# Understanding the Detrimental Class-level Effects of Data Augmentation

**Polina Kirichenko** [1 2] **Mark Ibrahim** [2] **Randall Balestriero** [2] **Diane Bouchacourt** [2]
**Ramakrishna Vedantam** [2] **Hamed Firooz** [2] **Andrew Gordon Wilson** [1]

## Abstract

Data augmentation (DA) encodes invariance and provides implicit regularization critical to a model's performance in image classification tasks. However, while DA improves average accuracy, recent studies have shown that its impact can be highly class dependent: achieving optimal average accuracy comes at the cost of significantly hurting individual class accuracy by as much as $20\%$ on ImageNet. In this work, we present a framework for understanding how DA interacts with class-level learning dynamics. Using higher-quality multi-label annotations on ImageNet, we systematically categorize the affected classes and find that the majority are inherently ambiguous, spuriously correlated, or involve fine-grained distinctions, while DA controls the model's bias towards one of the closely related classes. While many of the previously reported performance drops are explained by multi-label annotations, our analysis of class confusions reveals other sources of accuracy degradation. We show that simple class-conditional augmentation strategies informed by our framework improve performance on the negatively affected classes.

## 1. Introduction

Data augmentation (DA) provides numerous benefits for training of deep neural networks including promoting invariance and providing regularization. In particular, DA significantly improves the generalization performance in image classification problems when measured by average accuracy (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 DA, in particular, Random Resized Crop used in training of most modern 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 (see Figure 3 left). Performance degradation even on a small set of classes might result in poor generalization on downstream tasks related to the affected classes (Salman et al., 2022), while in other applications it would be unethical to sacrifice accuracy on some classes for improvements in average accuracy (Blodgett et al., 2016; Tatman, 2017; Buolamwini & Gebru, 2018). Balestriero et al. (2022a) attempted to address class-level performance degradation by applying DA selectively to the classes where the accuracy improves with DA strength. Surprisingly, they found that this augmentation policy did not address the issue and the performance on non-augmented classes still degraded with augmentation strength. In this work we perform detailed analysis and explore the mechanisms causing the class-level performance degradation. In particular, we identify the *interactions between class-conditional data distributions* as the cause of the class-level performance degradation with augmentation: DA creates an overlap between the data distributions associated with different classes. In particular, our contributions are the following:

- We refine the analysis of class-level effects of data augmentations by correcting for label ambiguity using multi-label annotations on ImageNet. Through this analysis, we find that class-level performance degradation reported in prior works is overestimated.
- We systematically categorize the class confusions exacerbated by strong augmentation and find that many affected classes are ambiguous or co-occurring and are often affected by label noise. We focus on addressing the remaining fine-grained and non-trivial class confusions.
- We show that for addressing DA biases it is important to consider the classes with an increasing number of *false positive mistakes*, and not only the classes negatively affected in accuracy. By taking into account our observations on DA affecting class interactions, we propose a simple class-conditional data augmentation strategy that leads to improvement on the affected group of classes by $2.5\%$ on ImageNet. This improvement is

[*]Equal contribution [1]New York University [2]Meta AI. Correspondence to: Polina Kirichenko <pk1822@nyu.edu>.

*Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability*. Honolulu, Hawaii, USA. 

in contrast to the previously explored class-conditional DA in Balestriero et al. (2022a) which failed to improve class-level accuracy.

## 2. Evaluation setup and notation

We closely follow the experimental setup from Balestriero et al. (2022a). We focus on ResNet-50 models trained on ImageNet (IN) and study how average and class-level performance changes depending on the Random Resized Crop (RRC) augmentation strength: it is by far the most widely adopted augmentation which leads to significant average accuracy improvements and is used for training the state-of-the-art computer vision models. The size of the crop in RRC DA is sampled from the uniform distribution $s_{RRC} \sim U[s, 100\%]$, and by varying the lower bound $s$ we control *the strength of augmentation*. In particular, $s = 8\%$ corresponds to the strongest augmentation (which is the default value in `pytorch` RRC implementation) and $s = 100\%$ corresponds to no augmentation.

**Evaluation metrics.** Beyer et al. (2020) released Reassessed Labels (ReaL) for ImageNet validation set which partially correct the label noise present in the original labels including mislabeled examples, multi-object images and ambiguous classes. We will use $l_{ReaL}(x)$ to denote the set of ReaL labels of example $x$. We aim to measure performance of model $f_s(x)$ as a function of DA strength $s$. We measure average accuracy $a(s)$, and per-class accuracy $a_k(s)$ with respect to both original IN labels and ReaL multi-label annotations given by: $a_k^{or}(s) = 1/|X_k| \sum_{x \in X_k} I[f_s(x) = k]$ and $a_k^{ReaL} = 1/|X_k| \sum_{x \in X_k} I[f_s(x) \in l_{ReaL}(x)]$, where $X_k$ are images from class $k$ in validation set. We will refer to $a^{or}$ and $a^{ReaL}$ as *original accuracy* and *ReaL accuracy*, respectively. In our analysis, we evaluate per-class accuracy drops comparing the maximum accuracy attained on a particular class $k$ across all augmentation levels $a_k(s_k^*) = \max_s a_k(s)$ and accuracy on that class when training with strongest DA $a_k(s = 8\%)$. We will refer to the classes with the highest $\Delta a_k = a_k(s_k^*) - a_k(s = 8\%)$ as the ones *most negatively affected* by DA. To summarize performance on the affected classes, we will evaluate average accuracy of classes with the highest $\Delta a_k$ (in many cases focusing on $5\%$ of IN classes with the highest accuracy drop following Balestriero et al. (2022a)). Due to space limitation, we provide more details on the setup in Appendix A and the related works discussion in Appendix C.

## 3. Per-class accuracy degradation with strong DA is overestimated due to label ambiguity

Previous studies reported that the performance of ImageNet models is effectively better when evaluated using multi-label annotations which address its label noise issues (e.g.

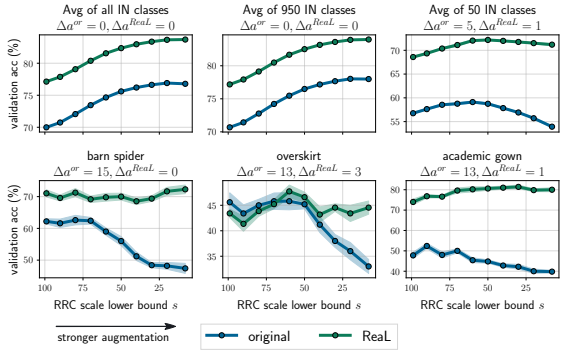


*Figure 1.* **We find that for many classes the negative effects of strong data augmentation are muted if we use high-quality multi-label annotations.** Average and per-class accuracy of ResNet-50 trained on ImageNet evaluated with original and ReaL labels against Random Resized Crop augmentation strength ($s = 8\%$ corresponds to the strongest and default augmentation). The top row shows the average accuracy of all ImageNet classes, the 50 classes with the highest original accuracy degradation and the remaining 950 classes. The bottom row shows the accuracy of 3 individual classes most significantly affected in original accuracy when using strong augmentation.

