The iNaturalist Species Classification and Detection Dataset — Source link ↗

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui ...+5 more authors

Institutions: California Institute of Technology, Google, Cornell University

Related papers:

- Deep Residual Learning for Image Recognition

- ImageNet: A large-scale hierarchical image database

- Microsoft COCO: Common Objects in Context

- ImageNet Large Scale Visual Recognition Challenge

- The Caltech-UCSD Birds-200-2011 Dataset

Share this paper:

View more about this paper here: https://typeset.io/papers/the-inaturalist-species-classification-and-detection-dataset-1mtww4fvzs

The iNaturalist Species Classification and Detection Dataset

Grant Van Horn[1]   Oisin Mac Aodha[1]   Yang Song[2]   Yin Cui[3]   Chen Sun[2]
Alex Shepard[4]   Hartwig Adam[2]   Pietro Perona[1]   Serge Belongie[3]

[1]Caltech   [2]Google   [3]Cornell Tech   [4]iNaturalist

Abstract

Existing image classification datasets used in computer vision tend to have a uniform distribution of images across object categories. In contrast, the natural world is heavily imbalanced, as some species are more abundant and easier to photograph than others. To encourage further progress in challenging real world conditions we present the iNaturalist species classification and detection dataset, consisting of 859,000 images from over 5,000 different species of plants and animals. It features visually similar species, captured in a wide variety of situations, from all over the world. Images were collected with different camera types, have varying image quality, feature a large class imbalance, and have been verified by multiple citizen scientists. We discuss the collection of the dataset and present extensive baseline experiments using state-of-the-art computer vision classification and detection models. Results show that current non-ensemble based methods achieve only 67% top one classification accuracy, illustrating the difficulty of the dataset. Specifically, we observe poor results for classes with small numbers of training examples suggesting more attention is needed in low-shot learning.

1. Introduction

Performance on existing image classification benchmarks such as [32] is close to being saturated by the current generation of classification algorithms [9, 37, 35, 46]. However, the number of training images is crucial. If one reduces the number of training images per category, typically performance suffers. It may be tempting to try and acquire more training data for the classes with few images but this is often impractical, or even impossible, in many application domains. We argue that class imbalance is a property of the real world and computer vision models should be able to deal with it. Motivated by this problem, we introduce the iNaturalist Classification and Detection Dataset (iNat2017). Just like the real world, it exhibits a large class imbalance, as some species are much more likely to be observed.

Figure 1. Two visually similar species from the iNat2017 dataset. Through close inspection, we can see that the ladybug on the left has *two* spots while the one on the right has *seven*.

It is estimated that the natural world contains several million species with around 1.2 million of these having already been formally described [26]. For some species, it may only be possible to determine the species via genetics or by dissection. For the rest, visual identification in the wild, while possible, can be extremely challenging. This can be due to the sheer number of visually similar categories that an individual would be required to remember along with the challenging inter-class similarity; see Fig. 1. As a result, there is a critical need for robust and accurate automated tools to scale up biodiversity monitoring on a global scale [4].

The iNat2017 dataset is comprised of images and labels from the citizen science website iNaturalist[1]. The site allows naturalists to map and share photographic observations of biodiversity across the globe. Each observation consists of a date, location, images, and labels containing the name of the species present in the image. As of November 2017, iNaturalist has collected over 6.6 million observations from 127,000 species. From this, there are close to 12,000 species that have been observed by at least twenty

---

people and have had their species ID confirmed by multiple annotators.

The goal of iNat2017 is to push the state-of-the-art in image classification and detection for 'in the wild' data featuring large numbers of imbalanced, fine-grained, categories. iNat2017 contains over 5,000 species, with a combined training and validation set of 675,000 images, 183,000 test images, and over 560,000 manually created bounding boxes. It is free from one of the main selection biases that are encountered in many existing computer vision datasets - as opposed to being scraped from the web all images have been collected and then verified by multiple citizen scientists. It features many visually similar species, captured in a wide variety of situations, from all over the world. We outline how the dataset was collected and report extensive baseline performance for state-of-the-art classification and detection algorithms. Our results indicate that iNat2017 is challenging for current models due to its imbalanced nature and will serve as a good experimental platform for future advances in our field.

2. Related Datasets

In this section we review existing image classification datasets commonly used in computer vision. Our focus is on large scale, fine-grained, object categories as opposed to datasets that feature common everyday objects, *e.g.* [6, 5, 21]. Fine-grained classification problems typically exhibit two distinguishing differences from their coarse grained counter parts. First, there tends to be only a small number of domain experts that are capable of making the classifications. Second, as we move down the spectrum of granularity, the number of instances in each class becomes smaller. This motivates the need for automated systems that are capable of discriminating between large numbers of potentially visually similar categories with small numbers of training examples for some categories. In the extreme, face identification can be viewed as an instance of fine-grained classification and many existing benchmark datasets with long tail distributions exist *e.g.* [13, 28, 8, 3]. However, due to the underlying geometric similarity between faces, current state-of-the-art approaches for face identification tend to perform a large amount of face specific pre-processing [38, 33, 28].

