Classes Are Not Equal: An Empirical Study on Image Recognition Fairness Supplementary Material

A. Class Examples in ImageNet

(a) Ostrich (b) Ostrich (c) MacaW (d) MacaW (e) Trilobite (f) Trilobite

Figure 8. **"Easy" class examples in ImageNet**

(a) Velvet (b) Velvet (c) Water jug (d) Water jug (e) Screwdriver (f) Screwdriver

Figure 9. **"Hard" class examples in ImageNet**

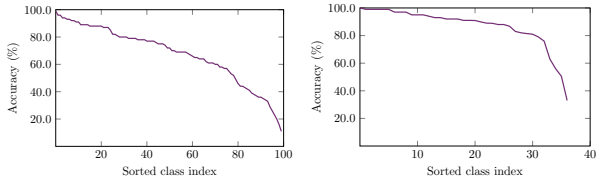
We show class examples in "Easy" and "Hard" classes in Figure 8 and Figure 9. "Easy" classes, like ostrich, macaw, and trilobite, are usually with simple scenarios. However, "Hard" classes can occur in much more complex scenarios. Take the "velvet" class as an example. Bags can be made of velvet. Velvet clothes for people or pets also belong to the "velvet" class. Thus, "Hard" classes can have overlap scenarios with other classes with a high probability, leading to model confusion and challenging optimization.

B. Diffusion Classifier

Oxford-IIIT Pet. Oxford-III pet dataset [36] consists of 37 category pets with roughly 200 images for each class. The images have large variations in scale, pose, and lighting. All images have an associated ground truth annotation of the breed.

Evaluation. The stable diffusion model [38] has become one of the most popular foundation models. We examine the fairness of the diffusion classifier [27] on CIFAR-100 and Oxford-IIIT Pet datasets. Checkpoint v2.0 of the stable diffusion model is adopted. With the pre-trained weights, we directly evaluate its zero-shot performance on CIFAR-100 and Oxford-IIIT Pet datasets. All configurations are the same as in diffusion classifier [27].

Results Analysis. The experimental results are listed in Figure 10. From Figure 10(a) and Figure 10(b), we can see that the class performance exhibits extreme disparity from 99.0% to 11.0% on CIFAR-100 and from 100.0% to 33.0% on Oxford-IIIT Pet. Although the stable diffusion model is trained with a huge amount of image-text pair data, it still faces the fairness challenge, demonstrating the prevalence of unfairness in vision-language models.



(a) CIFAR-100 (b) Oxford-IIIT Pet

(c) Zero-shot class accuracy of the stable diffusion model.

(d) Best and Worst class accuracy of the stable diffusion model.

Dataset	Best	Worst
CIFAR-100	99.0	11.0
Oxford-IIIT Pet	100.0	33.0

Figure 10. **Unfairness phenomenon exists in the stable diffusion model.**

C. More Other Datasets

We demonstrate that performance unfairness is prevalent in image classification. Besides ImageNet, CIFAR, and WIT-400M, other fine-grained benchmarks, including OxfordPets, StandfordCars, Flowers102, Food101, and FGVCAircraft, are also considered in our study.

StanfordCars. StanfordCars dataset [23] contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. Classes are typically at the level of Make, Model, Year, ex. 2012 Tesla Model S or 2012 BMW M3 coupe.

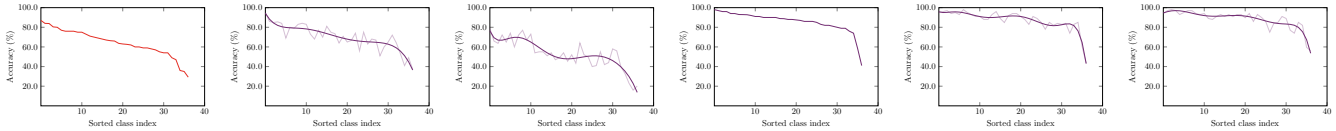
Flowers102. There are 102 flower categories in the Flowers102 dataset [33]. Each class consists of between 40 and 258 images. The images have large scale, pose, and light variations. In addition, there are categories that have large variations within the category and several very similar categories.