Shankar et al. (2020) and others), however, it is unclear how correcting for label ambiguity 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 many classes with severe drops in accuracy with original labels, the class-level ReaL accuracy is considerably less affected.** In Figure 4 we show the distributions of per-class accuracy drops $\Delta a_k^{or}$ and $\Delta a_k^{ReaL}$, where the distribution of $\Delta a_k^{or}$ has a much heavier tail. Using multi-label accuracy in evaluation reveals there are much fewer classes which have severe effective performance drop: e.g. only 37 classes with $\Delta a_k^{ReaL} > 4\%$ as opposed to 83 classes with $\Delta a_k^{or} > 4\%$. In Figure 1, we show how multi-label accuracy evaluation impacts the average and individual class performance across different DA strengths $s$. In the top row plots we see that while the average accuracy of all classes follows a similar trend when evaluated with either original or ReaL labels, the average accuracy of 50 most negatively affected classes only decreases by $1\%$ with ReaL labels as opposed to more significant $5\%$ drop with original labels. The bottom row shows the accuracy for "barn spider", "overskirt" and "academic gown" classes which have the highest $\Delta a_k^{or}$, and the trends for all 50 most negatively affected classes are shown in Appendix D. For many of these classes which are hurt in original accuracy by using stronger DA, the ReaL accuracy is much less affected. For example, for the class "barn spider" the original accuracy is decreased from $63\%$ to $47\%$ if we use the model trained with RRC $s = 8\%$ compared to $s = 70\%$, while the highest ReaL accuracy on this class is achieved with $s = 8\%$. However, there are still classes for which the ReaL accuracy degrades

with stronger DA, and in Appendix D we show ReaL accuracy trends against DA strength $s$ for 50 classes with the highest $\Delta a^{ReaL}$. In the next section, we aim to understand why strong DA hurts the performance on these classes.

## 4. DA most significantly affects 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 DA by analyzing the most common confusions the models make on the affected classes and how they evolve as we vary the DA strength. We focus on the 50 classes with the highest $\Delta a_k^{or}$ and 50 classes with the highest $\Delta a_k^{ReaL}$. For a pair of classes $k$ and $l$ we define the confusion rate (CR) as: $CR_{k\to l}(s) = 1/|X_k| \sum_{x\in X_k} I[f_s(x) = l]$, i.e. the ratio of examples from class $k$ misclassified as $l$. For each affected class, we identify most common confusions and track the CR against the RRC crop scale lower bound $s$. We also analyze the reverse confusion rate $CR_{l\to k}(s)$. 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 categorize the most common types of confusions on the classes which are significantly affected by DA into *ambiguous*, *co-occurring*, *fine-grained* or *semantically unrelated* (see Figure 2). We use semantic similarity and ReaL labels co-occurence as a criteria to approximately identify a confusion category for a pair of classes. We discuss each category in detail in the following paragraphs, and in Appendix E we give more details on computing the metrics to identify the confusion type and categorize the confusions of all affected classes.

**Intrinsically ambiguous classes.** Prior works (e.g. Beyer et al. (2020) and others) identified that some pairs of ImageNet classes are hardly distinguishable, e.g. "sunglasses" and "sunglass", or "monitor" and "screen". These pairs of classes generally have higher semantic similarity and higher overlap in ReaL labels. 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. In Figure 2 top left panel we show the confusion rates against the DA strength $s$ for an ambiguous pair of classes "sunglass" and "sunglasses". 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.

**Spuriously co-occurring or overlapping classes.** There is a number of classes in ImageNet which correspond to semantically different objects which often appear together, e.g. "academic gown" and "mortarboard", or "Windsor tie" and "suit". These pairs of classes have a rather high overlap in ReaL labels and their semantic similarity can vary. With RRC we may augment the sample such that only the spuriously co-occurring object, but not the main object, is left in the image, but the model would still be trained to predict the original label: e.g. we can crop just the mortarboard in an image labeled as "academic gown". 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 meaningful mistakes, not explained by label ambiguity. In Figure 2 top right panel we show how DA strength impacts model's bias towards predicting spuriously correlated "sandbar" or "seashore" classes.

*Table 1.* Class-conditional DA intervention results.

| Augmentation strategy | | Avg acc | Avg acc of 50 classes | Avg acc of 950 classes |
|---|---|---|---|---|
| Standard DA | $s = 8\%$ | $76.79_{\pm 0.03}$ | $53.93_{\pm 0.20}$ | $77.99_{\pm 0.02}$ |
| | $s = 60\%$ | $74.65_{\pm 0.03}$ | $59.11_{\pm 0.20}$ | $75.47_{\pm 0.02}$ |
| Class-cond. (Balestriero et al) | | $76.11_{\pm 0.05}$ | $43.02_{\pm 0.28}$ | $77.85_{\pm 0.04}$ |
| Our class-cond. DA | $m = 10$ | $76.70_{\pm 0.03}$ | $54.99_{\pm 0.15}$ | $77.84_{\pm 0.03}$ |
| | $m = 30$ | $76.70_{\pm 0.03}$ | $55.48_{\pm 0.23}$ | $77.82_{\pm 0.03}$ |
| | $m = 50$ | $76.68_{\pm 0.04}$ | $56.34_{\pm 0.14}$ | $77.75_{\pm 0.04}$ |

**Fine-grained categories.** There is a number of fine-grained categories in IN like "tobacco shop" and "barbershop", or "frying pan" and "wok", where objects appear in related contexts or share some visually similar features. These classes have high semantic similarity and are not significantly overlapping in ReaL labels. RRC can produce the augmented images from different categories that have visually similar features or are focused on similar backgrounds. In Figure 2 bottom left panel we show how model's confusion rates change depending on DA strength (for them only $12\%$ of confusions were corrected by ReaL labels).

**Semantically unrelated.** In the rare but most problematic cases, the stronger DA will result in the confusion of semantically unrelated classes (due to them having similar low-level features), for example, categories like "muzzle" and "sandal", or "bath towel" and "pillow". Figure 2 bottom right panel shows how confusions between unrelated classes "muzzle" and "sandal" emerge with stronger DA.