The vision community has released many fine-grained datasets covering several domains such as birds [44, 42, 2, 40, 18], dogs [16, 29, 23], airplanes [24, 41], flowers [27], leaves [20], food [10], trees [43], and cars [19, 22, 48, 7]. ImageNet [32] is not typically advertised as a fine-grained dataset, yet contains several groups of fine-grained classes, including about 60 bird species and about 120 dog breeds. In Table 1 we summarize the statistics of some of the most common datasets. With the exception of a small number *e.g.* [18, 7], many of these datasets were typically constructed

Dataset Name	# Train	# Classes	Imbalance
Flowers 102 [27]	1,020	102	1.00
Aircraft [24]	3,334	100	1.03
Oxford Pets [29]	3,680	37	1.08
DogSnap [23]	4,776	133	2.85
CUB 200-2011 [42]	5,994	200	1.03
Stanford Cars [19]	8,144	196	2.83
Stanford Dogs [16]	12,000	120	1.00
Urban Trees [43]	14,572	18	7.51
NABirds [40]	23,929	555	15.00
LeafSnap* [20]	30,866	185	8.00
CompCars* [48]	136,727	1,716	10.15
VegFru* [10]	160,731	292	8.00
Census Cars [7]	512,765	2,675	10.00
ILSVRC2012 [32]	1,281,167	1,000	1.78
iNat2017	579,184	5,089	435.44

Table 1. Summary of popular general and fine-grained computer vision classification datasets. 'Imbalance' represents the number of images in the largest class divided by the number of images in the smallest. While susceptible to outliers, it gives an indication of the imbalance found in many common datasets. *Total number of train, validation, and test images.

to have an approximately uniform distribution of images across the different categories. In addition, many of these datasets were created by searching the internet with automated web crawlers and as a result can contain a large proportion of incorrect images *e.g.* [18]. Even manually vetted datasets such as ImageNet [32] have been reported to contain up to 4% error for some fine-grained categories [40]. While current deep models are robust to label noise at training time, it is still very important to have clean validation and test sets to be able to quantify performance [40, 31].

Unlike web scraped datasets [18, 17, 45, 10], the annotations in iNat2017 represent the consensus of informed enthusiasts. Images of natural species tend to be challenging as individuals from the same species can differ in appearance due to sex and age, and may also appear in different environments. Depending on the particular species, they can also be very challenging to photograph in the wild. In contrast, mass-produced, man-made object categories are typically identical up to nuisance factors, *i.e.* they only differ in terms of pose, lighting, color, but not necessarily in their underlying object shape or appearance [49, 7, 50].

3. Dataset Overview

In this section we describe the details of the dataset, including how we collected the image data (Section 3.1), how we constructed the train, validation and test splits (Section 3.2), how we vetted the test split (Section 3.2.1) and how we collected bounding boxes (Section 3.3). Future researchers may find our experience useful when constructing their own datasets.

	Super-Class	Class	Train	Val	BBoxes
	Plantae	2,101	158,407	38,206	-
	Insecta	1,021	100,479	18,076	125,679
	Aves	964	214,295	21,226	311,669
	Reptilia	289	35,201	5,680	42,351
	Mammalia	186	29,333	3,490	35,222
	Fungi	121	5,826	1,780	-
	Amphibia	115	15,318	2,385	18,281
	Mollusca	93	7,536	1,841	10,821
	Animalia	77	5,228	1,362	8,536
	Arachnida	56	4,873	1,086	5,826
	Actinopterygii	53	1,982	637	3,382
	Chromista	9	398	144	-
	Protozoa	4	308	73	-
	Total	5,089	579,184	95,986	561,767

Table 2. Number of images, classes, and bounding boxes in iNat2017 broken down by super-class. 'Animalia' is a catch-all category that contains species that do not fit in the other super-classes. Bounding boxes were collected for nine of the super-classes. In addition, the public and private test sets contain 90,427 and 92,280 images, respectively.
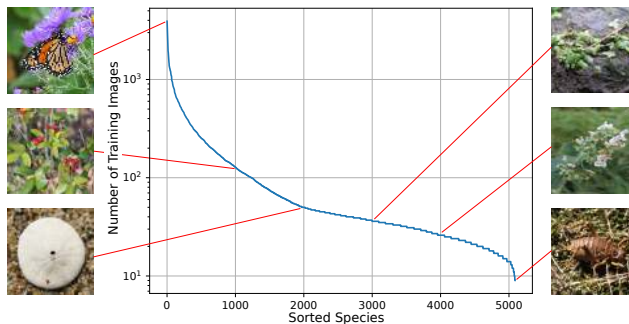

Figure 2. Distribution of training images per species. iNat2017 contains a large imbalance between classes, where the top 1% most populated classes contain over 16% of training images.

