2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00088

Elusive Images: Beyond Coarse Analysis for Fine-Grained Recognition

Connor Anderson[1] Matt Gwilliam[2] Evelyn Gaskin[1] Ryan Farrell[1]
[1]Brigham Young University, Provo, UT [2]University of Maryland, College Park, MD

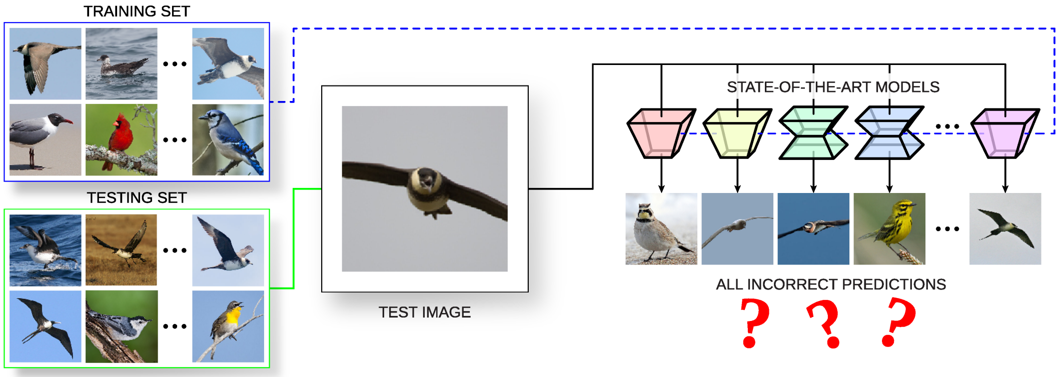

Figure 1. **Fine-grained analysis of *elusive* images.** Despite the cutting-edge performance of state-of-the-art models, there are images that are universally misclassified. Depicted above is a real example from the CUB-200-2011 dataset [57]: this image of a Pomarine Jaeger is misclassified by all state-of-the-art-models (the images on the right show the respective classes that are incorrectly predicted). There are various reasons why such images evade even our best classification models. This work seeks to identify the challenges behind these **elusive images** and focus future research on them.

Abstract

While the community has seen many advances in recent years to address the challenging problem of Fine-grained Visual Categorization (FGVC), progress seems to be slowing—new state-of-the-art methods often distinguish themselves by improving top-1 accuracy by mere tenths of a percent. However, across all of the now-standard FGVC datasets, there remain sizeable portions of the test data that none *of the current state-of-the-art (SOTA) models can successfully predict. This paper provides a framework for identifying and studying the errors that current methods make across diverse fine-grained datasets. Three models of difficulty—**Prediction Overlap**, **Prediction Rank** and **Pairwise Class Confusion**—are employed to highlight the most challenging sets of images and classes. Extensive experiments apply a range of standard and SOTA methods, evaluating them on multiple FGVC domains and datasets. Insights acquired from coupling these difficulty paradigms with the careful analysis of experimental results suggest crucial areas for future FGVC research, focusing critically on the set of **elusive images** that none of the current models can correctly classify. Code is available at* `catalys1.github.io/elusive-images-fgvc`.

1. Introduction

Fine-grained visual categorization (FGVC), with its high intra-class variation (*e.g.* different poses) and low inter-class variation (*e.g.* all classes have the same parts), is a uniquely challenging task within the broader area of image classification and object recognition. While the deep learning approaches developed recently have improved benchmark performance considerably, they still have sizeable shortcomings. Even the best methods mislabel images that a human expert would not; and when applied to real-world data, they leave much to be desired. Existing analysis, however, leaves such failures unexplored, focusing instead on top-level metrics (*e.g.* top-1 accuracy) for various datasets and tasks [27].

Ironically, the fine-grained recognition community tends to consider errors at a coarse (entire dataset) level. This lack of granularity and careful analysis limits progress; the minute improvement in top-1 accuracy of a new state-of-the-art (SOTA) method compared to its predecessors has very little explanatory power. As a result, we receive little insight into the actual progress being made in the field, in terms of which core FGVC problems are now solved that were unsolved before.

Identifying these unsolved problems starts with identifying the images that are misclassified and studying their

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00088

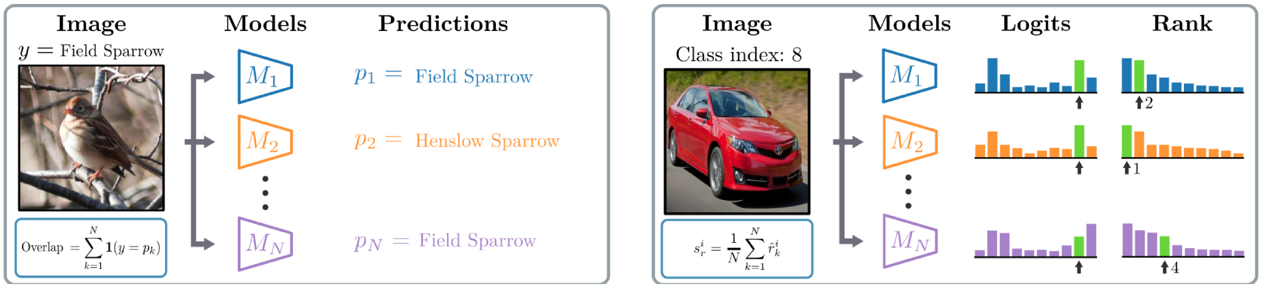

Figure 2. **Calculating prediction overlap and prediction rank**. (**left**): prediction overlap is the number of predictions from N different models that match the true class. (**right**): prediction rank is the average normalized rank (across N models) of the ground-truth class in the predicted distribution of the classifier.

characteristics. A given SOTA method can give inconsistent predictions across differing initialization seeds on both a per-image and per-class basis [22]. This can be leveraged to assess individual image difficulty in terms of whether one or more methods always predict correctly, compared to those where one or more methods sometimes or even never predict the correct class. We can consider other properties of the method predictions to help identify which images are hard, and, by aggregating information at the class level, the hard classes as well.