## 5. Class-conditional augmentation policy

Balestriero et al. (2022a) showed that a naive class-conditional DA approach is not sufficient for removing the negative effects of DA: they evaluated a DA strategy where augmentation is applied to all classes except the ones with degraded accuracy which are instead processed with Center Crop. Since this approach didn't recover the accuracy of the affected classes, they hypothesize DA induces a general invariance or an implicit bias that still negatively affects classes that are not augmented. In contrast, we explore a simple class-conditional augmentation strategy based on our insights regarding the class confusions, and by changing the augmentation strength for as few as 1 to $5\%$ of classes, we observe substantial improvements on the negatively affects

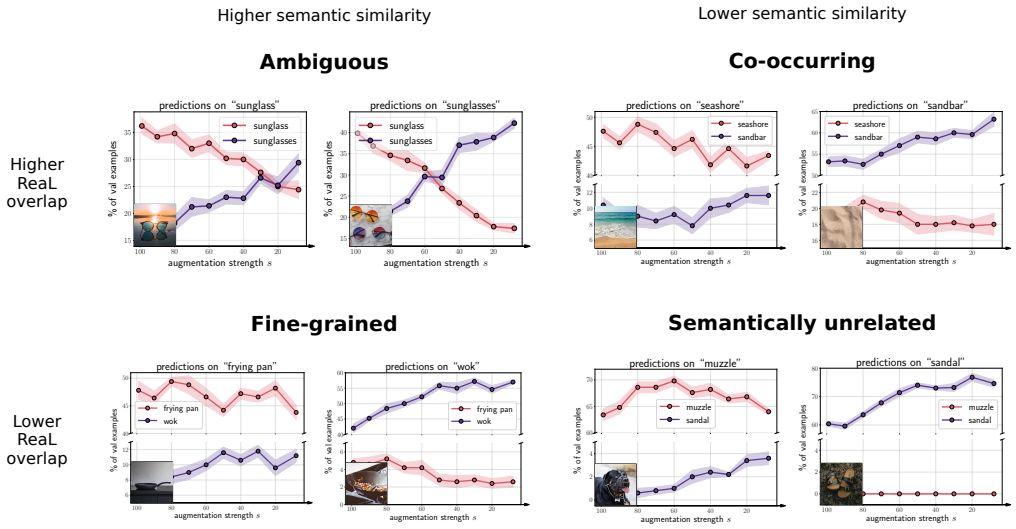


*Figure 2.* Each panel shows a pair of confused classes which we categorize into: *ambiguous*, *co-occurring*, *fine-grained* and *semantically unrelated*, depending on the inherent class overlap and semantic similarity. For each confused class pair, the left subplot corresponds to the class $k$ affected in accuracy by strong data augmentation (DA), e.g. "sunglass" on top left panel: the ratio of validation samples from that class that get classified as $k$ decreases with stronger DA, while the confusion rate with another class $l$ (e.g. class "sunglasses" on top left panel) increases. The right subplot shows the percent of examples from class $l$ that get classified as $k$ or $l$ against DA strength.

classed. We found in many cases that a class $k$ whose accuracy is affected by DA is misclassified as a related class $l$ with stronger augmentations. We can precisely describe these confusions in terms of *False Negative* (FN) mistakes for class $k$ (not recognizing an instance from class $k$) and *False Positive* (FP) mistakes for class $l$ (misclassifying an instance from another class as class $l$). We argue that to address the degraded accuracy of class $k$ it is also important to consider DA effect on class $l$. In Appendix F, we show the class-level False Positive mistake numbers compared to DA strength: these classes are often semantically related to and confused with the ones affected in accuracy, e.g. "barbershop" is confused with "tobacco shop". We explore a simple DA policy informed by the following observations: (1) generally stronger DA is helpful for the majority of classes and leads to learning more diverse features, (2) a substantially increased number of FP mistakes for a particular class likely indicates that its augmented data distribution overlaps with other classes and it might negatively affect their accuracy. Thus, by default we set the strongest data augmentation value $s = 8\%$ for the majority of classes, and change augmentation level for a small subset of classes for which FP mistakes grew the most with stronger DA. However, completely removing augmentations would hurt accuracy so we balance the tradeoff between learning diverse features and avoiding class confusions. As a heuristic, we set DA strength for each class to be $\arg\min$ of FP + FN mistakes of that class across DA levels. We vary the number of classes $m$ for which we change augmentations in the range $\{10, 30, 50\}$. We compare this intervention to the baseline model trained with the strongest DA $s = 8\%$, mild DA level $s = 60\%$ optimal for average accuracy on the affected set of classes, and the class-conditional augmentation approach studied in Balestriero et al. (2022a) where we remove augmentation from the negatively affected classes. The results are shown in Table 1. We find existing approaches sacrifice accuracy on the subset of negatively affected classes for overall average accuracy or vice versa. For example, as we previously observed the default model trained with $s = 8\%$ achieves high average accuracy on the majority of classes but suboptimal accuracy on the 50 classes affected by strong augmentation. Removing augmentation from the negatively affected classes only exacerbates the effect and decreases the accuracy both on the affected set and on average. At the same time, tuning down augmentation level on 1 to $5\%$ of classes with the highest FP mistakes increase improves the accuracy on the affected classes by $2.5\%$ for $m = 50$, and taking into account the tradeoff between False Positive and False Negative mistakes helps to maintain high average accuracy overall and on majority of classes. These results support our hypothesis and demonstrate how a simple intervention on a small number of classes informed by the appropriate metrics can substantially improve performance.

**Discussion.** In this work we provide new insights into the class-level accuracy degradation on ImageNet using standard augmentation. We show that to understand DA biases it is important to consider the interactions among class-conditional data distributions, and how DA affects these interactions. We systematically categorize the most significantly affected classes as ambiguous, co-occurring,

or involving fine-grained distinctions. In contrast to prior work, we show that a simple class-conditional DA policy based on our insights can significantly improve performance on the classes negatively affected by standard DA.

## References


Balestriero, R., Bottou, L., and LeCun, Y. The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*, 2022a.

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

Benton, G., Finzi, M., Izmailov, P., and Wilson, A. G. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33: 17605–17616, 2020.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Bitterwolf, J., Meinke, A., Boreiko, V., and Hein, M. Classifiers should do well even on their worst classes. In *ICML 2022 Shift Happens Workshop*, 2022.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

Botev, A., Bauer, M., and De, S. Regularising for invariance to data augmentation improves supervised learning. *arXiv preprint arXiv:2203.03304*, 2022.

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

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Cheung, T.-H. and Yeung, D.-Y. Adaaug: Learning class- and instance-adaptive data augmentation policies. In *International Conference on Learning Representations*, 2022.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fellbaum, C. *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.