3.1. Dataset Collection

iNat2017 was collected in collaboration with iNaturalist, a citizen science effort that allows naturalists to map and share observations of biodiversity across the globe through a custom made web portal and mobile apps. Observations, submitted by *observers*, consist of images, descriptions, location and time data, and community identifications. If the community reaches a consensus on the taxa in the observation, then a "research-grade" label is applied to the observation. iNaturalist makes an archive of research-grade observation data available to the environmental science community via the Global Biodiversity Information Facility (GBIF) [39]. Only research-grade labels at genus, species or lower are included in this archive. These archives contain the necessary information to reconstruct which photographs belong to each observation, which observations belong to each observer, as well as the taxonomic hierarchy relating the taxa. These archives are refreshed on a rolling basis and the iNat2017 dataset was created by processing the archive from October 3rd, 2016.

3.2. Dataset Construction

The complete GBIF archive had 54k classes (genus level taxa and below), with 1.1M observations and a total of 1.6M images. However, over 19k of those classes contained only one observation. In order to construct train, validation and test splits that contained samples from all classes we chose to employ a taxa selection criteria: we required that a taxa have at least 20 observations, submitted from at least 20 unique observers (i.e. one observation from each of the 20 unique observers). This criteria limited the candidate set to 5,089 taxa coming from 13 super-classes, see Table 2.

The next step was to partition the images from these taxa into the train, validation, and test splits. For each of the selected taxa, we sorted the *observers* by their number of observations (fewest first) and selected the first 40% of observers to be in the test split, and the remaining 60% to be in the "train-val" split. By partitioning the observers in this way, and subsequently placing all of their photographs into one split or the other, we ensure that the behavior of a particular user (*e.g.* camera equipment, location, background, *etc.*) is contained within a single split, and not available as a useful source of information for classification on the other split for a specific taxa. Note that a particular observer may be put in the test split for one taxa, but the "train-val" split for another taxa. By first sorting the observers by their number of observations we ensure that the test split contains a high number of unique observers and therefore a high degree of variability. To be concrete, at this point, for a taxa that has exactly 20 unique observers (the minimum allowed), 8 observers would be placed in the the test split and the remaining 12 observers would be placed in the "train-val" split. Rather than release all test images, we randomly sampled ∼183,000 to be included in the final dataset. The remaining test images were held in reserve in case we encountered unforeseen problems with the dataset.

To construct the separate train and validation splits for each taxa from the "train-val" split we again partition on the observers. For each taxa, we sort the observers by increasing observation counts and repeatedly add observers to the validation split until either of the following conditions occurs: (1) The total number of *photographs* in the validation set exceeds 30, or (2) 33% of the available *photographs* in the "train-val" set for the taxa have been added to the validation set. The remaining observers and all of their photographs are added to the train split. To be concrete, and continuing the example from above, exactly 4 images would be placed in the validation split, and the remaining 8 images would be placed in the train split for a taxa with 20 unique
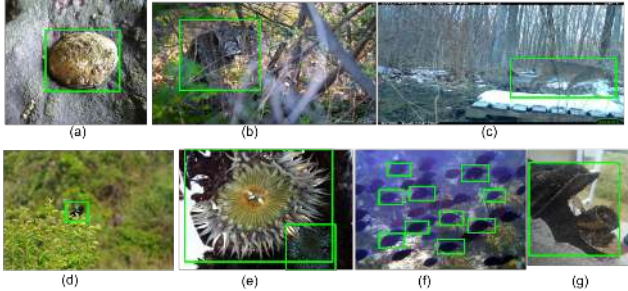
Figure 3. Sample bounding box annotations. Annotators were asked to annotate up to 10 instances of a super-class, as opposed to the fine-grained class, in each image.

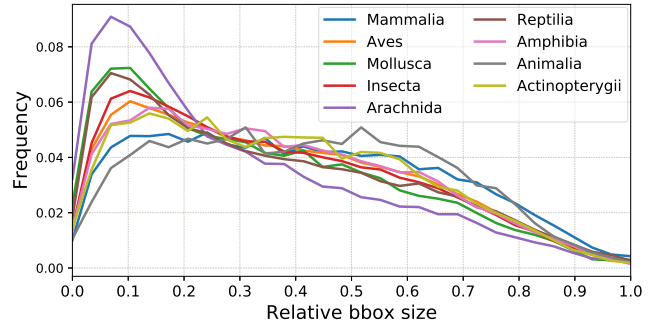Figure 4. The distribution of relative bounding box sizes (calculated by $\sqrt{w_{bbox} \times h_{bbox}}/\sqrt{w_{img} \times h_{img}}$) in the training set, per super-class. Most objects are relatively small or medium sized.

observers. This results in a validation split that has at least 4 and at most ∼30 images for each class (the last observer added to the validation split for a taxa may push the number of photographs above 30), and a train split that has at least 8 images for each class. See Fig. 2 for the distribution of train images per class.

At this point we have the final image splits, with a total of 579,184 training images, 95,986 validation images and 182,707 test images. All images were resized to have a max dimension of 800px. Sample images from the dataset can be viewed in Fig. 8. The iNat2017 dataset is available from our project website[2].