Once hard images have been identified, we can carefully analyze them to find the associated hard FGVC problems. Fig. 1 provides a great example of one of the key types of mistakes that these networks make. The struggle that classifiers have with this image serves as a reminder that their reliance on training data is both a gift and a curse. While increasing the amount of good data is a reliable way to increase performance, the under-representation of a given pose (such as a bird in flight) or other characteristic (cropping, occlusion, object scale, background, *etc*.) in the training data often guarantees a model will err when that characteristic occurs in the test set.

The fact that many such issues exist is well-documented and understood within the community. Our purpose in this work is not to simply restate these problems. Rather, we attempt to identify challenges at increasing levels of detail—moving from the dataset to its classes and even to individual images; to quantify them; and to provide analytic tools for understanding them. By doing this, we hope to provide tools and insight for future progress in the field.

In this paper, we seek to study what SOTA models have yet to solve: the **elusive** set of misclassified images. Fine-grained analysis, beyond the standard measure of overall accuracy, is a practice we advocate for and believe to be *critical* for further advancement of the field. We consider errors at a more granular level, performing rigorous analysis on the class- and image-level mistakes of state-of-the-art FGVC methods. In our effort to directly confront these long-standing issues, this work makes the following contributions:

- We introduce a novel framework for carefully analyzing challenging images, with three distinct views on difficulty (see Sec. 2). We will release this framework as a toolkit which can be used to analyze other models and datatsets.
- We standardize reimplementations of several state-of-the-art FGVC methods in a publicly available repository, allowing for fast, fair comparison and benchmarking.
- We present a novel real-world birds dataset, iCub (images of CUB categories gathered from iNaturalist.org) which is critical to our analysis, and which we release for further research.
- We employ our framework to analyze an extensive set of experiments (6 methods across 6 datasets), identifying the classes and images that SOTA methods struggle with.

2. Models of difficulty

To date, virtually every paper published in the FGVC literature—and in image classification, generally—uses a single performance measure for comparison on a given dataset. Typically, this measure is average accuracy (the fraction of all images correctly predicted) or class average accuracy (the per-class accuracy averaged across all classes). These measures are not only *coarse*, but limiting, providing only a minimal degree of objective comparison.

More *fine-grained* analysis affords valuable insight into the tradeoffs between approaches. Perhaps one model or classifier struggles on particular classes but another complementary model excels at those same classes. Moreover, detailed analysis about which images and classes are being classified correctly or incorrectly is rarely conducted; yet it can be important and illuminating.

In this section, we present and describe three approaches for assessing the difficulty of dataset images and classes.

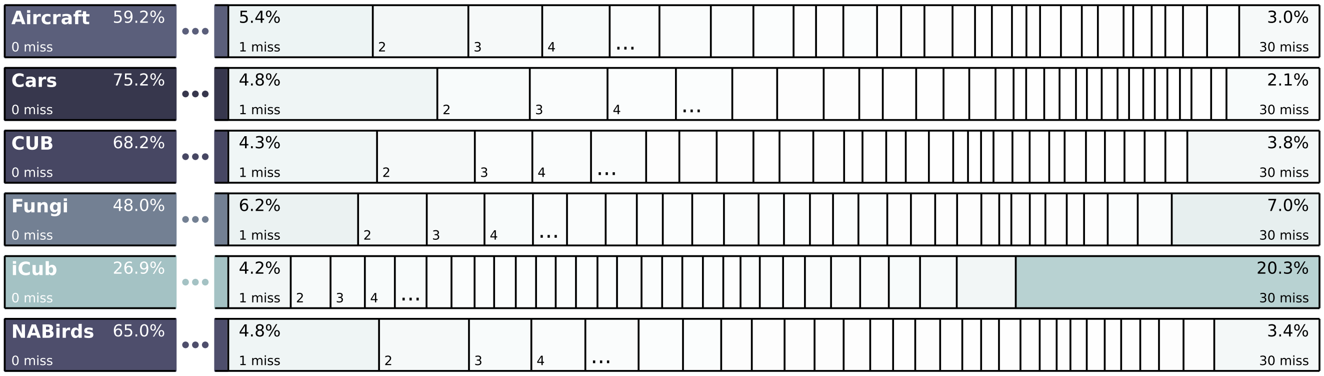

Figure 3. **Prediction overlap**. We show the overlap in predictions across all models. Each bin $k = [0, \ldots, N]$ gives the percentage of images from the dataset that were missed (incorrectly classified) by k out of N trained models (six models and five trials each).

The first method, **Prediction Overlap**, empirically estimates image and class difficulty based on a set of strong classifiers and how they collectively perform; the second approach, **Prediction Rank**, determines difficulty by examining the rank of an image's label within a classifier's prediction vector for the image; the final paradigm, **Pairwise Class Confusion**, is focused on pairs of classes, seeking to measure how distinct or confusing two classes are.

2.1. Prediction overlap

We propose **prediction overlap** as one method for quantifying the difficulty of an image. The idea is to take the predictions of N independently trained models and look at how many are correct. This process reveals a continuum of image difficulty within the dataset, where images with many correct predictions are inherently easier to classify, while images with few correct predictions are harder. This definition of difficulty depends on the type and number of models used, but as we show in Fig. 3 and explain in Sec. 4, there is a clear pattern of easy and hard subsets in the data that emerges across datasets, despite using many strong models.