Fujii, S., Ishii, Y., Kozuka, K., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. Data augmentation by selecting mixed classes considering distance between classes. *arXiv preprint arXiv:2209.05122*, 2022.

Geiping, J., Goldblum, M., Somepalli, G., Shwartz-Ziv, R., Goldstein, T., and Wilson, A. G. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. *arXiv preprint arXiv:2210.06441*, 2022.

Gontijo-Lopes, R., Smullin, S. J., Cubuk, E. D., and Dyer, E. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 1–16. Springer, 2020.

Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J., and Hansen, L. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial intelligence and statistics*, pp. 342–350. PMLR, 2016.

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

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

Hernández-García, A. and König, P. 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.

Ho, D., Liang, E., Chen, X., Stoica, I., and Abbeel, P. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning*, pp. 2731–2741. PMLR, 2019.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303.

Hooker, S., Courville, A., Clark, G., Dauphin, Y., and Frome, A. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Hu, M. and Li, J. Exploring bias in gan-based data augmentation for small samples. *arXiv preprint arXiv:1905.08495*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Idrissi, B. Y., Bouchacourt, D., Balestriero, R., Evtimov, I., Hazirbas, C., Ballas, N., Vincent, P., Drozdzal, M., Lopez-Paz, D., and Ibrahim, M. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*, 2022.

Kaplun, G., Ghosh, N., Garg, S., Barak, B., and Nakkiran, P. Deconstructing distributions: A pointwise framework of learning. *arXiv preprint arXiv:2202.09931*, 2022.

Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.

Leclerc, G., Ilyas, A., Engstrom, L., Park, S. M., Salman, H., and Madry, A. FFCV: Accelerating training by removing data bottlenecks. https://github.com/libffcv/ffcv/, 2022. commit xxxxxxx.

Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N. M., and Yang, Y. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

Lin, C.-H., Kaushik, C., Dyer, E. L., and Muthukumar, V. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *arXiv preprint arXiv:2210.05021*, 2022.

Luccioni, A. S. and Rolnick, D. Bugs in the data: How imagenet misrepresents biodiversity. *arXiv preprint arXiv:2208.11695*, 2022.

Mahan, S., Kvinge, H., and Doster, T. Rotating spiders and reflecting dogs: a class conditional approach to learning data augmentation distributions. *arXiv preprint arXiv:2106.04009*, 2021.

Miao, N., Mathieu, E., Dubois, Y., Rainforth, T., Teh, Y. W., Foster, A., and Kim, H. Learning instance-specific data augmentations. *arXiv preprint arXiv:2206.00051*, 2022.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Müller, S. G. and Hutter, F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 774–782, 2021.

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

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

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance.pdf.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Ratner, A. J., Ehrenberg, H., Hussain, Z., Dunnmon, J., and Ré, C. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30, 2017.

Rey-Area, M., Guirado, E., Tabik, S., and Ruiz-Hidalgo, J. Fucitnet: Improving the generalization of deep learning networks by the fusion of learned class-inherent transformations. *Information Fusion*, 63:188–195, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Salman, H., Jain, S., Ilyas, A., Engstrom, L., Wong, E., and Madry, A. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*, 2022.

Shah, H., Park, S. M., Ilyas, A., and Madry, A. Modeldiff: A framework for comparing learning algorithms. *arXiv preprint arXiv:2211.12491*, 2022.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pp. 8634–8644. PMLR, 2020.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tang, Z., Peng, X., Li, T., Zhu, Y., and Metaxas, D. N. Adatransform: Adaptive data transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3006, 2019.

Tatman, R. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59, 2017.

Teney, D., Lin, Y., Oh, S. J., and Abbasnejad, E. Id and ood performance are sometimes inversely correlated on real-world datasets. *arXiv preprint arXiv:2209.00613*, 2022.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pp. 9625–9635. PMLR, 2020.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. 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.

Vasudevan, V., Caine, B., Gontijo-Lopes, R., Fridovich-Keil, S., and Roelofs, R. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *arXiv preprint arXiv:2205.04596*, 2022.

Xu, M., Yoon, S., Fuentes, A., and Park, D. S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, pp. 109347, 2023.

Xu, Y., Noy, A., Lin, M., Qian, Q., Li, H., and Jin, R. Wemix: How to better utilize data augmentation. *arXiv preprint arXiv:2010.01267*, 2020.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., and Chun, S. 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.

Zheng, Y., Zhang, Z., Yan, S., and Zhang, M. Deep autoaugment. *arXiv preprint arXiv:2203.06172*, 2022.

Zhou, F., Li, J., Xie, C., Chen, F., Hong, L., Sun, R., and Li, Z. Metaaugment: Sample-aware data augmentation policy learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11097–11105, 2021.

