Understanding the class-specific effects of data augmentations

Polina Kirichenko[1,2], **Randall Balestriero**[2], **Mark Ibrahim**[2],
Ramakrishna Vedantam[2], **Hamed Firooz**[2], **Andrew Gordon Wilson**[1]
[1] New York University
[2] Meta AI

Abstract

Data augmentation (DA) is a major part of modern computer vision used to encode invariance and improve generalization. However, recent studies have shown that the effects of DA can be highly class dependent: augmentation strategies that improve average accuracy may significantly hurt the accuracies on a minority of individual classes, e.g. by as much as 20% on ImageNet. In this work, we explain this phenomenon from the perspective of interactions among class-conditional distributions. We find that most affected classes are inherently ambiguous, co-occur, or involve fine-grained distinctions. By using the higher-quality multi-label ImageNet annotations, we show the negative effects of data augmentation on per-class accuracy are significantly less severe.

1 Introduction

Data augmentation (DA) provides numerous benefits for training of deep neural networks including promoting invariance, providing regularization, and improving in- and out-of-distribution generalization and robustness (Hernández-García & König, 2018; Gontijo-Lopes et al., 2020; Balestriero et al., 2022b; Geiping et al., 2022). However, Balestriero et al. (2022a) and Bouchacourt et al. (2021) showed that strong DAs which are used by default in training of computer vision models may disproportionately hurt accuracies on some classes, e.g. with up to 20% class-level degradation on ImageNet compared to milder augmentation settings. Balestriero et al. (2022a) attempted to address this problem by only applying DA to the classes on which accuracy is not negatively affected and removing DA from the classes on which it leads to decreased performance. However, this strategy did not improve the accuracy on the affected classes, and Balestriero et al. (2022a) hypothesized that it is due to the model learning learning some general invariance from DAs being applied to the majority of classes that is not beneficial to the minority. Thus, several crucial open questions related to DA leading to class disparities remain unaddressed which we aim to understand: (1) why exactly the class-level performance degradation happens, (2) what kind of predictions and mistakes models make on those classes, and (3) why removing DA from those classes is not helpful for recovering performance (Balestriero et al., 2022a). In this work, we provide an explanation of the class-level performance degradation from the perspective of interactions between class-conditional distributions. In particular, our contributions are the following:

- We refine the per-class analysis of data augmentations correcting for label noise using multi-label annotations on ImageNet validation split (re-assessed labels from Beyer et al. (2020)) and systematically measure suboptimality of globally optimal data augmentation parameters for each class in terms of original and multi-label accuracy. Our analysis indicates that class-level performance degradation reported in Balestriero et al. (2022b) and Bouchacourt et al. (2021) is overestimated.
- We show that data augmentation significantly hurts top-1 classification accuracy specifically on ambiguous, co-occurring and fine-grained classes, which are often affected by label noise. We characterize each case in terms of the extent to which class-level performance drop can be attributed to label noise versus data augmentation.

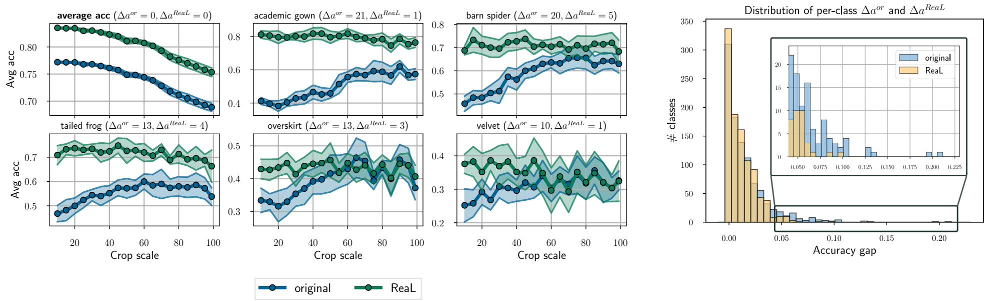


Figure 1: **Evaluating class-level performance using multi-label annotations reveals that negative effects of strong data augmentations are significantly muted. Left**: Average (top left panel) and individual class (remaining panels) validation top-1 accuracies of ResNet-50 on ImageNet computed with original and ReaL labels (Beyer et al., 2020) as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest Δa_k^{or}: the difference between the highest accuracy on that class $\max_s a_k^{or}(s)$ and accuracy of the model trained with $s = 10\%$ using original labels for evaluation. **Right**: Distribution of per-class accuracy gaps Δa_k for original and ReaL labels. The distribution of Δa_k^{or} has a heavier tail compared to the one computed with ReaL labels.

2 Related Work

Understanding data augmentation, invariance and regularization. Hernández-García & König (2018) analyzed the DA from the perspective of implicit regularization. Botev et al. (2022) propose an explicit regularizer that encourages invariance and show that it leads to improved generalization. Balestriero et al. (2022b) derive an explicit regularizer to simulate DA to quantify its benefits and limitations and estimate the number of samples for learning invariance. Gontijo-Lopes et al. (2020) and Geiping et al. (2022) study the mechanisms behind the effectiveness of DA, which include data diversity, exchange rates between real and augmented data, additional stochasticity and distribution shift. Bouchacourt et al. (2021) measure the learned invariances using DA. Lin et al. (2022) studied how data augmentation induces implicit spectral regularization which improves generalization. For a detailed review of DA techniques, see Xu et al. (2023).