3.2.1 Test Set Verification

Each observation on iNaturalist is made up of one or more images that provide evidence that the taxa *was present*. Therefore, a small percentage of images may not contain the taxa of interest but instead can include footprints, feces, and habitat shots. Unfortunately, iNaturalist does not distinguish between these types of images in the GBIF export, so we crowdsourced the verification of three super-classes (Mammalia, Aves, and Reptilia) that might exhibit these "non-instance" images. We found that less than 1.1% of the test set images for Aves and Reptilia had non-instance images. The fraction was higher for Mammalia due to the prevalence of footprint and feces images, and we filtered these images out of the test set. The training and validation images were not filtered.

3.3. Bounding Box Annotation

Bounding boxes were collected on 9 out of the 13 super-classes (see Table 2), totaling 2,854 classes. Due to the in-herit difficultly of asking non-expert crowd annotators to both recognize and box specific fine-grained classes, we instructed annotators to instead box all instances of the associated super-class for a taxa (*e.g.* "Box all Birds" rather than

"Box all Red-winged Black Birds"). We collected super-class boxes only on taxa that are part of that super-class. For some super-classes (*e.g.* Mollusca), there are images containing taxa which are unfamiliar to many of the annotators (*e.g.* Fig. 3(a)). For those cases, we instructed the annotators to box the prominent objects in the images.

The task instructions specified to draw boxes tightly around all parts of the animal (including legs, horns, antennas, *etc.*). If the animal is occluded, the annotators were instructed to draw the box around the visible parts (*e.g.* Fig. 3(b)). In cases where the animal is blurry or small (*e.g.* Fig. 3(c) and (d)), the following rule-of-thumb was used: "if you are confident that it is an animal from the requested super-class, regardless of size, blurriness or occlusion, put a box around it." For images with multiple instances of the super-class, all of them are boxed, up to a limit of 10 (Fig. 3(f)), and bounding boxes may overlap (Fig. 3(e)). We observe that 12% of images have more than 1 instance and 1.3% have more than 5. If the instances are physically connected (*e.g.* the mussels in Fig. 3(g)), then only one box is placed around them.

Bounding boxes were not collected on the Plantae, Fungi, Protozoa or Chromista super-classes because these super-classes exhibit properties that make it difficult to box the individual instances (*e.g.* close up of trees, bushes, kelp, *etc.*). An alternate form of pixel annotations, potentially from a more specialized group of crowd workers, may be more appropriate for these classes.

Under the above guidelines, 561,767 bounding boxes were obtained from 449,313 images in the training and validation sets. Following the size conventions of COCO [21], the iNat2017 dataset is composed of 5.7% small instances (area $< 32^2$), 23.6% medium instances ($32^2 \leq$ area $\leq 96^2$) and 70.7% large instances (area $> 96^2$), with area computed as 50% of the annotated bounding box area (since segmentation masks were not collected). Fig. 4 shows the distribution of relative bounding box sizes, indicating that a majority of instances are relatively small and medium sized.

4. Experiments

In this section we compare the performance of state-of-the-art classification and detection models on iNat2017.

4.1. Classification Results

To characterize the classification difficulty of iNat2017, we ran experiments with several state-of-the-art deep network architectures, including ResNets [9], Inception V3 [37], Inception ResNet V2 [35] and MobileNet [11]. During training, random cropping with aspect ratio augmentation [36] was used. Training batches of size 32 were created by uniformly sampling from all available training images as opposed to sampling uniformly from the classes. We fine-tuned all networks from ImageNet pre-trained weights with a learning rate of 0.0045, decayed exponentially by 0.94 every 4 epochs, and RMSProp optimization with momentum and decay both set to 0.9. Training and testing were performed with an image size of 299×299, with a single centered crop at test time.

Table 3 summarizes the top-1 and top-5 accuracy of the models. From the Inception family, we see that the higher capacity Inception ResNet V2 outperforms the Inception V3 network. The addition of the Squeeze-and-Excitation (SE) blocks [12] further improves performance for both models by a small amount. ResNets performed worse on iNat2017 compared to the Inception architectures, likely due to over-fitting on categories with small number of training images. We found that adding a 0.5 probability dropout layer (drp) could improve the performance of ResNets. MobileNet, designed to efficiently run on embedded devices, had the lowest performance.

Overall, the Inception ResNetV2 SE was the best performing model. As a comparison, this model achieves a single crop top-1 and top-5 accuracy of 80.2% and 95.21% respectively on the ILSVRC 2012 [32] validation set [35], as opposed to 67.74% and 87.89% on iNat2017, highlighting the comparative difficulty of the iNat2017 dataset. A more detailed super-class level breakdown is available in Table 4 for the Inception ResNetV2 SE model. We can see that the Reptilia super-class (with 289 classes) was the most difficult with an average top-1 accuracy of 45.87%, while the Protozoa super-class (with 4 classes) had the highest accuracy at 89.19%. Viewed as a collection of fine-grained datasets (one for each super-class) we can see that the iNat2017 dataset exhibits highly variable classification difficulty.

In Fig. 5 we plot the top one public test set accuracy against the number of training images for each class from the Inception ResNet V2 SE model. We see that as the number of training images per class increases, so does the test accuracy. However, we still observe a large variance in accuracy for classes with a similar amount of training data, revealing opportunities for algorithmic improvements in both the low data and high data regimes.