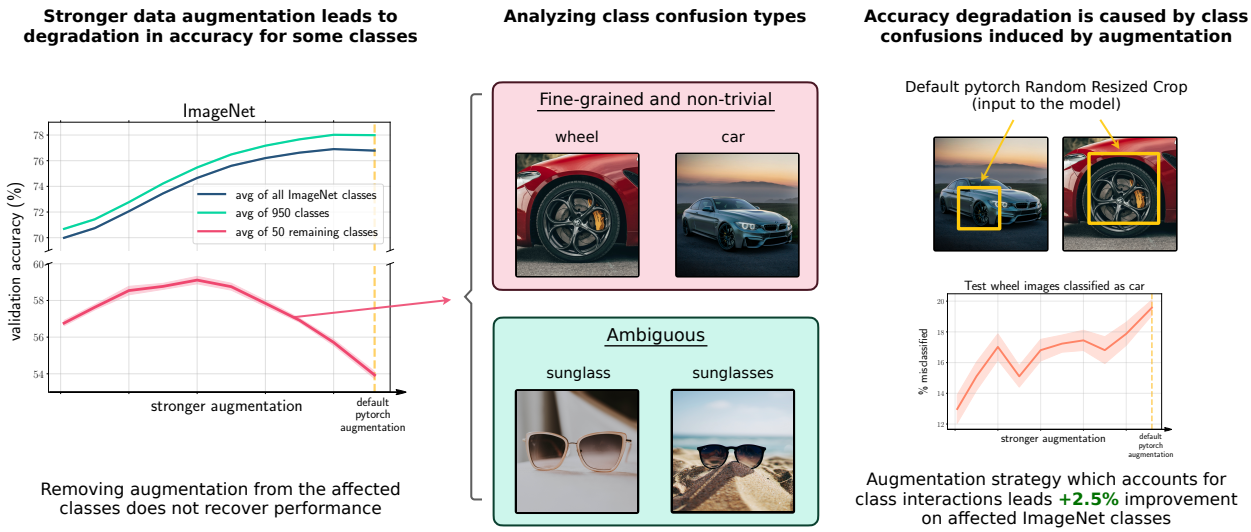


*Figure 3.* **We show that the classes negatively affected by data augmentation are often ambiguous, co-occurring or fine-grained categories and analyze how data augmentation exacerbates class confusions. Left:** Average accuracy of ResNet-50 on ImageNet against Random Resized Crop (RRC) data augmentation strength: average of all classes (blue), average of the 50 classes on which stronger RRC hurts accuracy the most (red), and the average of the remaining 950 classes (green). Yellow line indicates the default RRC setting used in training of most computer vision models. **Middle:** We systematically categorize the types of class confusions exacerbated by strong data augmentation: while some of them include ambiguous or correlated classes, there is a number of fine-grained and non-trivial confusions. **Right:** Often the class-level accuracy drops due to overlap with other classes after applying augmentation: e.g. heavily augmented samples from "car" class can look like typical images from "wheel" class. As a result, the model learns to predict "car" on "wheel" images, and the accuracy on the "wheel" class drops. To resolve the negative effect of strong augmentation on classes like "wheel", we should modify augmentation strength of classes like "car".

## Contribution summary

In this work we perform detailed analysis and explore the mechanisms causing the class-level performance degradation emerging with strong data augmentation (DA). In particular, we identify the *interactions between class-conditional data distributions* as the cause of the class-level accuracy drops: DA creates an overlap between the data distributions associated with different classes. As a simple example, in Figure 3 (right) we show that the standard Random Resized Crop operation creates an overlap between the "car" and "wheel" classes. As a result, the model learns to predict "car" on "wheel" images, and the performance on the "wheel" class drops. Importantly, if we want to improve the performance on the "wheel" class, we need to modify the augmentation policy on the class "car" and not "wheel" as was done in prior work (Balestriero et al., 2022a). We also identify the types of mistakes models make on the affected classes, and explain why the selective DA policy of Balestriero et al. (2022a) fails to improve class-level performance. Finally, we show that a simple class-conditional DA policy motivated by our analysis improves performance on the classes that are negatively affected by standard augmentation. We summarize our findings in Figure 3

## A. Setup details

Following Balestriero et al. (2022a), we train ResNet-50 models (He et al., 2016) on ImageNet (Russakovsky et al., 2015) for 88 epochs with SGD with momentum 0.9, using batch size 1024, weight decay $10^{-4}$, and label smoothing 0.1 (Szegedy et al., 2016). We use cyclic learning rate schedule starting from the initial learning rate $10^{-4}$ with the peak value 1 after 2 epochs and linearly decaying to 0 until the end of training. We use `PyTorch` (Paszke et al., 2017), automatic mixed precision training with `torch.amp` package[1], `ffcv` package (Leclerc et al., 2022) for fast data loading. We use image resolution 176 during training, and resolution 224 during evaluation, following Balestriero et al. (2022a), Touvron et al.

[1] https://pytorch.org/docs/stable/amp.html

(2019) and `torchvision` training recipe[2]. Balestriero et al. (2022a) also use different image resolution at training and test time: ramping up resolution from 160 to 192 during training and evaluating models on images with resolution 256. We train 10 independent models with different random seeds for each augmentation strength $s \in \{8, 20, 30, 40, 50, 60, 70, 80, 90, 99\}$ where $s = 8\%$ corresponds to the strongest and default augmentation.

**Data augmentation.** We apply random horizontal flips and Random Resized Crop (RRC) DA when training our models. In particular, for an input image of size $h \times w$ the RRC transformation samples the crop scale $s_{RRC} \sim U[s_{low}, s_{up}]$ and the aspect ratio $r \sim U[r_{low}, r_{up}]$, where $U[a, b]$ denotes a uniform distribution between $a$ and $b$. RRC then takes random a crop of size $\sqrt{s_{RRC}hwr} \times \sqrt{s_{RRC}hw/r}$ and resizes it to a chosen resolution $R \times R$. We use the standard values for $s_{up} = 100\%$ and aspect ratios $r_{low} = 3/4, r_{up} = 4/3$, and vary the lower bound of the crop scale $s_{low}$ (for simplicity, we will further use $s$) between $8\%$ and $100\%$ which controls *the strength of augmentation*: $s = 8\%$ corresponds to the strongest augmentation (note this is the default value in `pytorch` (Paszke et al., 2019) RRC implementation) and $s = 100\%$ corresponds no cropping hence no augmentation.

**ReaL labels.** Beyer et al. (2020) used large-scale vision models to generate new label proposals for ImageNet validation set which were then evaluated by human annotators. These Reassessed Labels (ReaL) 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 image, model's prediction is considered correct if it matches one of the plausible labels.

We use `NLTK` library (Bird et al., 2009) for WordNet and `spaCy` library (Honnibal et al., 2020) for embeddings similarity. Example images in Figures 2 and 3 are taken from `https://unsplash.com/`.

## B. Evaluation metrics

To understand the biases introduced or exacerbated by data augmentation, we use a number of fine-grained metrics and evaluate them for models trained with different augmentation levels. We compute these metrics using original ImageNet validation labels and ReaL multi-label annotations (Beyer et al., 2020). We use $f_s(\cdot)$ to denote a neural network trained with augmentation parameter $s$, $l_{ReaL}(x)$ a set of ReaL labels for a validation example $x$, $X$ a set of all validation images, $X_k$ the validation examples with the original label $k$.

**Accuracy.** The average accuracy across for original and ReaL labels is defined as:

$$a^{or}(s) = 1/|X| \sum_{x \in X} I[f_s(x) = k] \quad \text{and} \quad a^{ReaL} = 1/|X| \sum_{x \in X} I[f_s(x) \in l_{ReaL}(x)],$$

while for per-class accuracies $a_k^{or}(s)$ and $a_k^{ReaL}(s)$ the summation is over the set $X_k$ instead of all validation examples $X$. The accuracy on class $k$ with original labels $a_k^{or}(s)$ also correspond to *recall* of the model on that class.

**Confusion.** In Section 4 we looked at class confusions, in particular for a pair of classes $k$ and $l$ the confusion rate (CR) is defined as:

$$CR_{k \to l}(s) = 1/|X_k| \sum_{x \in X_k} I[f_s(x) = l],$$

i.e. the ratio of examples from class $k$ misclassified as $l$. We are only discussing confusions $CR_{k \to l}$ in the context of original labels.