Biases of data augmentations. While DA is commonly applied to improve generalization and robustness, a number of prior works identified its potential negative effects. Hermann et al. (2020) showed that decreasing minimum crop size in Random Resized Crops leads to increased texture bias. Shah et al. (2022) showed that using standard DA amplifies model's reliance on spurious features compared to model trained without augmentations. Idrissi et al. (2022) provided a thorough analysis on how the strength of DA for different transformations has a disparate effect on subgroups of data corresponding to different factors of variation. Kapoor et al. (2022) suggested that DA can cause models to misinterpret uncertainty. Izmailov et al. (2022) showed that DA can hurt the quality of learned features on some classification tasks with spurious correlations. Balestriero et al. (2022a) and Bouchacourt et al. (2021) showed that strong DA may disproportionately hurt accuracies on some classes on ImageNet, and in this work we focus on understanding this class-level performance degradation through the lens of interactions between classes.

Multi-label annotations on ImageNet. A number of prior works identified that ImageNet dataset contains label noise such as ambiguous classes, multi-object images and mislabeled examples (Beyer et al., 2020; Shankar et al., 2020; Vasudevan et al., 2022; Northcutt et al., 2021b; Stock & Cisse, 2018; Northcutt et al., 2021a). Tsipras et al. (2020) found that nearly 20% of ImageNet validation set images contain objects from multiple classes. Hooker et al. (2019) ran a human study and showed that examples most affected by pruning a neural network are often mislabeled, multi-object or fine-grained. Yun et al. (2021) generate pixel-level multi-label annotations for ImageNet train split using a large-scale computer vision model. Beyer et al. (2020) provide re-assessed (ReaL) multi-label annotations for ImageNet validation split which aim to resolve label noise issues, and we use ReaL labels in our analysis to refine the understanding of per-class effects of DA.

3 Experimental Setup

Since we aim to understand class-level accuracy degradation emerging with strong data augmentations reported in Balestriero et al. (2022a), we closely follow their experimental setup to train models.

Training details. We trained a ResNet-50 on ImageNet for 88 epochs with SGD using a cyclic learning rate schedule, batch size 256, and weight decay 10^{-4}. For each hyper-parameter setting, we train networks with 10 different random seeds.

Data augmentation. For our analysis, we used random horizontal flips and random resized crop (RRC) DA when training our models which are the most commonly used transformations. In particular, for an input image of size $h \times w$ the RRC transformation first samples the crop scale $s \sim U[s_{low}, s_{up}]$ and the aspect ratio $r \sim U[r_{low}, r_{up}]$, takes random a crop of size $\sqrt{shwr} \times \sqrt{shw/r}$ and resizes it back to the input size required for the model. In our experiments we use the standard values for $s_{up} = 100\%, r_{low} = 3/4, r_{up} = 4/3$, and we vary the lower bound on the crop scale s_{low} (for simplicity, we will further use s to denote the lower bound on crop scale) between 10% and 100% which controls the strength of DA (with $s = 10\%$ and $s = 100\%$ corresponding to strongest and weakest DA strength respectively). Note that the default value in `pytorch` RRC implementation is $s = 8\%$.

ReaL labels. Beyer et al. (2020) used large-scale models to generate new label proposals for the validation split of ImageNet which were then evaluated by human annotators. These Reassessed Labels (ReaL) aim to correct the label noise present in the original labels including mislabeled examples, multi-object images and ambiguous classes. Since there are possibly multiple ReaL labels for each images, the prediction is considered correct if it falls in the set of the plausible labels.

Metrics and notation. We are interested in evaluating the accuracy $a_k(s)$ on class k as a function of the hyper-parameter of DA, e.g. RRC scale lower bound s. Balestriero et al. (2022a) and Bouchacourt et al. (2021) reported that on some classes RRC can hurt the accuracy by more than 20%. Balestriero et al. (2022a) compare the per-class accuracy of the models trained with the strongest DA with $s = 8\%$ and the model trained with no DA ($s = 100\%$ which effectively just resizes input images without cropping), while Bouchacourt et al. (2021) compared the models trained with RRC with $s = 8\%$ and models trained with fixed size center crop. In our analysis, we want to measure how suboptimal the choice of DA hyper-parameter s is for a particular class if we choose s using average accuracy on validation data as opposed to validation accuracy on that class (the optimal s in terms of the average accuracy is the strongest DA $s = 10\%$ which improves accuracy on the majority of the classes). More formally, we evaluate $\Delta a_k = a_k(s_k^*) - a_k(s^*)$, where s^* is the optimal value of s based on the average accuracy: $s^* = \arg\max \sum_k a_k(s)$ (if we assume balanced classes), and s_k^* is the optimal s for class k: $s_k^* = \arg\max_s a_k(s)$. We report $a_k(s)$ and Δa_k for both original and ReaL labels.

4 Per-class accuracy degradation under strong data augmentation is overestimated due to label noise

Previous studies reported that the performance of ImageNet models is effectively better if we evaluate it using re-assessed multi-label annotations which address label noise issues in ImageNet (Beyer et al., 2020; Shankar et al., 2020; Vasudevan et al., 2022). These works showed that recent performance improvements on ImageNet might be saturating, but the effects of such label noise on per-class performance has not been previously studied. In particular, it is unclear how label noise would affect the results of Balestriero et al. (2022a) and Bouchacourt et al. (2021) on the effects of DA on class-level performance.