Let us be precise in our definition of prediction overlap. We start with a dataset D, divided into train and evaluation sets D_T and D_E. D consists of images and their corresponding class labels $\{(x_i, y_i)\}$. We train N different models to predict y_i from x_i using D_T. Once the models are trained, we get N class predictions—one from each model—for each image x_i in the evaluation set D_E: $(p_1^i, \ldots, p_N^i)$. We then compute an overlap value o_i for each image as the number of correct predictions:

$$o_i = \sum_{k=1}^{N} \mathbf{1}(p_k^i, y_i) \tag{1}$$

where $\mathbf{1}(a, b) = 1$ if $a = b$ else 0. We use the overlap value o_i, or equivalently the overlap ratio o_i/N, as a measure of image difficulty. Fig. 2 (left) shows how this works.

In Sec. 4 we show prediction overlap results for multiple datasets and models. We find that within each dataset there are generally a large number of **easy** images, which are predicted correctly by all models. There is also a smaller set of **elusive** images, which are *never* correctly classified. In between the two extremes are images varying from **medium** to **hard**, which are sometimes classified correctly, with varying levels of frequency.

2.2. Prediction rank

We propose **prediction rank** as an additional, complimentary method for quantifying the difficulty of an image. Similar to prediction overlap, we start with the output of N independently trained models, but we consider the predictive distribution over classes (the logits) instead of just the top prediction. We order the logits $\mathbf{q}_i$ for each image x_i from highest to lowest, and calculate the rank of the ground truth class (its position in the ordered list):

$$r^i = \text{rank}(y_i, \text{argsort}(\mathbf{q}_i)) \tag{2}$$

where `argsort` returns the indices that sort the sequence and $\text{rank}(a, \mathbf{b})$ returns the index of a in $\mathbf{b}$. Using N models we obtain a vector of ranks for each image, $(r_1^i, \ldots, r_N^i)$, and we normalize each value by the maximum observed rank across all models and images: $\hat{r}_k^i = \frac{r_k^i}{\max_j \max_n r_n^j}$. We define the prediction rank score for image x_i as its average normalized rank

$$s_r^i = \frac{1}{N} \sum_{k=1}^{N} \hat{r}_k^i \tag{3}$$

The score s_r^i gives a continuous value of hardness for each image. The process is depicted in Fig. 2 (right). We also compute a class-level difficulty score by averaging the prediction rank scores over all images in a class.

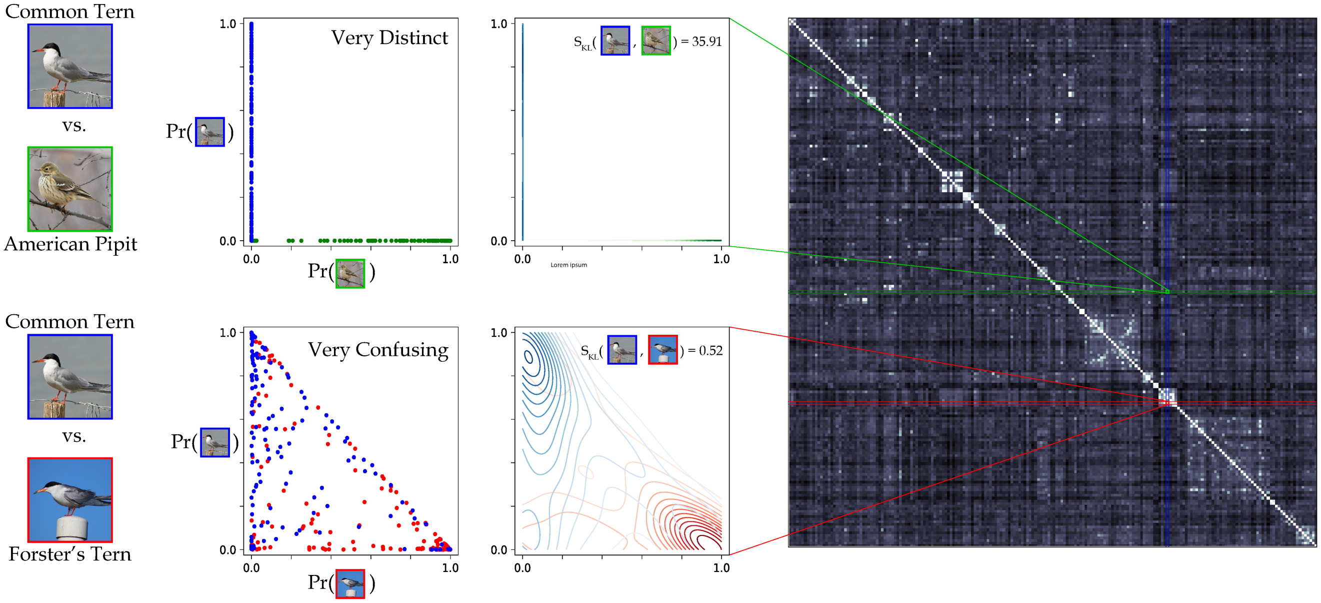

Figure 4. **Pairwise class confusion**. The top and bottom rows respectively consider very distinct and very confusing pairs of classes. In each row, prediction models estimate class label probabilities (each point is a model's prediction on one image of label probability for each of the two classes). These predictions have little to no overlap (distinct case) or have highly overlapping distributions (confusing case). A symmetric KL divergence measure (denoted S_{KL}) quantifies the degree to which models find a pair of classes confusing. At the right, pairwise S_{KL} values for the full CUB dataset (all 200 classes) are shown; higher brightness indicates similarity and several block diagonal regions show similar classes that are easily confused (often within a taxonomic family).