**False Positive and False Negative mistakes.** In Section 5, we emphasized the importance of looking at how data augmentation impacts not only per-class accuracy but also the number of *False Positive* (FP) mistakes for a particular class:

$$FP_k^{or}(s) = \sum_{(x \in X) \cap (x \notin X_k)} I[f_s(x) = k] \quad \text{and} \quad FP_k^{ReaL}(s) = \sum_{(x \in X) \cap (k \notin l_{ReaL}(x))} I[f_s(x) = k]$$

for original and Real labels respectively. The number of *False Negative* mistakes on class $k$ in terms of the original labels are directly related to the accuracy, or recall, on that class:

[2] `https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/`

$$FN_k^{or}(s) = \sum_{x \in X_k} I[f_s(x) \neq k] = |X_k|(1 - a^{or}(s)),$$

while for multi-label annotations we define it as:

$$FN_k^{ReaL}(s) = \sum_{(x \in X) \cap (k \in l_{ReaL}(x))} I[f_s(x) \notin l_{ReaL}(x)],$$

i.e. the number of examples $x$ which were misclassidied by the model where $k$ was in the ReaL label set $l_{ReaL}(x)$. In Section 5 we explored $s_k^* = \arg\min FN_k(s) + FN_k(s)$ as a proxy for optimal class-conditional augmentation level which emphasizes the inherent tradeoff between class-level accuracy and the number of False Positive mistakes.

**Affected classes.** We are focusing on analyzing model's behavior on the classes which were negatively affected by strong (default) augmentation in terms of original or ReaL accuracy, i.e. classes where the accuracy drop $\Delta a_k = a_k(s_k^*) - a_k(s = 8\%)$ from $a_k(s_k^*) = \max_s a_k(s)$ to $a_k(s = 8\%)$ is the highest. We focus on $5\%$ of classes (50 classes) with the highest $\Delta a_k$ following Balestriero et al. (2022a) and measure the average accuracy on this set of classes as a function of $s$ and after interventions in Section 5.

In Section 5, we also look at classes where the number of FP mistakes increased the most with strong DA, i.e. with the highest $\Delta FP_k = FP_k(s = 8\%) - FP_k(s_k^*)$ where $FP_k(s_k^*) = \min_s FP_k(s)$.

**Prior work evaluation.** To quantify the class accuracy drops, Balestriero et al. (2022a) compare the per-class accuracy of models trained with the strongest DA ($s = 8\%$) and models trained without augmentation ($s = 100\%$ which effectively just resizes input images without cropping), while Bouchacourt et al. (2021) compared class-level accuracy of models trained with RRC with $s = 8\%$ and models trained with fixed size Center Crop.

## C. 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 models 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.

**Adaptive and learnable data augmentation.** Xu et al. (2020) showed that data augmentation may exacerbate data bias which may lead to model' suboptimal performance on the original data distribution. They propose to train the model on a mix of augmented and unaugmented samples and then fine-tune it on unaugmented data after training which showed improved performance on CIFAR dataset. Raghunathan et al. (2020) showed standard error in linear regression could increase when training with original data and data augmentation, even when data augmentation is label-preserving. Rey-Area et al. (2020) and Ratner et al. (2017) learn DA transformation using GAN framework, while Hu & Li (2019) study the bias of GAN-learned data augmentation. Fujii et al. (2022) take into account the distances between classes to adapt mixed-sample

DA. Hauberg et al. (2016) learn class-specific DA on MNIST. Numerous works, e.g. Cubuk et al. (2018); Lim et al. (2019); Ho et al. (2019); Hataya et al. (2020); Li et al. (2020); Cubuk et al. (2020); Tang et al. (2019); Müller & Hutter (2021) and Zheng et al. (2022) find dataset-dependent augmentation strategies. Benton et al. (2020) proposed Augerino framework to learn augmentation form training data. Zhou et al. (2021); Cheung & Yeung (2022); Mahan et al. (2021) and Miao et al. (2022) learn class- or input-dependent augmentation policies. Yao et al. (2022) propose to modify mixed-sample augmentation to improve out-of-domain generalization.

**Robustness and model evaluation beyond average accuracy.** While Miller et al. (2021) showed that model's average accuracy is strongly correlated with its out-of-distribution performance, there have been a number of works that showed that only evaluating average performance can be deceptive. Teney et al. (2022) showed counter-examples for "accuracy-on-the-line" phenomenon. Kaplun et al. (2022) show that while model's average accuracy improves during training, it may decrease on a subset of examples. Sagawa et al. (2019) show that training with Empirical Risk Minimization may lead to suboptimal performance in the worst case. Bitterwolf et al. (2022) evaluated ImageNet models' performance in terms of a number of metrics beyond average accuracy, including worst-class accuracy and precision.

**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 set using a large-scale computer vision model. Beyer et al. (2020) provide re-assessed (ReaL) multi-label annotations for ImageNet validation set 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.

## D. Accuracy of the classes most negatively affected by data augmentation

We show the per-class accuracies as a function of data augmentation strength $s$ for (1) the 50 classes most negatively affected in original accuracy, i.e. with the highest $\Delta a_k^{or}$ in Figure 5, and (2) 50 classes most negatively affected in ReaL accuracy in Figure 6.

In Figure 4 we show the distributions of per-class accuracy drops $\Delta a_k^{or}$ and $\Delta a_k^{ReaL}$. Using multi-label accuracy in evaluation reveals there are much fewer classes which have severe effective performance drop: e.g. only 37 classes with $\Delta a_k^{ReaL} > 4\%$ as opposed to 83 classes with $\Delta a_k^{or} > 4\%$. moreover, there are no classes with $\Delta a_k^{ReaL} > 11\%$.

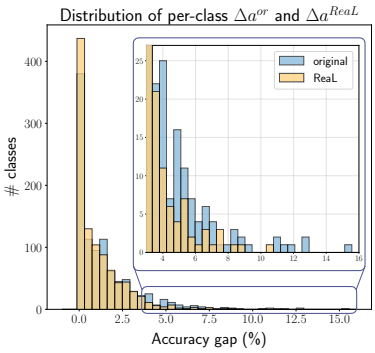


*Figure 4.* **We find that for many classes the negative effects of strong data augmentation are muted if we use high-quality multi-label annotations.** Distribution of per-class accuracy drops $\Delta a_k$ for original and ReaL labels. The distribution of $\Delta a_k^{or}$ has a heavier tail compared to the one computed with ReaL labels.

## E. Class confusion types

We consider the classes most affected by strong DA (see Figures in Appendix D) 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 the 50 classes with the highest $\Delta a_k^{or}$ (corresponding to $\Delta a_k^{or} > 5\%$), and 50 classes with the highest $\Delta a_k^{ReaL}$. (corresponding to $\Delta a_k^{ReaL} > 4\%$).