We observe that **for most classes with significant drops in accuracy on original labels, the class-level ReaL multi-label accuracy is considerably less affected.** First to compare the class-level effects of DAs between original and ReaL labels in an aggregated way, we plot the distributions of Δa_k^{or} and Δa_k^{ReaL} values on the right panel of Figure 1. Note that the distribution of the accuracy gap scores computed with the original labels has a heavier tail. For ReaL labels, there are much fewer classes which have a drop in accuracy over 4% when using the model with the strongest augmentation (34 classes according to ReaL accuracy as opposed to 98 classes with the original labels), and there are no classes with a drop in accuracy over 10%.

On the left panel of Figure 1, we show the average and individual class accuracies, both using original and ReaL labels, against RRC crop scale lower bound s for classes with the highest Δa^{or}, i.e. the classes with the highest drop in accuracy when choosing $s = 10\%$ (optimal for the average accuracy) instead of optimal value $s^*_k = \arg\max_s a_k(s)$ for each class. Both a^{or}_k and a^{ReaL}_k are evaluated on images from ImageNet validation splits that have the original label k. The larger selection of classes with the highest Δa^{or} is shown in Appendix Figure 3. For many classes which are hurt by using stronger DA, the ReaL accuracy is much less affected. For example, for the class "academic gown" the original top-1 accuracy is decreased by 21% from 62% to 41% if we use the model trained with RRC with $s = 10\%$ compared to the optimal for that class model with $s = 90\%$, while ReaL multi-label accuracy is not significantly affected with $\Delta a^{ReaL} = 1\%$ and the model trained with $s = 10\%$ is comparable to the optimal one. For most classes in Appendix Figure 3 (with a few exceptions like "Siberian husky", "tobacco shop", "bighorn") Δa^{ReaL} is significantly lower than Δa^{or} and overall in most cases $a^{ReaL}(s)$ either improves as we increase DA strength or is not significantly affected by the choice of s.

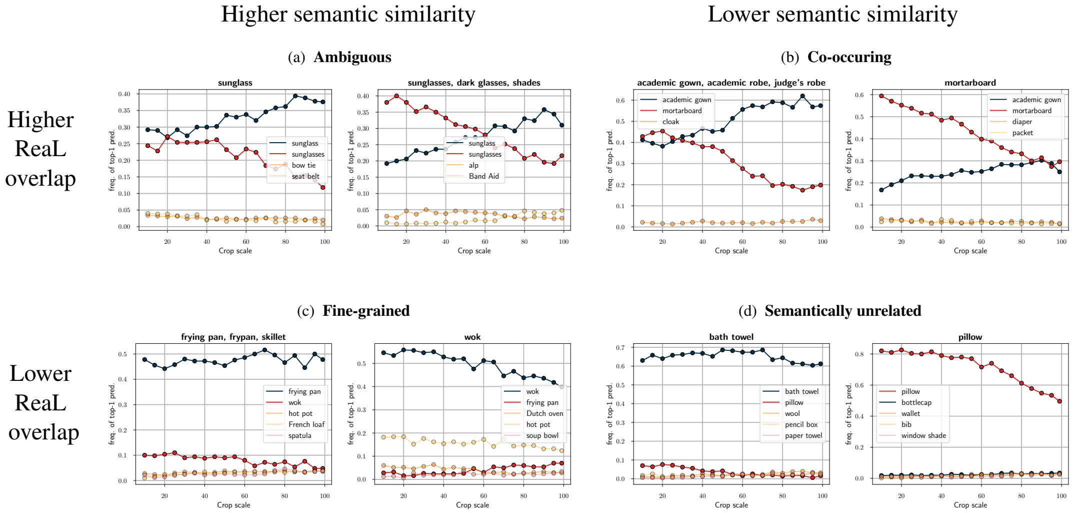
However, there are still some classes for which it is beneficial to use the crop scale higher than $s = 10\%$ for the optimal ReaL accuracy, and in Appendix Figure 4 we show per-class accuracy trends against s for the classes with the highest Δa^{ReaL}. Some of them (especially classes from the "animal" categories) may still be affected by the remaining label noise (Vasudevan et al., 2022; Van Horn et al., 2015; Shankar et al., 2020; Luccioni & Rolnick, 2022; Beyer et al., 2020), while for other classes it is in fact suboptimal for their recall to use the strongest DA. It remains puzzling (1) why strong DA hurt some classes in terms of ReaL accuracy, and (2) what leads to the high discrepancy between original and ReaL accuracy trends on most classes with the highest Δa^{or} since correcting for the label noise could have just shifted the class-level trends higher, equally improving the accuracy for all DA levels. We study these questions from the perspective of interactions among class-conditional distributions induced by DAs.

5 The framework of class-conditional distributions induced by data augmentations for reasoning about class-specific effects

To aid our understanding of the class-specific effects of data augmentations, it can be helpful to reason about them in terms of how they affect class-conditional distributions of the training data. In particular, this perspective can help us categorize the effects of data augmentations (see Section 6). We are interested in understanding how a particular parametrized class of transformations $\mathcal{T}_\alpha(\cdot)$ changes the data distribution for each class $X_k \sim p_k(x)$, e.g. in our case $\mathcal{T}_\alpha(\cdot)$ represents the family of RRC parametrized by the lower bound on the crop scale s. We denote the train data distribution by $p(x) = 1/K \sum_i p_i(x)$ and the augmented class distributions by $\mathcal{T}_\alpha(p_k)(\cdot)$. We assume that the support of p_k is a subset of the support of $\mathcal{T}_\alpha(p_k)$ (i.e. the original images are included in the set of all their possible RRC augmentations). Typically, prior works discussed potential harmful effects of DA in cases when it is not *label preserving*, i.e. $P[f^*(X^\alpha_k) = k] \ll 1$ where $X^\alpha_k \sim \mathcal{T}_\alpha(p_k)$ and $f^*(\cdot)$ is a true labeling function. However, it may not necessarily be problematic if the some samples from $\mathcal{T}_\alpha(p_k)$ are out-of-distribution for class k if they are in general out-of-distribution for $\mathcal{T}_\alpha(p)$. At the same time, if supports of $\mathcal{T}_\alpha(p_k)$ and $\mathcal{T}_\alpha(p_l)$ for two classes k and l overlap, especially in high-density regions, the model might be optimized to predict different labels k and l on similar inputs corresponding to features from both classes k and l which will lead to performance degradation. Some class distributions p_k and p_l are intrinsically almost coinciding or highly overlapping in ImageNet dataset, while others have distinct supports, but in all cases the parameters of the transformation class α will control the overlap of the induced class distributions, and thus the biases of the model when making predictions on such classes.