2.3. Pairwise class confusion

Our third approach for estimating difficulty uses **pairwise class confusion** to identify similar classes. For a pair of classes, $1 \leq i, j \leq |C|$, we construct two image sets I_i and I_j containing the images with ground-truth labels of i and j, respectively. For an image x_k, with ground-truth class label y_k, a model M produces a prediction vector $P_M(x_k) = \left(p_1^k, p_2^k, p_3^k, \ldots, p_{|C|}^k\right)$, $|P_M| = 1$. For each x_k, we care about only two of the values in the prediction vector, namely p_i^k and p_j^k. These represent model M's calculations of the probability that image x_k belongs to class i and class j, respectively. These prediction values are computed and extracted, effectively mapping an image x_k to a probability pair (p_i^k, p_j^k).

The images in I_i are mapped in this way to probability pairs, forming a non-parametric distribution $Pr(y_k = i)$ vs $Pr(y_k = j)$ over the unit square $[0, 1]_2$. In Fig. 4, a pair of distinct classes are shown in the top row, and a pair of similar classes are shown in the bottom row. The probability pairs are plotted in the top-left and bottom-left. The top-right and bottom-right plots show these same sets of points as distributions (via kernel density estimation). Note that for the distinct classes, these distributions are entirely disjoint (top-right plot). For the similar classes, there is significant overlap between the blue and red point sets, indicating that the model struggles to correctly classify the instances from classes i and j (actually misclassifying many instances). This confusion is again observable in the overlapping distributions (lower-right plot).

To quantify how similar (confusing) or dissimilar (distinct) a pair of classes is, the Kullback-Leibler divergence is used between their respective distributions to measure similarity. The KL divergence D_{KL} is defined for two distributions $P(x)$ and $Q(x)$ as

$$D_{KL}(P, Q) = \sum_{x \in \mathcal{X}} P(x) \cdot log\left(\frac{P(x)}{Q(x)}\right) \tag{4}$$

KL divergence is inherently asymmetric, $D_{KL}(P, Q) \neq D_{KL}(Q, P)$. Therefore, a symmetric KL divergence S_{KL} is defined to calculate similarity between two classes' distributions:

$$S_{KL}(P, Q) = \tfrac{1}{2}\left(D_{KL}(P, Q) + D_{KL}(Q, P)\right) \tag{5}$$

For every pair of classes i and j, these distributions Pr_i and Pr_j are generated and the symmetric KL divergence $S_{KL}\left(Pr_i, Pr_j\right)$ is calculated and used as a measure of similarity. A pair of classes with distinct distributions has a very high divergence, while a pair of similar classes will have a low divergence.

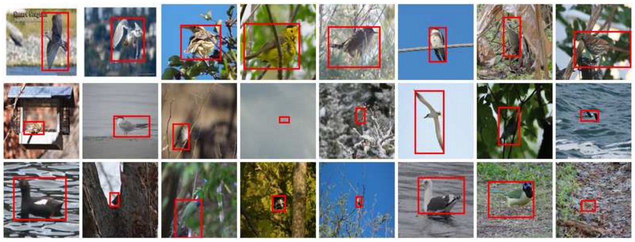

Figure 5. **iCub example images**. Zoom in for detail.

3. Hard problems in FGVC

There are multiple factors that can make certain images or classes difficult to classify. We give a brief overview of some of the prominent ones here. Many are hard to quantify, and we don't attempt to measure them all in this work.

Similar class confusion Within FGVC datasets, there exist groups of classes that are especially hard to distinguish from one another due to visual similarity within the domain; in fact, this is the core challenge of FGVC. When a model's error involves predicting a particularly similar-looking class, we consider this to be an instance of similar class confusion. We compute how many similar classes each dataset has in Sec. 4.

Object size and location We hypothesize that the size and location of the object in relation to the image may account for some challenging images, where classifiers may struggle more for objects that are very small (i.e. distant from the camera), or not well-centered. We use bounding box area and offset to approximate object size and location.

Pose Data driven classification methods naturally struggle when they encounter out of distribution poses during inference. However, assessing the actual prevalence and impact of this issue is very challenging. It would require both annotations of poses as well as algorithms that group such poses to determine which are in or out of distribution. Due to the unavailability of the keypoint labels and the challenges associated with clustering poses, we propose such analysis for future work.

Occlusion While we recognize the importance of occlusion as a source of error, we do not address it in this work for two reasons: first, it is difficult to precisely define occlusion in a satisfying way; and second, previous definitions (*e.g.* from Hoeim *et al.* [25]) rely entirely on manual annotation.

Distractors The background or other objects in the image have the potential to confuse a classifier; this issue is also hard to measure, and we don't address it here.

Photometric Issues such as lack of contrast, poor illumination and blurriness can make images challenging to classify, but we do not address those here.

4. Experiments and analysis

4.1. Datasets

We use four prominent and well-studied fine-grained classification datasets—FGVC Aircraft [41], Stanford Cars [33], Caltech Birds (CUB) [57] and NABirds [55]—along with the recently introduced Danish Fungi Mini dataset [44]. Statistics of these datasets are shown in Table 1 in the Supplementary Material. We also introduce a new birds dataset specifically for use in evaluation, which we call **iCub**. iCub has the same categories as CUB, but more images and additional challenges. We describe iCub in more detail next.