We roughly outline the most common types of confusions on the classes which are significantly affected by DA. The different types of confusion differ in in the extent to which the accuracy degradation can be attributed to label noise versus the presence of DA. We also characterize how DA effectively changes the data distribution of these classes leading to changes in 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 DA. We use *semantic similarity* and *ReaL labels co-occurence* as a criteria to identify a confusion category for a pair of classes. We can measure semantic similarity by (a) WordNet class similarity, given by the Wu-Palmer score which relies on the categories' most specific common ancestor in the WordNet tree, and (b) similarity of the class name embeddings.

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*. Examples of how confusion rates for each class confusion category change with DA strength are shown in Figures 2 and 7, and categorization of confusions on all affected classes is in Table 2.

**Class-conditional distributions induced by DA.** To aid our understanding of the class-specific effects of DA, it is helpful to reason about how a parametrized class of DA transformations $\mathcal{T}_s(\cdot)$ changes the distributions of each class in the training data $p_k(x)$. We denote the augmented class distributions by $\mathcal{T}_s(p_k)$. In particular, if supports of $\mathcal{T}_s(p_k)$ and $\mathcal{T}_s(p_l)$ for two classes $k$ and $l$ overlap, the model is trained 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 the ImageNet dataset, while others have distinct supports, but in all cases the parameters of DA $s$ will control the overlap of the induced class distributions, and thus the biases of the model when making predictions on such classes.

**Intrinsically ambiguous or semantically 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. 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 $p_k$ and $p_l$ highly overlap or even coincide, but with varying $\alpha$ depending on how the supports of $\mathcal{T}_\alpha(p_k)$ and $\mathcal{T}_\alpha(p_l)$ overlap the model would be biased towards predicting one of the classes. In Figure 2 top left panel 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 $C_{kl} = 91.1\%$ 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 DA 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.

**Co-occurring or overlapping classes.** There is a number of classes in 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", "seashore" and "sandbar". 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 in the image, but the model would still be trained to predict the original label: we can crop just the mortarboard in an image labeled as "academic gown". 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 meaningful mistakes, not explained by label ambiguity. In Figure 2 top right panel we show how DA strength impacts model's bias towards predicting "sandbar" or

"seashore" class (for which $C_{kl} = 72\%$ and $96\%$ confusions resolved by ReaL labels).

We emphasize that unlike the ambiguous classes discussed ealier, the co-occuring classes cause meaningful mistakes on the test data, which are not resolved by multi-label annotations. For example, the model will be biased to predict "academic gown" even when shown an image of just the mortarboard.

**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 $\mathcal{T}(p_k)$ and $\mathcal{T}(p_l)$ 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 bottom left panel we show how model's most common predictions change depending on RRC crop scale for fine-grained classes "frying pan" and "wok" (for which $C_{kl} = 10\%$, 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 $p_k$ and $p_l$ 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 bottom right panel shows how confusions between unrelated classes "muzzle" and "sandal" 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.

In Table 2 we show the classes most negatively affected in accuracy by strong data augmentation (column "Affected class $k$") and the confusions the model starts making more frequently with stronger augmentation ("Confused class $l$"). In particular, we study the union of $50$ classes most affected in original accuracy and $50$ classes most affected in ReaL accuracy (see Section D) which do not belong to the animal subtree in WordNet tree. We focus on the confusions $l$ where confusion rate difference

$$\Delta CR_{k\to l} = CR_{k\to l}(s = 8\%) - \min_s CR_{k\to l}(s)$$

is the highest for class $k$ and above $2.5\%$ (see Section B for definition of confusion rate $CR_{k\to l}(s)$). Additionally for each pair of confused classes $k$ and $l$ we also look at

$$\Delta CR^*_{l\to k} = \max_s CR_{l\to k}(s) - CR_{l\to k}(s = 8\%)$$

which characterizes to what extent the model trained with weaker augmentation starts making the reverse confusion more often compared to the strong DA model.

To quantitatively estimate the confusion type for each pair of classes, we measure the intrinsic distribution overlap of the classes and their semantic similarity. We compute one sided overlap for classes $k$ and $l$, which is the ratio of examples that have both labels $k$ and $l$ among the examples with the label $k$:

$$C_{kl} = \sum_{x\in X} I[k \in l_{ReaL}(x)] \times I[l \in l_{ReaL}(x)] / \sum_{x\in X} I[k \in l_{ReaL}(x)]$$

and intersection-over-union of the two classes:

$$IoU_{kl} = \sum_{x\in X} I[k \in l_{ReaL}(x)] \times I[l \in l_{ReaL}(x)] / \sum_{x\in X} I[k \in l_{ReaL}(x) \text{ or } l \in l_{ReaL}(x)].$$

Assuming that train and test are coming from similar distributions, we can treat $C_{kl}$ as a measure of overlap between distributions $p_k$ and $p_l$. We use WordNet class similarity and similarity of word embeddings from spacy (Honnibal et al., 2020) to measure semantic similarity. Note that these metrics only serve as approximate measures of distribution overlap and semantic distance since (1) the ReaL labels still contain some amount of label noise and may contain mislabelled examples

or examples that are missing some of the plausible labels, (2) the WordNet distance sometimes is low for classes that are semantically very similar, and (3) spacy doesn't have a representation for all words and is underestimating the similarity of closely related concepts. However, all together these metrics can point towards one of the appropriate confusion type categories.

In Figure 7 we show more examples of the confusion rates for different pairs of classes $k$ and $l$ as a function of data augmentation strength $s$ where $k$ is among the ones most negatively affected in accuracy and $l$ is the class the model misclassified examples from the class $k$ to. We show example pairs from different confusion types defined in Section 4.

## F. Class-conditional augmentation intervention experiments

In Figures 8 and 9 we show how the number of False Positive (FP) mistakes changes with data augmentation strength for the set of classes where FP number increased the most with strong DA (see Figure 8 for the set of classes where original FP mistakes increased the most and Figure 9 for ReaL FP mistakes). In Section 5, we conducted class-conditional data augmentation interventions changing the DA strength for these sets of classes and showed that it improved the accuracy on the classes negatively affected in accuracy.

While in Section 5 we show results for adapting augmentation level for classes using original labels to evaluate False Positive and False Negative mistakes, in Table 3 we show analogous results when using ReaL labels which also shows that this targeted intervention into augmentation policy for a small number of classes leads to improvement in ReaL average accuracy on the affected classes (we specifically consider the set of classes affected in ReaL accuracy).