6 Data augmentation most significantly affects classification of ambiguous, co-occurring and fine-grained categories

In this section, we aim to understand the reasons behind per-class accuracy degradation when using stronger data augmentation and the high discrepancy between class-level accuracy trends using original and ReaL labels analyzing the most common mistakes the models make on these classes and how they evolve as we vary the data augmentation strength. We consider the classes most affected by



(a) **Ambiguous**

(b) **Co-occuring**

(c) **Fine-grained**

(d) **Semantically unrelated**

Figure 2: **Types of class pair confusions affected by data augmentations with varied semantic similarity and data distribution overlap.** Each panel shows the frequency of the most commonly predicted labels for a given class against the Random Resized Crop's scale hyper-parameter. We categorize types of class confusions into (a) ambiguous, (b) co-occurring, (c) fine-grained and (d) semantically unrelated, depending on inherent class overlap and similarity. In each of these categories data augmentation either controls the model's prediction preference among the plausible labels or biases model towards mistakes due to the overlap of the data augmentation induced class distributions as in the panel (d).

strong DA (see Appendix Figures 3 and 4) which do not belong to the "animal" subtree category in the WordNet hierarchy (Fellbaum, 1998) since fine-grained animal classes were reported to have higher label noise in previous studies (Van Horn et al., 2015; Shankar et al., 2020; Luccioni & Rolnick, 2022; Beyer et al., 2020). We focus on 50 classes with the highest Δa^{or} (corresponding to $\Delta a^{or} > 5.8\%$), and 50 classes with the highest Δa^{ReaL} (corresponding to $\Delta a^{ReaL} > 3.8\%$).

We analyze the set of predictions generated by 10 independently trained models for each RRC crop scale parameter s value on all validation images from a particular class according to the original labels (there are 50 images for each class in ImageNet validation). For each class, among all generated predictions, we identify 5 most common ones (filtering out those which occur in less than 4% cases for any crop scale level) and track the frequency of these predictions against RRC crop scale. We additionally perform the same analysis on some of the frequently confused classes (which may not be among the classes most hurt by strong DA).

In general, we observe that in many cases DA strength controls the model's preference in predicting one or another plausible ReaL label or among semantically similar classes. We roughly outline the most common types of confusions on the classes which are significantly affected by DAs which differ in the extent to which the accuracy degradation can be attributed to label noise versus the presence of DA, and characterize how DA effectively changes the data distribution of these classes which affects performance. These categories are closely related to common mistake types on ImageNet identified by Beyer et al. (2020) and Vasudevan et al. (2022), but we focus on class-level interactions as opposed to instance-level mistakes and particularly connect them to the impact of DAs. We use the following criteria to identify a confusion category for a pair of classes:

(1) **Semantic similarity** which we can measure by (a) WordNet class similarity, in particular, we use Wu-Palmer score (between 0 and 1) which relies on the categories' most specific common ancestor in the WordNet tree, and (b) similarity of the class name embeddings[1].

(2) **ReaL labels co-occurence** between classes i and j: we consider intersection over union $IoU_{ij} = \sum_L I[i \in L] \times I[j \in L] / \sum_L I[i \in L \text{ or } j \in L]$ as well as one-sided overlap $C_{ij} = \sum_L I[i \in L] \times I[j \in L] / \sum_L I[i \in L]$ where the summation is over ReaL label

[1] We use NLTK library for WordNet and spaCy library for embeddings similarity.

sets L for all examples. Assuming that train and test are coming from similar distributions, we can treat this as an approximate measure of class overlap or distance between distributions p_i and p_j. We expect that for classes with high overlap in ReaL labels, the majority of confusions made by model $f : \mathcal{X} \to \mathcal{Y}$ of class i for j will be resolved by ReaL labels, and measure it by: $R_{ij} = \sum_{(x,L):\ x \in X_i} I[f(x) = j] \times I[j \in L] / \sum_{(x,L):\ x \in X_i} I[f(x) = j]$.

Using these metrics, depending on a higher or lower semantic similarity and higher or lower ReaL labels overlap, we categorize confused class pairs as *ambiguous*, *co-occurring*, *fine-grained* or *semantically unrelated*. The following subsections discuss each category in detail, and the examples are shown in Figure 2 and Appendix Figure 5.