Food101. Food101 dataset [3] consists of 101 food categories, with 101,000 images. For each class, 250 manually reviewed test images are provided as well as 750 training images. On purpose, the training images were not cleaned, and thus still contain some amount of noise. This comes mostly in the form of intense colors and sometimes wrong labels.

FGVCAircraft. The dataset contains 10,200 images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes. It is divided into three equally sized training, validation, and test subsets.

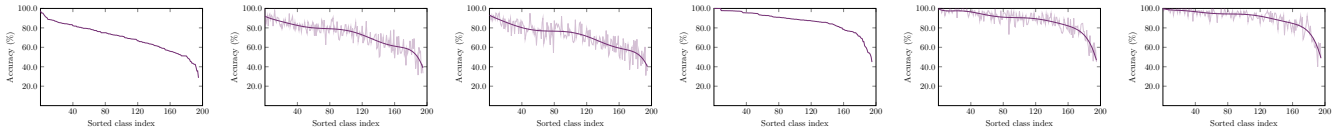
Table 5. **Unfairness phenomenon exists in the popular fine-grained recognition benchmarks**. "Best" represents the best class accuracy (%) while "Worst" denotes the worst class performance.

Backbone	Oxford-IIIT Pet		StanfordCars		Flowers102		Food101		FGVCAircraft	
	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst
Train from scratch										
ResNet-18	87.0	29.2	97.7	28.6	100.0	4.8	98.4	48.8	97.1	20.6
ResNet-34	95.0	37.1	100.0	31.4	100.0	5.6	99.2	52.0	97.0	21.2
ResNet-50	79.0	16.0	95.3	31.0	100.0	4.7	98.4	55.2	91.2	9.0
Train with initialization from ImageNet pre-train weights										
ResNet-18	98.0	41.0	100.0	44.8	100.0	50.0	100.0	57.2	100.0	33.3
ResNet-34	98.0	43.0	100.0	44.8	100.0	40.0	98.4	51.2	100.0	45.4
ResNet-50	98.0	56.0	100.0	40.0	100.0	50.0	99.2	48.4	100.0	39.3



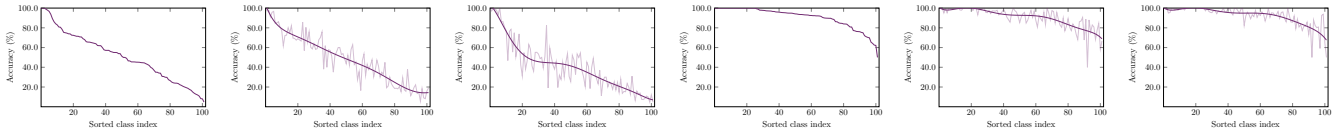
(a) ResNet-18 (b) ResNet-34 (c) ResNet-50 (d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 11. **Unfairness on the Oxford-IIIT Pet dataset.** (a), (b), and (c) are trained from scratch. (d), (e), and (f) are initialized with ImageNet pre-train weights.



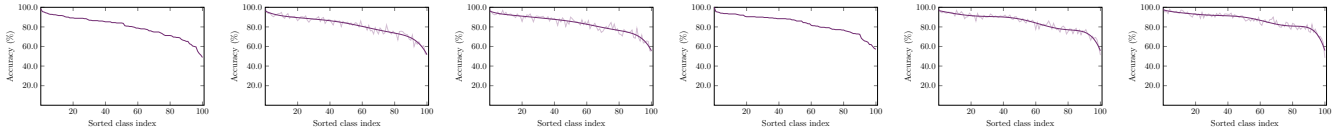
(a) ResNet-18 (b) ResNet-34 (c) ResNet-50 (d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 12. **Unfairness on the StanfordCars dataset.** (a), (b), and (c) are trained from scratch. (d), (e), and (f) are initialized with ImageNet pre-train weights.