	Validation		Public Test		Private Test	
	Top1	Top5	Top1	Top5	Top1	Top5
IncResNetV2 SE	**67.3**	**87.5**	**68.5**	**88.2**	67.7	**87.9**
IncResNetV2	67.1	**87.5**	68.3	88.0	**67.8**	87.8
IncV3 SE	65.0	85.9	66.3	86.7	65.2	86.3
IncV3	64.2	85.2	65.5	86.1	64.8	85.7
ResNet152 drp	62.6	84.5	64.2	85.5	63.1	85.1
ResNet101 drp	60.9	83.1	62.4	84.1	61.4	83.6
ResNet152	59.0	80.5	60.6	81.7	59.7	81.3
ResNet101	58.4	80.0	59.9	81.2	59.1	80.9
MobileNet V1	52.9	75.4	54.4	76.8	53.7	76.3

Table 3. Classification results for various CNNs trained on only the training set, using a single center crop at test time. Unlike some current datasets where performance is near saturation, iNat2017 still poses a challenge for state-of-the-art classifiers.

Super-Class	Avg Train	Public Test	
		Top1	Top5
Plantae	75.4	69.5	87.1
Insecta	98.4	77.1	93.4
Aves	222.3	67.3	88.0
Reptilia	121.8	45.9	80.9
Mammalia	157.7	61.4	85.1
Fungi	48.1	74.0	92.3
Amphibia	67.9	51.2	81.0
Mollusca	81.0	72.4	90.9
Animalia	67.9	73.8	91.1
Arachnida	87.0	71.5	88.8
Actinopterygii	37.4	70.8	86.3
Chromista	44.2	73.8	92.4
Protozoa	77.0	89.2	96.0

Table 4. Super-class level accuracy (computed by averaging across all species within each super-class) for the best performing model Inception ResNetV2 SE [12]. "Avg Train" indicates the average number of training images per class for each super-class. We observe a large difference in performance across the different super-classes.
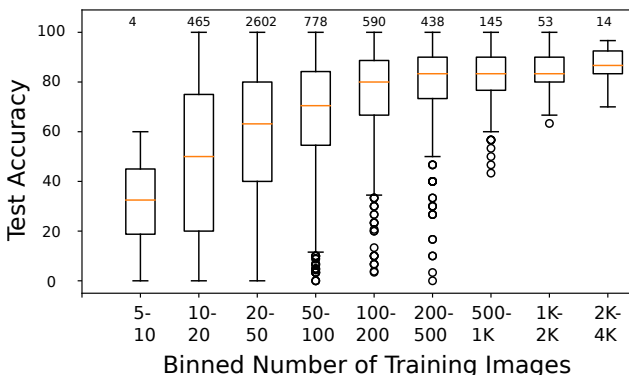

Figure 5. Top one public test set accuracy per class for IncResNet V2 SE [12]. Each box plot represents classes grouped by the number of training images. The number of classes for each bin is written on top of each box plot. Performance improves with the number of training images, but the challenge is how to maintain high accuracy with fewer images?

4.2. Detection Results

To characterize the detection difficulty of iNat2017, we adopt Faster-RCNN [30] for its state-of-the-art performance as an object detection setup (which jointly predicts object bounding boxes along with class labels). We use a TensorFlow [1] implementation of Faster-RCNN with default hyper-parameters [14]. Each model is trained with 0.9 momentum, and asynchronously optimized on 9 GPUs to expedite experiments. We use an Inception V3 network, initialized from ImageNet, as the backbone for our Faster-RCNN models. Finally, each input image is resized to have 600 pixels as the short edge while maintaining the aspect ratio.

As discussed in Section 3.3, we collected bounding boxes on 9 of the 13 super-classes, translating to a total of 2,854 classes with bounding boxes. In the following experiments we only consider performance on this subset of classes. Additionally, we report performance on the the validation set in place of the test set and we only evaluate on images that contained a single instance. Images that contained only evidence of the species' presence and images that contained multiple instances were excluded. We evaluate the models using the detection metrics from COCO [21].

We first study the performance of fine-grained localization and classification by training the Faster-RCNN model on the 2,854 class subset. Fig. 7 shows some sample detection results. Table 5 provides the break down in performance for each super-class, where super-class performance is computed by taking an average across all classes within the super-class. The precision-recall curves (again at the super-class level) for 0.5 IoU are displayed in Fig. 6. Across all super-classes we achieve a comprehensive average precision (AP) of 43.5. Again the Reptilia super-class proved to be the most difficult, with an AP of 21.3 and an AUC of 0.315. At the other end of the spectrum we achieved an AP of 49.4 for Insecta and an AUC of 0.677. Similar to the classification results, when viewed as a a collection of datasets (one for each super-class) we see that iNat2017 exhibits highly variable detection difficulty, posing a challenge to researchers to build improved detectors that work across a broad group of fine-grained classes.