6.1 Classes where the negative effects of data augmentation are resolved after removing label noise

Intrinsically ambiguous or semantically (close to) identical classes. Prior works e.g. Beyer et al. (2020); Shankar et al. (2020); Vasudevan et al. (2022); Tsipras et al. (2020) identified that some pairs of ImageNet classes are practically indistinguishable, e.g. "sunglasses" and "sunglass", "monitor" and "screen", "maillot" and "maillot, tank suit". These pairs of classes would generally have higher semantic similarity and higher overlap in ReaL labels (in particular, high IoU for equivalent classes and high one-sided overlap for subcategories). We observe that in many cases the accuracy on one class within the ambiguous pair degrades with stronger augmentations, while the accuracy on another one improves. The supports of distributions of these class pairs p_i and p_j highly overlap or even coincide, but with varying α depending on how the supports of $\mathcal{T}_\alpha(p_i)$ and $\mathcal{T}_\alpha(p_j)$ overlap the model would be biased towards predicting on of the classes. In Figure 2(a) we show how the frequencies of most commonly predicted labels change on an ambiguous pair of classes "sunglass" and "sunglasses" as we vary the crop scale parameter (these classes overlap with $C_{ij} = 91.1\%$ and 99% of confusions are corrected by ReaL labels). We note that for images from both classes the frequency of "sunglasses" label increases with stronger DAs while "sunglass" predictions have the opposite trend.

Models trained on ImageNet often achieve a better-than-random-guess accuracy when classifying between these classes due to overfitting to marginal statistical differences and idiosyncrasies of their labeling pipeline. While DA strength controls model's bias towards predicting one or another plausible label, the models are not effectively making mistakes when confusing such classes.

6.2 Classes where data augmentation may amplify or cause problematic misclassification

For the categories described below, the classes become more ambiguous or overlapping *in particular* when strong DA is applied during training.

Co-occurring or overlapping classes. There is a number of classes on ImageNet which correspond to semantically different objects which often appear together, e.g. "academic gown" and "mortarboard", "Windsor tie" and "suit", "assault rifle" and "military uniform". These pairs of classes have rather high overlap in ReaL labels (depending on the spurious correlation strength) and their semantic similarity can vary (but it would be lower than for ambiguous classes). The class distributions of co-occurring classes inherently overlap, however, stronger DAs may increase this overlap in class distribution supports, for example, with RRC we may augment the sample such that only the spuriously co-occurring object is left on the image, but the model would still be optimized to predict the original label. It was previously shown that RRC can increase model's reliance on spurious correlations (Hermann et al., 2020; Shah et al., 2022) which can lead to real mistakes. In Figure 2(b) we show how DA strength impacts model's bias towards predicting "academic gown" or "mortarboard" class (for which $C_{ij} = 72\%$ and 96% confusions resolved by ReaL labels).

Fine-grained categories. There is a number of semantically related class pairs like "tobacco shop" and "barbershop", "frying pan" and "wok", "violin" and "cello", where objects appear in related contexts, share some visually similar features and generally represent fine-grained categories of a similar object type. These classes have high semantic similarity and are not significantly overlapping (sometimes they are affected by mislabeling but generally not multi-object). The class distributions for such categories are close to each other or slightly overlapping, but strong DA pulls them closer, and $\mathcal{T}(p_i)$ and $\mathcal{T}(p_k)$ would be more overlapping due to e.g. RRC resulting in the crops of the visually similar features or shared contexts in the augmented images from different categories. In Figure 2(c) we show how model's most common predictions change depending on RRC crop scale for fine-grained classes "frying pan" and "wok" (for which $C_{ij} = 10\%$, only 12% of confusions were corrected by ReaL labels, while their WordNet distance is 0.92).

Semantically unrelated. In the rare but most problematic cases, the stronger DA will result in confusion of semantically unrelated classes (while they could possibly share some low-level features, they are semantically dissimilar and their distributions p_i and p_j and ReaL labels do not overlap, and they get confused with one another specifically because of strong DA), for example, categories like "muzzle" and "sandal", "bath towel" and "pillow". Figure 2(d) shows how confusions between unrelated classes "bath towel" and "pillow" emerge with stronger DA.

In Appendix we show a larger selection of example pairs from each category. Among the confusions on the classes most significantly hurt in original accuracy approximately 55% are co-occurring, 35% are fine-grained and 10% are ambiguous classes, while on the classes most affected in their ReaL accuracy around a half of the confusions correspond to fine-grained with another half corresponding to co-occurring classes. The confusion of semantically unrelated categories is rare, while it is potentially most concerning since it corresponds to more severe mistakes.

7 Effects of data augmentation on per-class precision and recall

As discussed in Sections 5 and 6, if class-conditional distributions induced by data augmentation $\mathcal{T}_\alpha(\cdot)$ start having overlapping support, we might observe a degradation in class-level performance. In the specific case when $f^*(x_k^\alpha) = l$ for some augmented samples $x_k^\alpha \sim \mathcal{T}_\alpha(p_k)$ from class k, the model will be optimized to predict the label k effectively on the examples from class l, which will result in reduced precision for class k and reduced accuracy (or recall) on class l. For example, when the model is trained to classify between the images from classes "cars" and "wheels", strong RRC will sometimes produce augmented car images that are zoomed in on the wheel (effectively coming from the "wheel" data distribution). However, the model has to predict the "car" label on said images which would result in the model sometimes predicting "car" for images from the "wheel" during evaluation (and, thus, reduced accuracy for "wheels" and reduced precision for "cars"). In other cases x_k^α will not belong to any of the training classes, but if augmented samples x_k^α and x_l^α from different classes $k \neq l$ will be focus on visually similar features, the model will start confusing these classes, and might be biased towards predicting one or another depending on their statistical differences in train data.