4.1.1 The iCub dataset

We introduce iCub, a new *evaluation-only* bird dataset that consists of the same 200 bird species as CUB, but with images sourced from iNaturalist.[1] The dataset is intended to be used as an additional validation set for models trained on CUB; thus, none of the images are designated for training. We provide iCub as an aid in analyzing model errors.

iCub was built by downloading all research-grade iNaturalist images for CUB categories, available as of January 2020. Images are certified as "research grade" when the label has been verified by consensus among the iNaturalist citizen-scientist community. To keep the dataset size manageable, we randomly selected 100 images per category and filtered them by hand, removing images with large flocks of birds or multiple birds of different species, then annotated the remaining images with bounding boxes. In total, iCub contains 16,876 images. The last row in Table 1 shows details for iCub and Fig. 5 shows randomly chosen iCub images.

Our purpose in introducing iCub is to provide an additional, large source of evaluation data that can test a CUB-trained model's performance on the same classes but on a different, more challenging visual distribution. iCub contains a greater number of difficult and *elusive* instances than does CUB for the same categories. iCub also contains many more images where the bird is small relative to the area of the image. The distributions of object size, aspect ratio, and centeredness (distance from the image center to the bounding box center), estimated from the bounding boxes, are shown in Fig. 12 (Supplementary) for CUB and iCub. The distributions of object size in particular are quite different.

4.2. Models

We use a small representative set of FGVC-specific methods—PMG [12], WSDAN [28], SIMTrans [51] and IELT [64]—as well as ResNet-50 [24] and ViT [11], which serve as the backbone networks for the other approaches.

[1] www.inaturalist.org

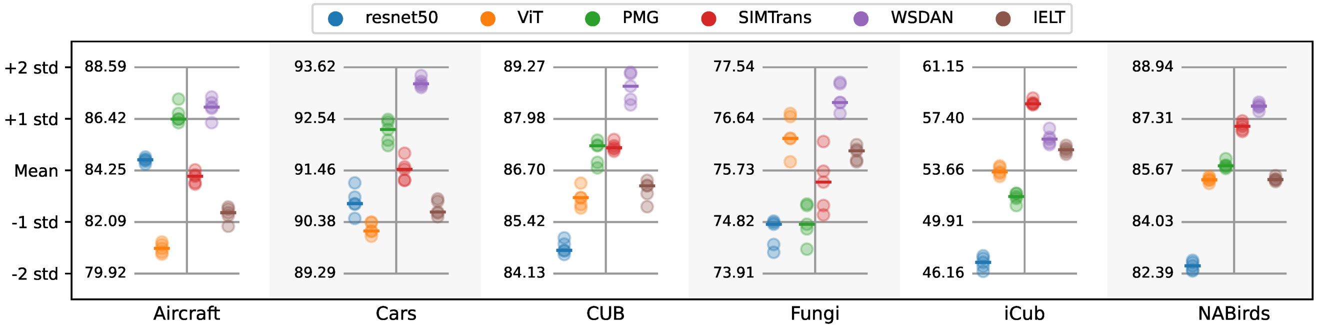

Figure 6. **Model performance**. Accuracy for five trials of each model on each dataset.

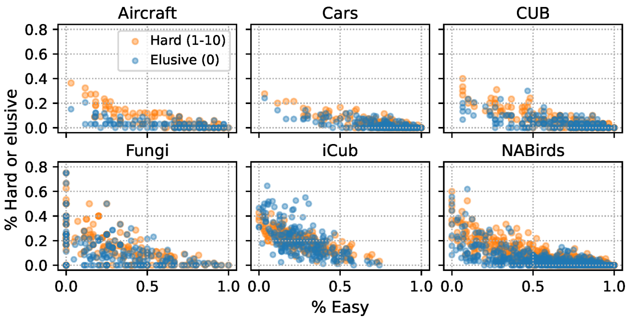

Figure 7. **Per-class easy vs. hard and elusive images**. Each class has an orange dot and a blue dot: the orange dot shows the percentage of hard images (1-10 models predict correctly) vs. easy images, while the blue dot shows the percentage of elusive images (no models predict correctly) vs. easy images.

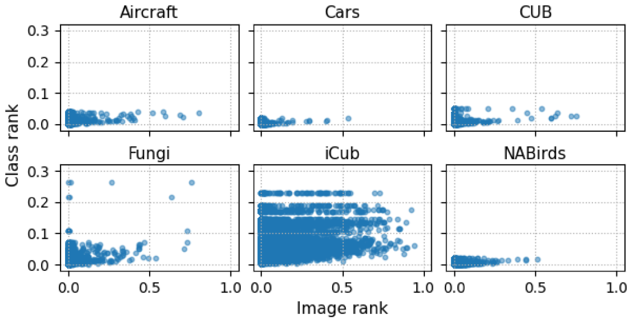

Figure 8. **Image vs. class rank**. Using prediction rank estimation (Sec. 2.2), we show the image rank compared to the class rank, which shows that certain images are much more challenging than the class average.

We choose these methods since together they are representative of major approaches to FGVC (see Sec. 5); a jigsaw-solving method (PMG), a feature-fusion and spatial-dependency learning method (SimTrans), a data augmentation and attention method (WSDAN) and an internal ensembling transformer method (IELT).

4.3. Training procedure

We use a standardized training procedure for all models and datasets, with a few specific adjustments where needed. We choose to keep things as standardized as possible across all methods, rather than using the exact backbone, schedule, or hyperparameters used by the original model implementations. We feel that this helps to remove confounding factors from the analysis. We include the precise details in the Supplementary Material A.

We train each model 5 times using a fixed set of 5 random seeds, which are the same across methods. We refer to the same model trained with a different seed as distinct *trials* or *runs*. We include these multiple trials in our analysis. Fig. 6 shows the overall accuracy across all models and runs on each dataset. When the pre-training, training and hyperparameters are standardized, WSDAN outperforms more recent works such as SIMTrans. As a community, we must carefully consider the impact of factors like data augmentation, optimizer, and learning rate when comparing SOTA FGVC methods.