Next we explored the effect of label granularity on detection performance. We trained two more Faster-RCNN models, one trained to detect super classes rather fine-grained classes (so 9 classes in total) and another model trained with all labels pooled together, resulting in a generic object / not object detector. Table 6 shows the resulting AP scores for the three models when evaluated at different granularities. When evaluated on the coarser granularity, detectors trained on finer-grained categories have lower detection performance when compared with detectors trained at coarser labels. The performance of the 2,854-class detector is particularly poor on super-class recognition and object localization. This suggests that the Faster-RCNN algorithm
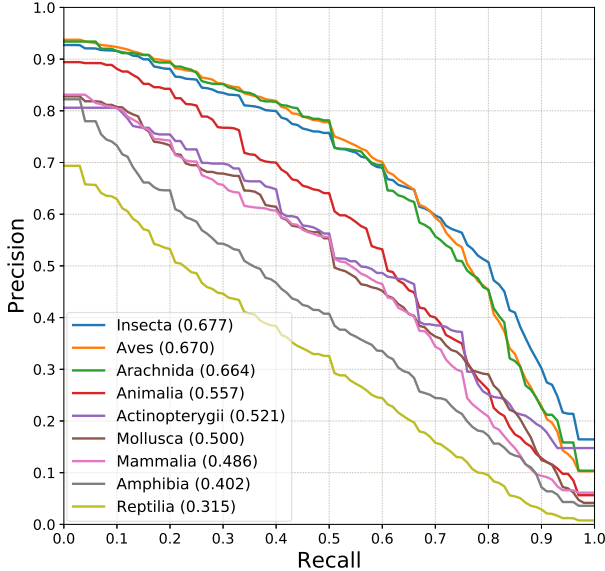

Figure 6. Precision-Recall curve with 0.5 IoU for each super-class, where the Area-Under-Curve (AUC) corresponds to AP^{50} in Table 5. Super-class performance is calculated by averaging across all fine-grained classes. We can see that building a detector that works well for all super-classes in iNat2017 will be a challenge.

	AP	AP^{50}	AP^{75}	AR^1	AR^{10}
Insecta	49.4	**67.7**	**59.3**	**64.5**	**64.9**
Aves	**49.5**	67.0	59.1	63.3	63.6
Reptilia	21.3	31.5	24.9	44.0	44.8
Mammalia	33.3	48.6	39.1	49.8	50.6
Amphibia	28.7	40.2	35.0	52.0	52.3
Mollusca	34.8	50.0	41.6	52.0	53.0
Animalia	35.6	55.7	40.8	48.3	50.5
Arachnida	43.9	66.4	49.6	57.3	58.6
Actinopterygii	35.0	52.1	41.6	49.1	49.6
Overall	43.5	60.2	51.8	59.3	59.8

Table 5. Super-class-level Average Precision (AP) and Average Recall (AR) for object detection, where AP, AP^{50} and AP^{75} denotes AP@[IoU=.50:.05:.95], AP@[IoU=.50] and AP@[IoU=.75] respectively; AR^1 and AR^{10} denotes AR given 1 detection and 10 detections per image.

Training	**Evaluation**		
	2854-class	9-super-class	1-generic
2854-class	43.5	55.6	63.7
9-super-class	-	65.8	76.7
1-generic	-	-	78.5

Table 6. Detection performance (AP@[IoU=.50:.05:.95]) with different training and evaluation class granularity. Using finer-grained class labels during training has a negative impact on coarser-grained super-class detection. This presents an opportunity for new detection algorithms that maintain precision at the fine-grained level.