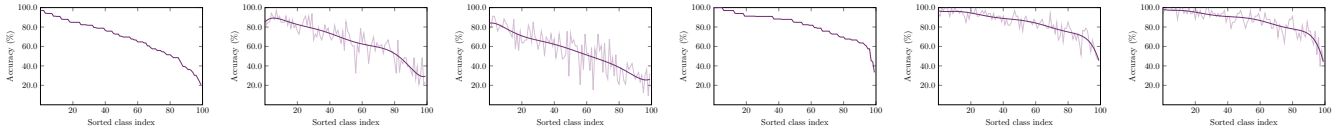
(a) ResNet-18 (b) ResNet-34 (c) ResNet-50 (d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 13. **Unfairness on the Flowers102 dataset.** (a), (b), and (c) are trained from scratch. (d), (e), and (f) are initialized with ImageNet pre-train weights.



(a) ResNet-18 (b) ResNet-34 (c) ResNet-50 (d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 14. **Unfairness on the Food101 dataset.** (a), (b), and (c) are trained from scratch. (d), (e), and (f) are initialized with ImageNet pre-train weights.



(a) ResNet-18 (b) ResNet-34 (c) ResNet-50 (d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 15. **Unfairness on the FGVCAircraft dataset.** (a), (b), and (c) are trained from scratch. (d), (e), and (f) are initialized with ImageNet pre-train weights.

Training and Evaluation. We use ResNet-18, ResNet-34, and ResNet-50 as our backbones. Following the training schedule on ImageNet, we use the same pre-process, *i.e.*, randomly crop and resize to 224×224 and then randomly horizontal flip. Models are trained in 100 epochs with a cosine learning rate strategy and a batch size of 256 on 8 GPUs. The initial learning rate and the weight decay are set to 0.1 and 1e-4 separately. The SGD optimizer with a momentum of 0.9 is used.

We evaluate the performance of models trained from scratch and initialized with ImageNet pre-train weights. ImageNet pre-train weights embed training data information of ImageNet. Thus, initialization with the pre-train weights can disturb the original training data distribution. Considering that, we use a ResNet-18 model trained from scratch as the reference model to sort classes under the case without ImageNet pre-train weight initialization. Otherwise, we use a ResNet-18 model trained with the weight initialization as the reference model to sort classes. As shown in Figures 1, 11, 12, 13, 14, 15, various models exhibit similar trends on the same dataset, implying that the unfairness highly depends on training data distribution.

Results Analysis. Our results are summarized in Figs 11, 12, 13, 14, 15 and Table 5. For models training from scratch, the performance unfairness is obvious on Oxford-IIIT pet, StandardCars, Flowers102, and FGVCAircraft datasets. Particularly, there is over 70% performance disparity between the best class and the worst class on the FGVCAircraft dataset, demonstrating the severe fairness issue. On models trained with initialization from ImageNet pre-train weights, the worst class performance significantly increases. However, the extreme performance imbalance still exists, specifically on Oxford-IIIT Pet, StanfordCars, and FGVCAircraft datasets.

Without initialization from ImageNet pre-train weights, we observe that the model performance can decrease as the capacity increases. The accuracy of the ResNet-50 model is lower than that of ResNet-18 and ResNet-34 on the Oxford-IIIT Pet dataset. This phenomenon can be caused by limited training data.

D. Equalized Odds Evaluation

Following the definition of Equalized Odds (EO), we extend it with a tighter constrain:

$$P(\hat{Y} = y_i | Y = y_i, A = y_i) = P(\hat{Y} = y_j | Y = y_j, A = y_j), \tag{12}$$

where y_i, $y_j \in$ 1,2,...,C. C is the number of classes. $\hat{Y}$ is the prediction. Y is the true label, and A refers to group membership. Here, we treat classes as groups. We report the maximum False Positive Error Rate (FPR) and False Negative Error Rate (FNR) disparities among C groups in Table 6.

Table 6. **EO for fairness on ImageNet.**

EO metrics	ResNet-50	ResNet-101	ViT-B
FPR balance	0.78	0.74	0.80
FNR balance	0.84	0.84	0.80