Since non-label-preserving augmentations result in model being optimized to predict the label k on an input that is not coming from the distribution p_k, it is also important to measure class-level *precision* as a function of data augmentation strength, and study the classes whose precision decreases as the model is trained with stronger DAs. In Appendix Figure 6, we plot per-class precision (computed with original and ReaL labels) against RRC crop scale lower bound s for classes with the highest drop in accuracy between model with $s = 10\%$ and the model with optimal value s^* in terms of precision computed with ReaL labels on that class. Interestingly, some of these classes are confused with classes that are most hurt in recall (see Appendix Figure 4), e.g. "barbershop" and "tobacco shop", or "honeycomb" and "apiary". We hypothesize that taking into account the classes most affected in precision is important for removing the class-level negative effects introduced by data augmentations.

8 Discussion

In this work, we provide insights into the class-level accuracy degradation on ImageNet when applying stronger augmentations from the perspective of interactions among class-conditional distributions. We observe that the most significantly affected classes are inherently ambiguous, co-occur, or involve fine-grained distinctions. These categories often suffer from label noise and thus the overall negative effect is significantly muted when evaluating performance with cleaner multi-label annotations.

References

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*, 2022a.

Randall Balestriero, Ishan Misra, and Yann LeCun. A data-augmentation is worth a thousand samples: Exact quantification from analytical augmented sample moments. *arXiv preprint arXiv:2202.08325*, 2022b.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Aleksander Botev, Matthias Bauer, and Soham De. Regularising for invariance to data augmentation improves supervised learning. *arXiv preprint arXiv:2203.03304*, 2022.

Diane Bouchacourt, Mark Ibrahim, and Ari Morcos. Grounding inductive biases in natural images: invariance stems from variations in data. *Advances in Neural Information Processing Systems*, 34: 19566–19579, 2021.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 05 1998. ISBN 9780262272551. doi: 10.7551/mitpress/7287.001.0001. URL https://doi.org/10.7551/mitpress/7287.001.0001.

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. *arXiv preprint arXiv:2210.06441*, 2022.

Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 19000–19015, 2020.

Alex Hernández-García and Peter König. Further advantages of data augmentation on convolutional neural networks. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pp. 95–103. Springer, 2018.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*, 2022.

Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew Gordon Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.

Chi-Heng Lin, Chiraag Kaushik, Eva L Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *arXiv preprint arXiv:2210.05021*, 2022.

Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity. *arXiv preprint arXiv:2208.11695*, 2022.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021a.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021b.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms. *arXiv preprint arXiv:2211.12491*, 2022.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pp. 8634–8644. PMLR, 2020.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–512, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pp. 9625–9635. PMLR, 2020.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *arXiv preprint arXiv:2205.04596*, 2022.

Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, pp. 109347, 2023.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2340–2350, 2021.

A Additional plots

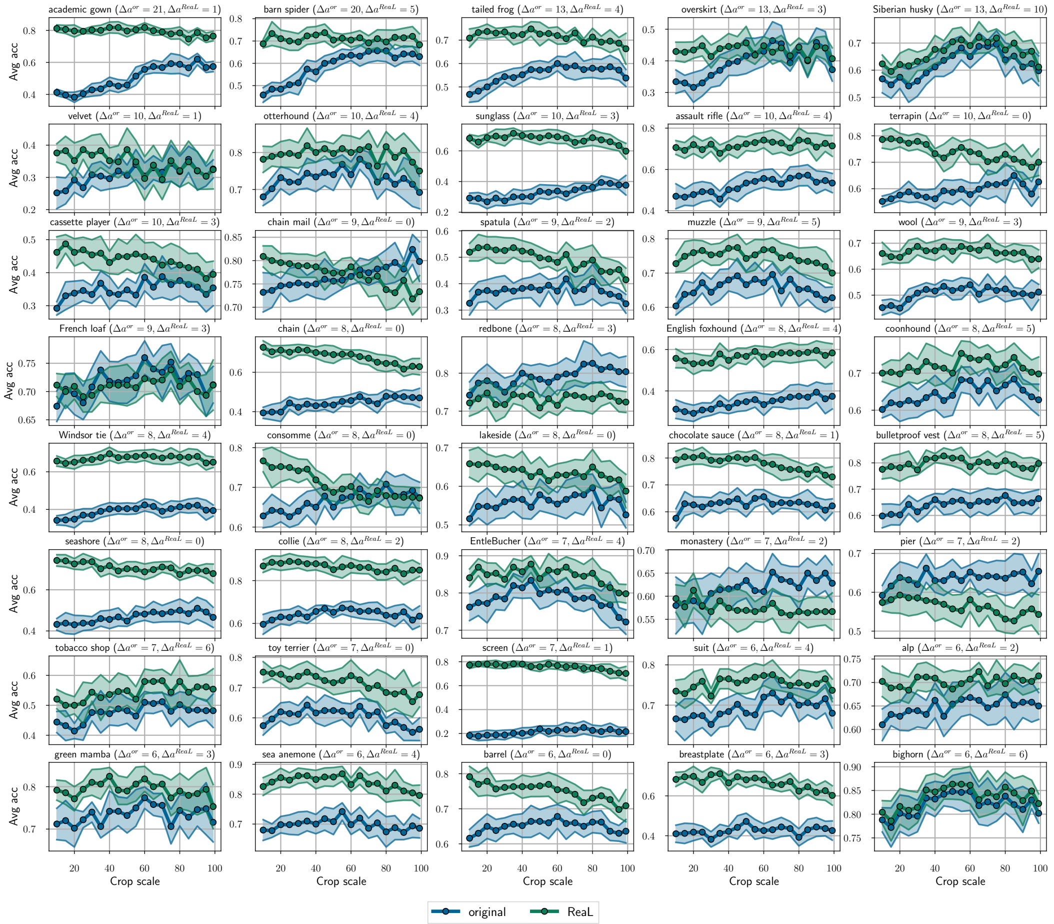


Figure 3: Individual class validation top-1 accuracies of ResNet-50 on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest Δa_k^{or}: the difference between the highest accuracy on that class $\max_s a_k^{or}(s)$ and accuracy of the model trained with $s = 10\%$ using original labels for evaluation. We report the mean and standard deviation over 10 independent runs of the network.