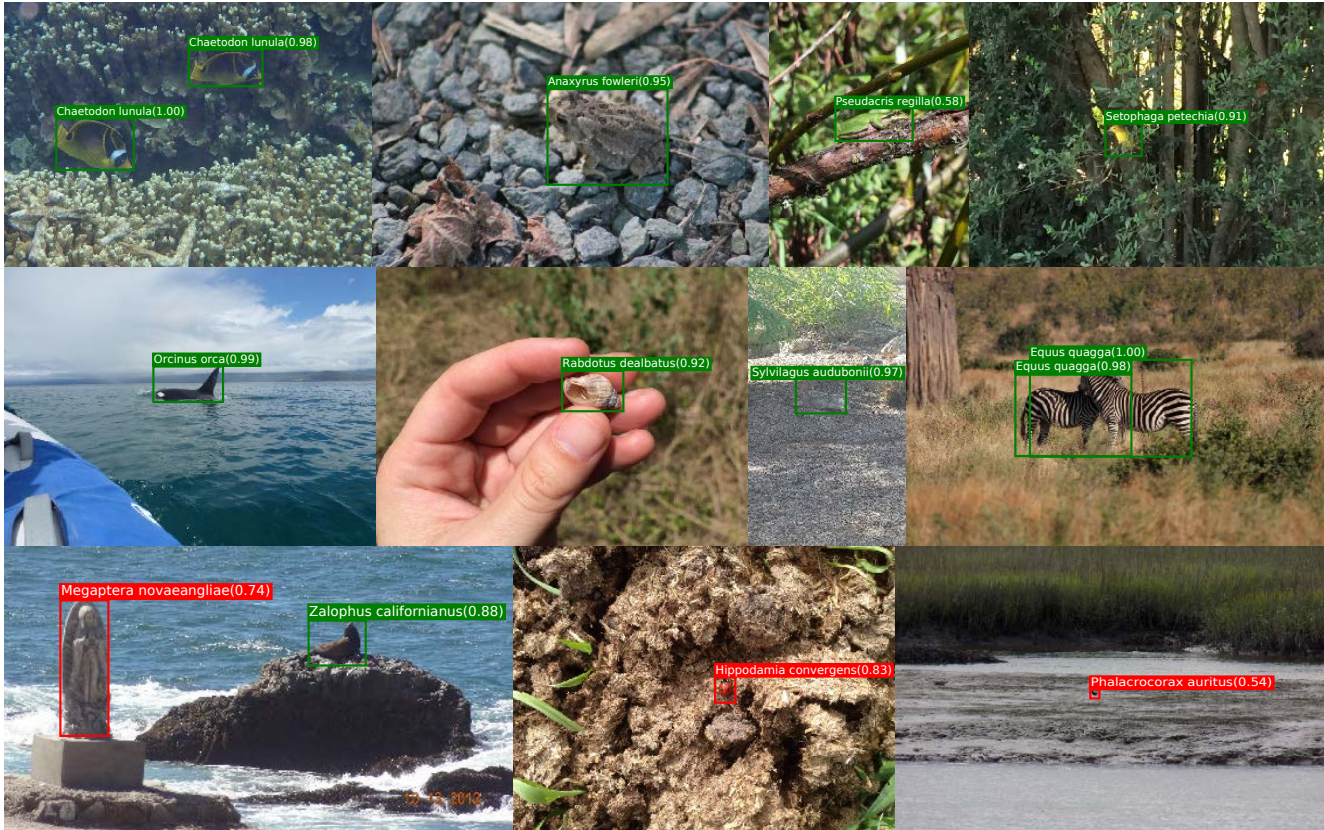has plenty of room for improvements on end-to-end fine-grained detection tasks.

Figure 7. Sample detection results for the 2,854-class model that was evaluated across all validation images. Green boxes represent correct species level detections, while reds are mistakes. The bottom row depicts some failure cases. We see that small objects pose a challenge for classification, even when localized well.

5. Conclusions and Future Work

We present the iNat2017 dataset, in contrast to many existing computer vision datasets it is: 1) unbiased, in that it was collected by non-computer vision researchers for a well defined purpose, 2) more representative of real-world challenges than previous datasets, 3) represents a long-tail classification problem, and 4) is useful in conservation and field biology. The introduction of iNat2017 enables us to study two important questions in a real world setting: 1) do long-tailed datasets present intrinsic challenges? and 2) do our computer vision systems exhibit transfer learning from the well-represented categories to the least represented ones? While our baseline classification and detection results are encouraging, from our experiments we see that state-of-the-art computer vision models have room to improve when applied to large imbalanced datasets. Small efficient models designed for mobile applications and embedded devices have even more room for improvement [11].

Unlike traditional, researcher-collected datasets, the iNat2017 dataset has the opportunity to grow with the iNaturalist community. Currently, every 1.7 hours another species passes the 20 unique observer threshold, making it available for inclusion in the dataset (already up to 12k as of November 2017, up from 5k when we started work on the dataset). Thus, the current challenges of the dataset (long tail with sparse data) will only become more relevant.