4.4. Analysis

4.4.1 Measures of difficulty

Prediction overlap Fig. 3 shows the results of the analysis we describe in Sec. 2.1. In general, each dataset has a large set of easy images, and a smaller set of elusive images. Not all datasets are equally challenging; Fungi and iCub have a much smaller easy set than other methods, and larger elusive sets. We believe careful examination of the hard (few models predict correctly) and elusive sets is key to future model design, and we start this examination in our work in Sec. 4.4.2.

We also measure difficulty at the class level. In Fig. 7 we measure the extent to which each class is comprised of easy, hard (1–10 correct predictions) or elusive images. We find that while datasets such as Cars tend to be made up mostly of easy classes, other datasets (Fungi, NABirds, iCub) have a large set of classes that include no easy images; and iCub

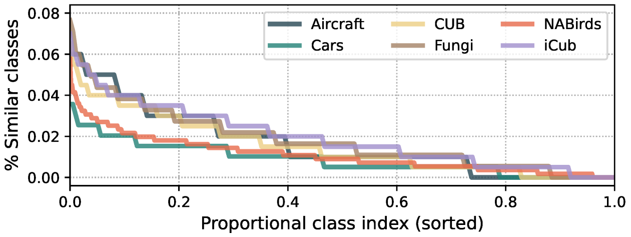

Figure 9. **Distribution of similar classes**. For each dataset, we show the distribution of similar classes (vertical axis) for each class (sorted on the horizontal axis).

has no trivial classes. We take a closer look at easy classes by model in Fig. 14 in the Supplementary Material.

Prediction rank We show both image and class difficulty using prediction rank estimation in Fig. 8. These results reinforce our prediction overlap findings; datasets like iCub and Fungi have harder images and harder classes, while Cars and CUB tend to have easier images and easier classes—also note that NABirds may look skewed, due to the much larger number of classes and thus smaller likelihood of a high image-rank score. We observe that there are certain images which are much harder than the class average; but there are also easy images belonging to more challenging classes. Certain classes are intrinsically more challenging—likely due to similar class confusion; but it's interesting that certain images are still easy.

Pairwise class confusion Using Pairwise Class Confusion to quantify similarity between classes, we calculate for each class how many other classes are highly similar. We define highly similar class pairs as those whose S_{KL} values (Eq. 4) are at least three standard deviations below the mean. Fig. 9 shows this distribution for each dataset. The distributions are similar across datasets, with the exception of Cars and NABirds. NABirds has proportionally fewer similar classes, but that is likely due to the fact that it has many more classes overall; the actual number of similar classes may be the same as or higher than in other datasets. For instance, 4% of classes for NABirds corresponds to 22 classes. Clearly, similar class confusion remains a major challenge in FGVC.

4.4.2 Diagnostics

Spatial distribution Certain properties of the object in an image, such as size or location, could affect classifier performance. To measure this, we correlate an objects' spatial properties with prediction overlap. Fig. 15 in the Supplementary Material shows these correlations for object size, aspect ratio and distance from the image center for each dataset. Surprisingly, in most cases there isn't any correlation. However, iCub and NABirds do show a correlation between object size (area of the bounding box) and prediction overlap; this is shown in Fig. 10. This confirms, unsurprisingly, that objects that are smaller with respect to the image are more challenging to recognize, and these types of images are more prevalent in iCub and NABirds.

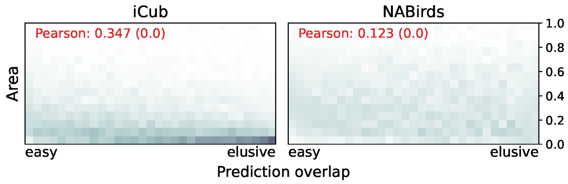

Figure 10. **Spatial distribution**. Distribution of relative bounding box area for each prediction overlap group in iCub and NABirds. Columns in each plot correspond to the prediction overlaps from Fig. 3, with 31 columns total ranging from easy to elusive. The Pearson correlation (and p-value) is shown in red.

Hard and elusive image examples Figure 11 shows some examples of images from the elusive and hard subsets, where we've defined "hard" in this case to be images where less than a third of the predictions are correct (1–10 correct predictions out of 30). The hard subset is particularly interesting because it contains images that are sometimes predicted correctly, but frequently aren't. The elusive subset may contain images that are mislabeled or not representative of the dataset; for example, **B2** and **B3** in Fig. 11. Some commonalities we observe in these challenging images include uncommon pose or viewpoint (**D1**, **E4**, **F3-4**, **G1-6**, **H5-8**); "camouflage" (**C1**, **E7**, **F4-5**); and distractor objects (**C3**, **F1**, **H4**), to name a few.

Discussion In Sec. 3 we discuss properties that can make images challenging, but which are difficult to measure. We see evidence of those properties in Fig. 11; but, we need to be careful about assuming causality. A more thorough examination of images across all prediction overlap groups—as well as correlation with the training set—would be needed for substantiation. There are also many images that don't have an obvious reason for being challenging; for example, **E8**, **F7** and **G8**. These are arguably the most interesting, because they should be solvable; understanding and addressing the reason behind their difficulty is an important direction for improving FGVC.

