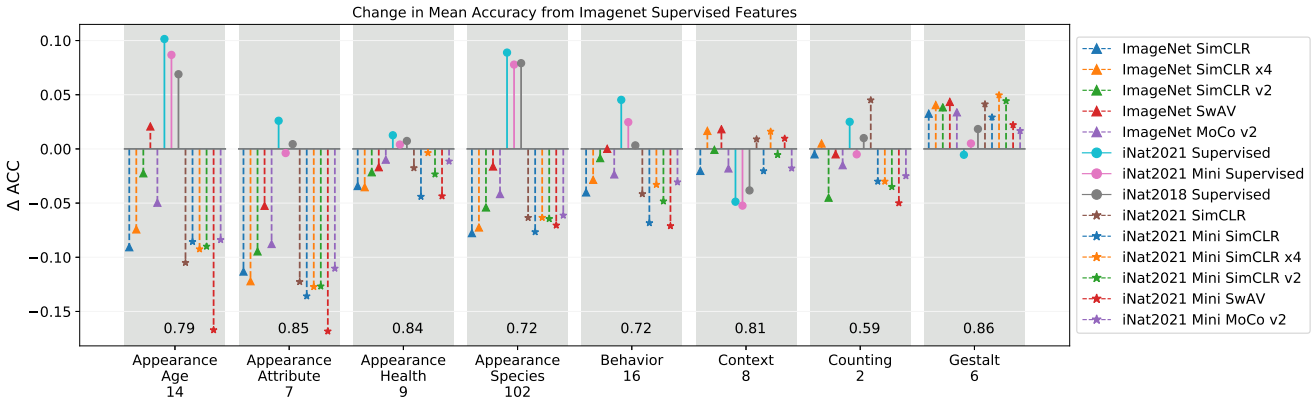

Figure 4. **NeWT evaluation.** The mean top-1 accuracy difference between "off-the-shelf" supervised ImageNet features and various other pretraining strategies on the NeWT dataset, divided into related groups. See Figure 3 for information regarding the plot organization and interpretation. Several patterns are apparent: (1) Supervised learning using iNaturalist data achieves better performance on NeWT tasks that focus on species appearance and behavior. (2) Self-supervised learning achieves better performance compared to supervised methods on the *Gestalt* tasks, i.e. tasks that do not focus on a particular individual. (3) For self-supervision, we do not see a consistent benefit to using iNat2021 over ImageNet (unlike Figure 3); sometimes pretraining on iNat2021 leads to better performance than pretraining on ImageNet, other times it is reversed. For detailed accuracy numbers see the supplementary material.

5.4. Discussion

We summarize our main findings:

Supervised ImageNet features are a strong baseline. The off-the-shelf supervised ImageNet features were often much better than the features derived from self-supervised models trained on either ImageNet or iNat2021. This applies to supervised iNat2021 features as well. It is currently easier to achieve downstream performance gains from a model trained with a supervised objective (assuming it is possible to get labels).

Fine-grained classification is challenging for self-supervised models. For most self-supervised methods performance is not close to supervised methods for the fine-grained datasets tested, see Figure 3. However, the SwAV method has closed the gap and is better in some cases (e.g. Stanford Cars). This trend did not hold when SwAV was trained on iNat2021 mini data.

Not all tasks are equal. Self-supervised features can be more effective compared to supervised ones for certain tasks (e.g. see the *Gestalt* tasks in NeWT in Figure 4). This highlights the value of benchmarking performance on a varied set of classification tasks, in addition to conventional object classification.

More data does not help methods as much for downstream tasks. While we observe a large boost in accuracy on the iNat2021 test set when we increase the amount of training data (+11 percentage points, see Tables 2 and 3), this boost is much smaller for both supervised and self-supervised models on the fine-grained datasets and NeWT (see the differences between iNat2021 large and mini for the supervised and SimCLR experiments in Figures 3 and 4).

Self-supervised ImageNet training settings do not necessarily generalize. The performance gap between supervised and self-supervised features on downstream tasks is closing when the feature extractor is trained on ImageNet. However, the gap between supervised and self-supervised features is much larger when the the feature extractor is trained on iNat2021. This potentially points to self-supervised training settings being overfit to ImageNet e.g. via hyperparameters or the image augmentations used.

6. Conclusion

We presented, and benchmarked, the iNat2021 and NeWT datasets. The iNat2021 dataset contains 2.7M training images covering 10k species. As a large-scale image dataset we have shown its utility as a powerful pretraining network for a variety of existing fine-grained datasets as well as the NeWT dataset. Our NeWT dataset expands beyond the question of "What species is this?", to incorporate questions that challenge models to identify behaviors, health, and context questions as they relate to wildlife captured in photographs. Our experiments on NeWT reveal interesting performance differences between supervised and self-supervised learning methods. While supervised learning appears to still have an edge over existing self-supervised approaches, new methods are constantly being introduced by the research community. The iNat2021 and NeWT datasets should serve as a valuable resource for benchmarking these new techniques as they expose challenges not present in the standard datasets currently in use.

Acknowledgments Thanks to the iNaturalist team and community for providing access to data, Eliot Miller and Mitch Barry for helping to curate NeWT, and to Pietro Perona for valuable feedback.

References

[1] iNaturalist. www.inaturalist.org, accessed Nov 14 2020.

[2] iNaturalist Challenge Datasets. https://github.com/visipedia/inat_comp, accessed Nov 14 2020.

[3] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, et al. Biodiversity loss and its impact on humanity. *Nature*, 2012.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020.

[11] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *ICCV Workshops*, 2019.

[12] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018.

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[15] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, 2021.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[17] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, 2017.

[18] Ciara Giaimo. Hold me, squeeze me, bite my head. *The New York Times*, Sep 2020.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[20] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.

[22] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[23] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[24] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[25] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[28] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *ICCV*, 2017.

[29] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[30] iNaturalist. Seek by iNaturalist, 2020. https://apps.apple.com/us/app/seek-by-inaturalist/id1353224144.

[31] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020.

[32] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization*, 2011.

[33] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019.

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, 2019.

[35] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper

Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[36] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, 2016.

[37] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[39] Neeraj Kumar, Peter N Belhumeur, Arijit Biswas, David W Jacobs, W John Kress, Ida C Lopez, and João VB Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*. 2012.

[40] 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 *ECCV*, 2014.

[41] Yen-Liang Lin, Vlad I Morariu, Winston Hsu, and Larry S Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*. 2014.

[42] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *ECCV*. 2012.

[43] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019.

[44] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013.

[45] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

[46] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *CVPR*, 2020.

[47] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[48] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

[49] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[50] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, 2015.

[51] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[53] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020.

[54] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 2009.

[55] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

[56] Cornell University. Merlin bird id, 2020. https://apps.apple.com/us/app/merlin-bird-id-by-cornell-lab/id773457673.

[57] 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 *CVPR*, 2015.

[58] 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 *CVPR*, 2018.

[59] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, 2014.

[60] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[61] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[62] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015.

[63] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014.

[64] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv:1910.04867*, 2019.

[65] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017.

[67] T. Züchner, C.J. Sharpe, and P.F.D. Boesman. Pink-throated brilliant (heliodoxa gularis). *Birds of the World*, 2020.