In the future we plan to investigate additional annotations such as sex and life stage attributes, habitat tags, and pixel level labels for the four super-classes that were challenging to annotate. We also plan to explore the "open-world problem" where the test set contains classes that were never seen during training. This direction would encourage new error measures that incorporate taxonomic rank [25, 47]. Finally, we expect this dataset to be useful in studying how to teach fine-grained visual categories to humans [34, 15], and plan to experiment with models of human learning.
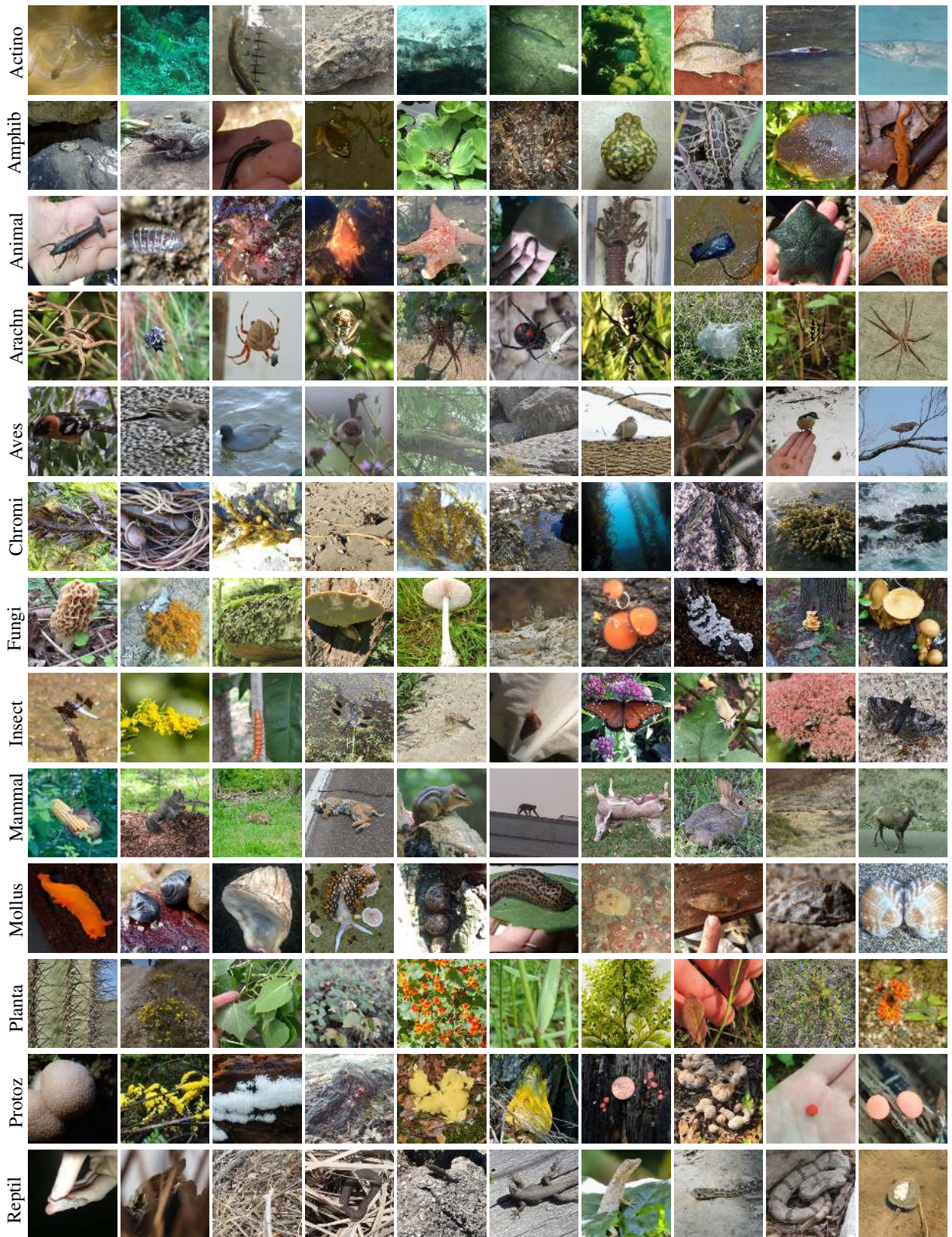
Figure 8. Example images from the training set. Each row displays randomly selected images from each of the 13 different super-classes. For ease of visualization we show the center crop of each image.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 6

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014. 2

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. 2

[4] B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, et al. Biodiversity loss and its impact on humanity. *Nature*, 2012. 1

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 2

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *CVIU*, 2007. 2

[7] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, 2017. 2

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[10] S. Hou, Y. Feng, and Z. Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *ICCV*, 2017. 2

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5, 7

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017. 5

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 2

[14] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 6

[15] E. Johns, O. Mac Aodha, and G. J. Brostow. Becoming the expert-interactive multi-class machine teaching. In *CVPR*, 2015. 7

[16] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *FGVC Workshop at CVPR*, 2011. 2

[17] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, et al. Openimages: A public dataset for large-scale multi-label and multiclass image classification. *Dataset available from https://github. com/openimages*, 2016. 2

[18] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, 2016. 2

[19] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *3D Representation and Recognition Workshop at ICCV*, 2013. 2

[20] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*. 2012. 2

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 4, 6

[22] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*. 2014. 2

[23] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. In *ECCV*. 2012. 2

[24] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[25] A. Mittal, M. Blaschko, A. Zisserman, and P. Torr. Taxonomic multi-class prediction and person layout using efficient structured ranking. In *ECCV*, 2012. 7

[26] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm. How many species are there on earth and in the ocean? *PLoS Biol*, 2011. 1

[27] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006. 2

[28] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, 2015. 2

[29] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 2

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *PAMI*, 2017. 6

[31] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 2

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1, 2, 5

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[34] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014. 7

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 1, 5

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 5

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1, 5

[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[39] K. Ueda. iNaturalist Research-grade Observations. iNaturalist.org. Occurrence Dataset. *https://doi.org/10.15468/ab3s5x*, 2017. 3

[40] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015. 2

[41] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. Blaschko, D. Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, 2014. 2

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[43] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona. Cataloging public objects using aerial and street-level images-urban trees. In *CVPR*, 2016. 2

[44] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010. 2

[45] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *ICCV*, 2017. 2

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CVPR*, 2017. 1

[47] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV*, 2015. 7

[48] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015. 2

[49] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 2

[50] X. Zhang, Y. Cui, Y. Song, H. Adam, and S. Belongie. The iMaterialist Challenge 2017 Dataset. *FGVC Workshop at CVPR*, 2017. 2