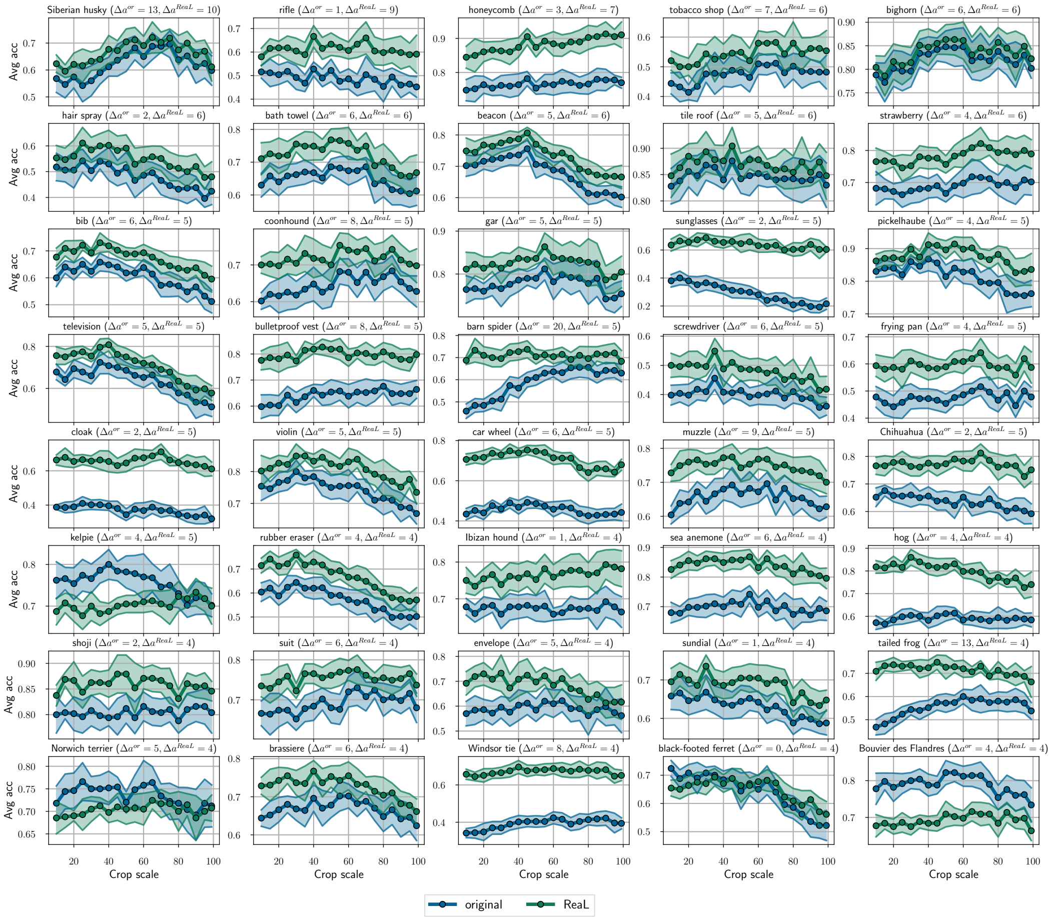


Figure 4: Individual class validation top-1 accuracies of ResNet-50 on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest Δa_k^{ReaL}: the difference between the highest accuracy on that class $\max_s a_k^{ReaL}(s)$ and accuracy of the model trained with $s = 10\%$ using ReaL labels for evaluation (i.e. similar to Figure 3 but choosing classes based on highest Δa^{ReaL} instead of Δa^{or}). We report the mean and standard deviation over 10 independent runs of the network.