5. Related Work

For general image classification, categories can be distinguished based on major differences such as the presence or absence of key parts (*e.g.* a human has legs while a car has wheels). In FGVC, however, categories may have subtle differences in the shape or color of parts they have in common (*e.g.* differences in beak color or length between different sea bird species). While large datasets like ImageNet [9, 47] and iNat [56] have subdomains that are fine-grained, FGVC research more commonly uses

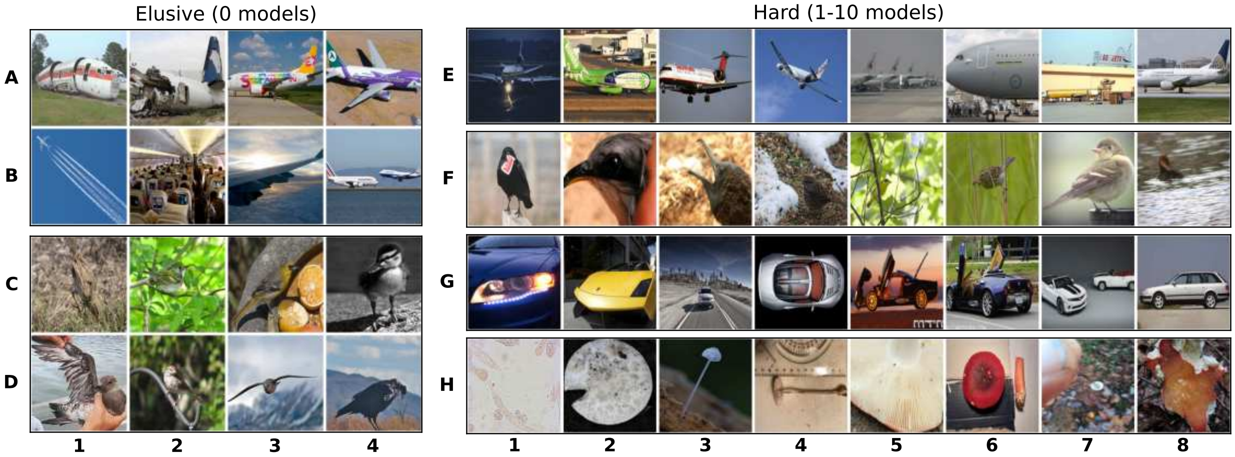

Figure 11. **Challenging images**. We show several examples of elusive images from Aircraft (**A**, **B**) and CUB (**C**, **D**). We also show examples of hard images, which were correctly predicted by between 1 and 10 models (out of 30), for Aircraft (**E**), CUB (**F**), Cars (**G**), and Fungi (**H**).

single-domain datasets such as Aircraft [41], Birds [55, 57], Cars [33], Dogs [30], and Flowers [43] as benchmarks, each having hundreds of classes with dozens of images each.

Deep learning has become the dominant approach to FGVC [5, 14, 19, 40, 60, 70], like computer vision in general. Most models are pre-trained on ImageNet [9, 47], while some approaches train on other, related datasets before the target FGVC task (*e.g*. [32]). Data augmentation has become ubiquitous in FGVC, and recent work explores optimizing the set of transformations [7].

The main approaches to FGVC include segmentation [4, 66], part-model [29, 50], and pose-alignment [17, 18, 20, 21, 38] methods, which attempt to isolate and model important class-specific features in a pose-invariant way. Pooling methods, such as bilinear [31, 36, 37, 67], Grassman [62], covariance [34, 35], and several learnable pooling methods [1, 8, 49], attempt to leverage second order statistics between deep CNN features. Attention methods [45, 46, 58] have also received a lot of "attention", with different methods employing recurrent models [16, 48, 69], reinforcement learning [39], metric learning [52], and part discovery [28, 70–72].

Other approaches modify the learning objective to account for high similarity and ambiguity between classes. Label smoothing [42, 53] and taxonomy-based schemes [3, 54] redistribute probability mass in the target distributions. Pairwise confusion [13] and maximum entropy [14] help reduce overconfidence by regularizing predictions. A recent plugin module combines feature maps with top-k predictions to overcome ambiguity [10]. Some methods resolve such differences via mixture of experts [68].

Feature fusion methods [6, 23, 45, 59, 65] have become increasingly popular with the introduction of the vision transformer [11]. Transformer-based [51, 61] and graph-based [2] approaches learn dependencies between image patches in order to attend to the most relevant fine-grained differences. Others learn dependencies by solving jigsaw puzzles [5, 12].

Previous FGVC analysis focuses primarily on the dataset or task itself as the vehicle for analysis, and analysis is limited to accuracy benchmarking [27]. Similar to Hoiem *et al*. [26] diagnosis of object detection, we attempt to categorize different types of errors.

6. Conclusion

In this paper, we carefully explore current FGVC challenges across a range of popular fine-grained datasets and models. Through overlap analysis and rank estimation, we establish a spectrum of image difficulty at a granular level, allowing us to examine the impact of specific challenges within the data on modern FGVC methods. Our code will be made publicly available as a toolkit that can be used to analyze additional datasets and models, giving FGVC researchers a rich set of tools for moving the field forward. As part of this toolkit, we provide a standardized training recipe for FGVC methods. This sort of fair comparison reveals the inadequacy of existing leaderboards, which conflate progress on general purpose optimizers, learning rate schedules, and data augmentation with progress on core FGVC problems. Finally, we contribute iCub to allow for additional analysis of bird classification methods.

Acknowledgements We would like to thank Matthew Gabbitas for helpful feedback on the manuscript. This work was supported by the National Science Foundation under Grant No. IIS1651832.

References

[1] Ardhendu Behera, Zachary Wharton, Pradeep Hewage, and Asish Bera. Context-aware attentional pooling (cap) for fine-grained visual classification, 2021. 8

[2] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31:6017–6031, 2022. 8

[3] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. *arXiv preprint arXiv:1912.09393*, 2019. 8

[4] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic Segmentation and Part Localization for Fine-Grained Categorization. In *ICCV*, 2013. 8

[5] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[6] Po-Yung Chou, Cheng-Hung Lin, and Wen-Chung Kao. A novel plug-in module for fine-grained visual classification, 2022. 8

[7] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[8] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel Pooling for Convolutional Neural Networks. In *CVPR*, 2017. 8

[9] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 7, 8

[10] Tuong Do, Huy Tran, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Fine-grained visual classification using self assessment classifier, 2022. 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 8

[12] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches. *arXiv preprint*, 2020. 5, 8

[13] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise Confusion for Fine-Grained Visual Classification. In *ECCV*, 2018. 8

[14] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-Entropy Fine Grained Classification. In *NeurIPS*, 2018. 8

[15] WA Falcon. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019. 12

[16] Jianlong Fu, Heliang Zheng, and Tao Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *CVPR*, 2017. 8

[17] E. Gavves, B. Fernando, C.G.M. Snoek, A.W.M. Smeulders, and T. Tuytelaars. Fine-Grained Categorization by Alignments. In *ICCV*, 2013. 8

[18] Efstratios Gavves, Basura Fernando, Cees G M Snoek, Arnold W M Smeulders, and Tinne Tuytelaars. Local Alignments for Fine-Grained Categorization. *IJCV*, 111(2):191–212, 2015. 8

[19] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[20] Christoph Goring, Erik Rodner, Alexander Freytag, and Joachim Denzler. Nonparametric Part Transfer for Fine-grained Recognition. In *CVPR*, 2014. 8

[21] Pei Guo and Ryan Farrell. Aligned to the Object, not to the Image: A Unified Pose-aligned Representation for Fine-grained Recognition. In *WACV*, 2019. 8

[22] Matthew Gwilliam, Adam Teuscher, Connor Anderson, and Ryan Farrell. Fair comparison: Quantifying variance in results for fine-grained visual categorization. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, jan 2021. 2

[23] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition, 2021. 8

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5

[25] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing Error in Object Detectors. In *ECCV*, 2012. 5

[26] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 8

[27] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections, 2021. 1, 8

[28] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. In *arXiv preprint*, 2019. 5, 8

[29] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1173–1182, 2016. 8

[30] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for Fine-Grained Image Categorization. In *CVPR Workshops (FGVC)*, 2011. 8

[31] Shu Kong and Charless C Fowlkes. Low-rank Bilinear Pooling for Fine-Grained Classification. In *CVPR*, 2017. 8

[32] 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. 8

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *ICCV Workshops (3DRR)*, 2013. 5, 8, 12

[34] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization. In *CVPR*, 2018. 8

[35] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2078, 2017. 8

[36] Tsung-Yu Lin and Subhransu Maji. Improved Bilinear Pooling with CNNs. In *BMVC*, 2017. 8

[37] T Y Lin, A RoyChowdhury, and S Maji. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *PAMI*, 2018. 8

[38] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. In *ECCV*, 9 2018. 8

[39] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. Localizing by Describing: Attribute-Guided Attention Localization for Fine-Grained Recognition. In *AAAI*, 2017. 8

[40] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 8

[41] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv.org*, 2013. 5, 8, 12

[42] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4696–4705, 2019. 8

[43] Maria-Elena Nilsback and Andrew Zisserman. A Visual Vocabulary for Flower Classification. In *CVPR*, 2006. 8

[44] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020-not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1525–1535, 2022. 5, 12

[45] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification, 2021. 8

[46] Pau Rodríguez, Josep M. Gonfaus, Guillem Cucurull, F. Xavier Roca, and Jordi Gonzàlez. Attend and Rectify: A Gated Attention Mechanism for Fine-Grained Recovery. In *ECCV*, 2018. 8

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 12 2015. 7, 8, 12

[48] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for Fine-Grained Categorization. In *ICLR Workshops*, 2015. 8

[49] Marcel Simon, Yang Gao, Trevor Darrell, Joachim Denzler, and Erik Rodner. Generalized Orderless Pooling Performs Implicit Salient Matching. In *ICCV*, 2017. 8

[50] Marcel Simon and Erik Rodner. Neural Activation Constellations: Unsupervised Part Model Discovery With Convolutional Networks. In *ICCV*, 2015. 8

[51] Hongbo Sun, Xiangteng He, and Yuxin Peng. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5853–5861, 2022. 5, 8

[52] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In *ECCV*, 2018. 8

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. 8

[54] Michael J Trammell, Priyanka Oberoi, James Egenrieder, and John Kaufhold. Contextual label smoothing with a phylogenetic tree on the inaturalist 2018 challenge dataset. *Washington Academy of Sciences. Journal of the Washington Academy of Sciences*, 105(1):23–45, 2019. 8

[55] 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. 5, 8, 12

[56] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 7

[57] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 1, 5, 8, 12

[58] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual Attention Network for Image Classification. In *CVPR*, pages 6450–6458, 2017. 8

[59] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization, 2022. 8

[60] Yaming Wang, Vlad I. Morariu, and Larry S. Davis. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition. In *CVPR*, 2018. 8

[61] Yu Wang, Shuo Ye, Shujian Yu, and Xinge You. R2-trans:fine-grained visual categorization with redundancy reduction, 2022. 8

[62] Xing Wei, Yihong Gong, Yue Zhang, Nanning Zheng, and Jiawei Zhang. Grassmann Pooling for Fine-Grained Visual Classification. In *ECCV*, 2018. 8

[63] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 12

[64] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 2023. 5

[65] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, pages 1–14, 2023. 8

[66] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing*, 25(10):4858–4872, 2016. 8

[67] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In *ECCV*, 2018. 8

[68] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, 2019. 8

[69] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified Visual Attention Networks for Fine-Grained Object Classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, 6 2017. 8

[70] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In *ICCV*, 2017. 8

[71] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In *CVPR*, 2019. 8

[72] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification, 2022. 8