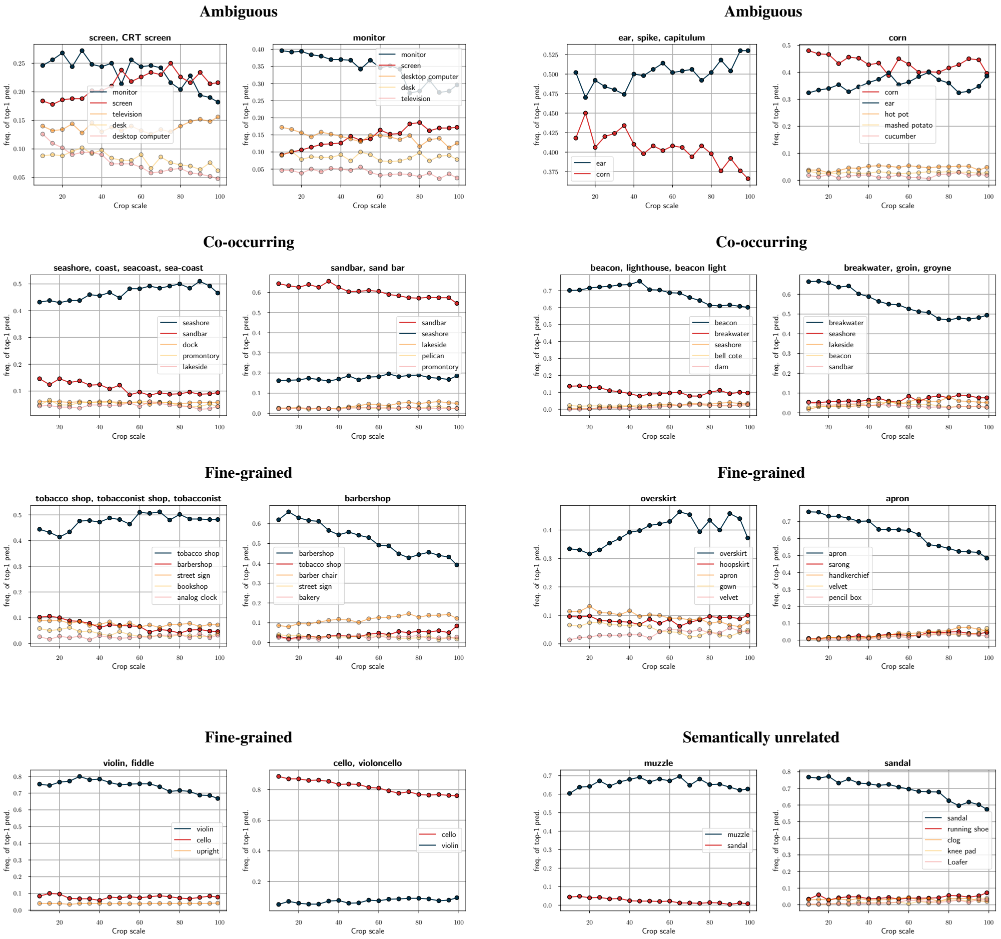


Figure 5: Each panel shows the frequency of the most commonly predicted labels for a given class against the Random Resized Crop's scale hyper-parameter. We categorize types of class confusions into ambiguous, co-occurring, fine-grained and semantically unrelated, depending on inherent class overlap and similarity. In each of these categories data augmentation either controls the model's prediction preference among the plausible labels or biases model towards mistakes due to the overlap of the data augmentation induced class distributions.

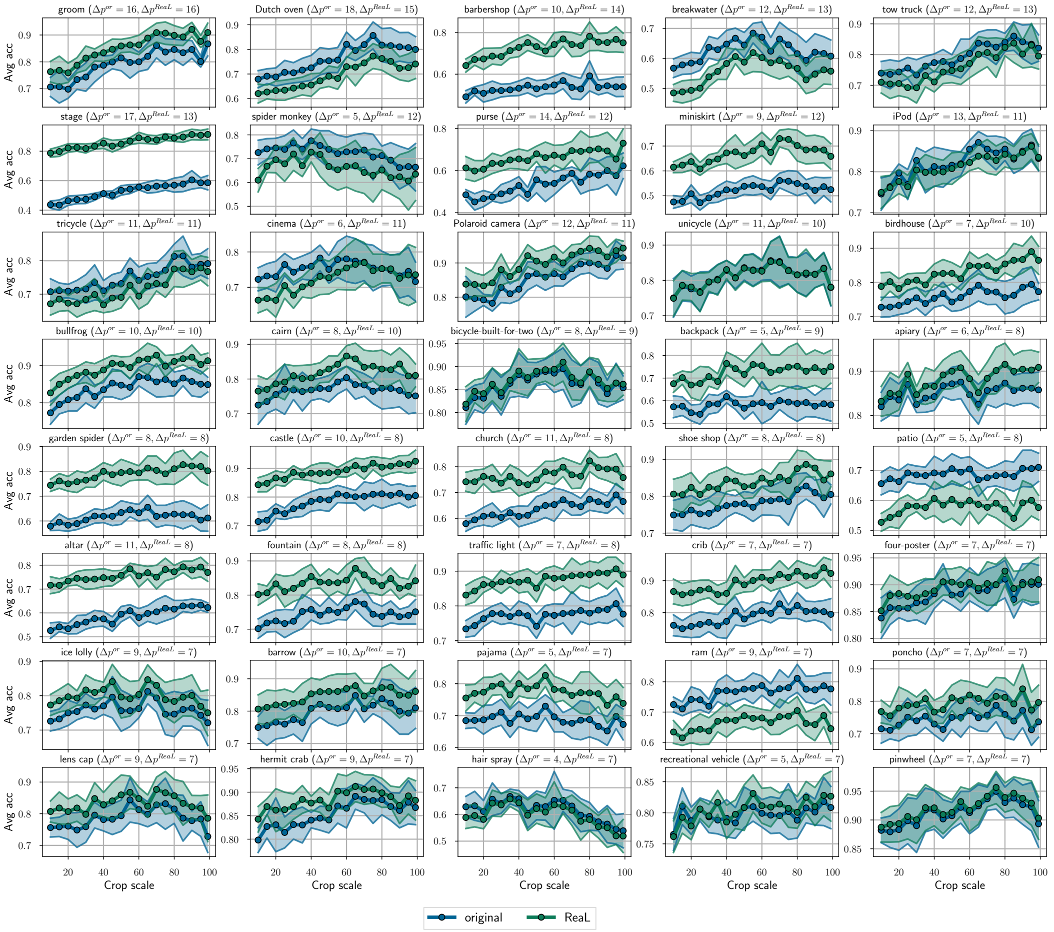


Figure 6: Individual per-class precision for classes most affected in terms of their precision evaluated with ReaL labels.