## G. Broader impact and limitations

**Limitations.** In this paper we consider the impact of Random Resized Crop (RRC) data augmentation which is the most commonly used augmentation transformation which is also often used in combination with other automatic augmentation policies (Cubuk et al., 2018; Müller & Hutter, 2021). RRC DA also leads to most substantial improvements in average accuracy, unlike other transformations such as color-based augmentation which usually leads to limited improvements. For the main analysis we focus on ResNet-50 architecture, however, Balestriero et al. (2022a) showed that per-class biases persist in other architectures like Vision Transformers (Dosovitskiy et al., 2020) and DenseNets (Huang et al., 2017) and for colorjitter augmentation. While we provide a deep analysis of RRC per-class effects in ResNet models, the same framework can be extended to better understand the biases of other augmentations and other architectures in the future work.

As discussed in Section E while we provide quantitative metrics to describe each confusion type affected by data augmentation, the categorization is not strict due to the remaining noise in ReaL labels and imprecise word similarity metrics.

**Broader impact.** A potential negative outcome that can result from misinterpretation of our analysis in Section 3 is if the practitioners assume that data augmentation does not have any negative effects since we discover that previously reported performance drops were overestimated due to label noise. We emphasize that while some of the class-level accuracy drops were indeed due to label ambiguity or co-occurring objects, data augmentation does exacerbate model's bias and introduces class confusions (often between fine-grained categories but sometimes even for semantically unrelated classes that share visually similar features). We encourage researchers to carefully study the negative impact of DA using fine-grained metrics beyond average accuracy (such as per-class accuracy, False Positive mistakes and class confusions) to better understand its biases.

**Practical recommendations.** When evaluating model performance, one should not only check average accuracy, which may conceal class-level learning dynamics. Instead, we recommend researchers also consider other metrics such as False Positive rates to better detect which confusions DA introduces or exacerbates. In particular, when training a model with strong augmentations, one should train another model with weaker augmentations to check whether finer-grained metrics such as FP rates degraded as an indicator DA is biasing learning dynamics. We can then design targeted augmentation policies to improve performance on the groups negatively affected by standard augmentations.

**Compute.** We estimate the total compute used in the process of working on this paper at roughly $5000$ GPU hours. The compute usage is dominated by training models for different augmentation strengths (Section 3). The experiments were run on GPU clusters on Nvidia Tesla V100, Titan RTX, RTX8000, 3080 and 1080Ti GPUs.

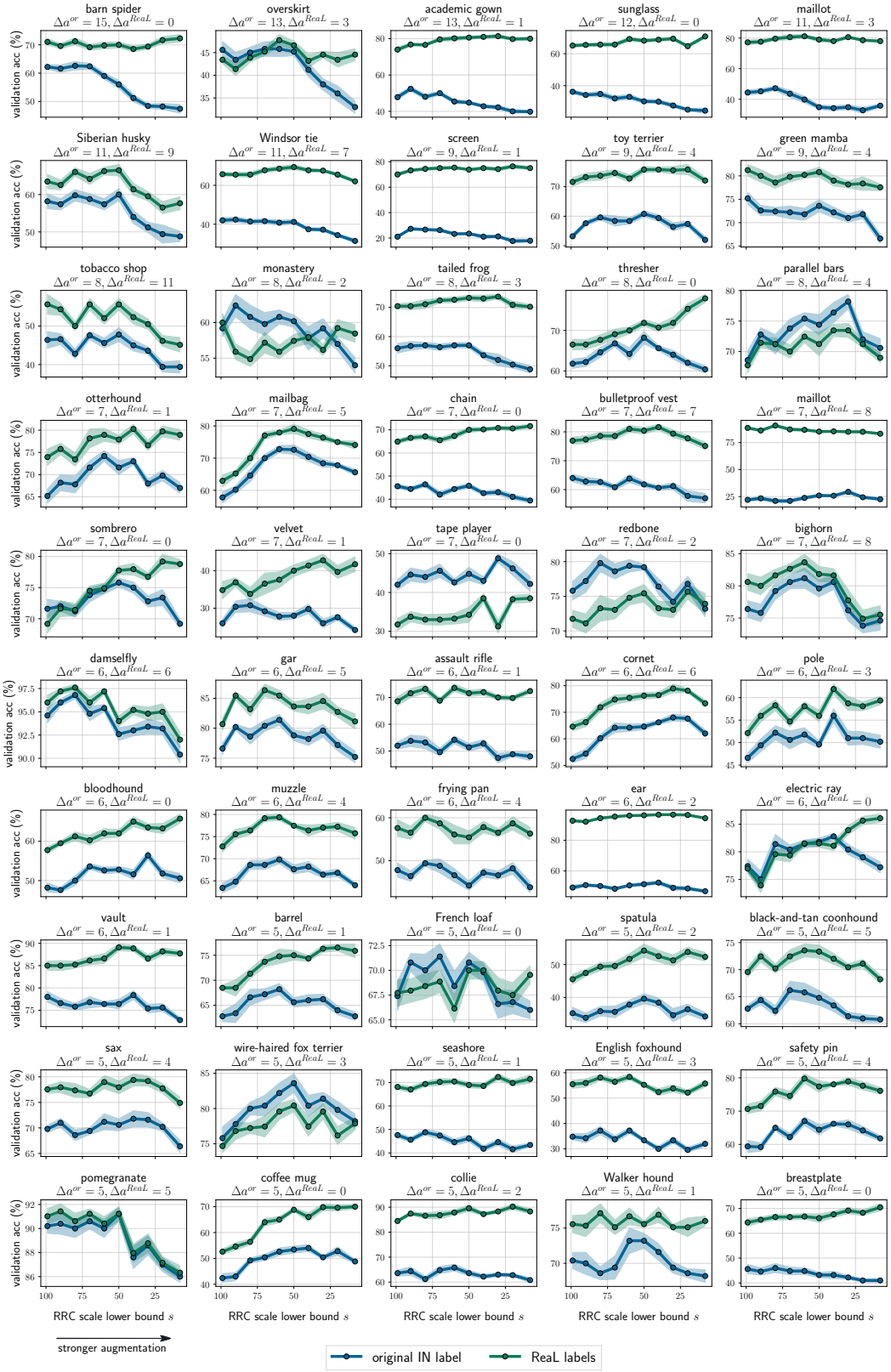


*Figure 5.* Per-class class validation accuracies of ResNet-50 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound $s$. We show the accuracy trends for the classes with the highest difference between the maximum accuracy on that class across augmentation levels $\max_s a_k^{or}(s)$ and the accuracy of the model trained with $s = 8\%$. On each subplot below the name of the class we show the accuracy drops with respect to original and ReaL labels: $\Delta a_k^{or}$ and $\Delta a_k^{ReaL}$. We report the mean and standard error over 10 independent runs of the network.

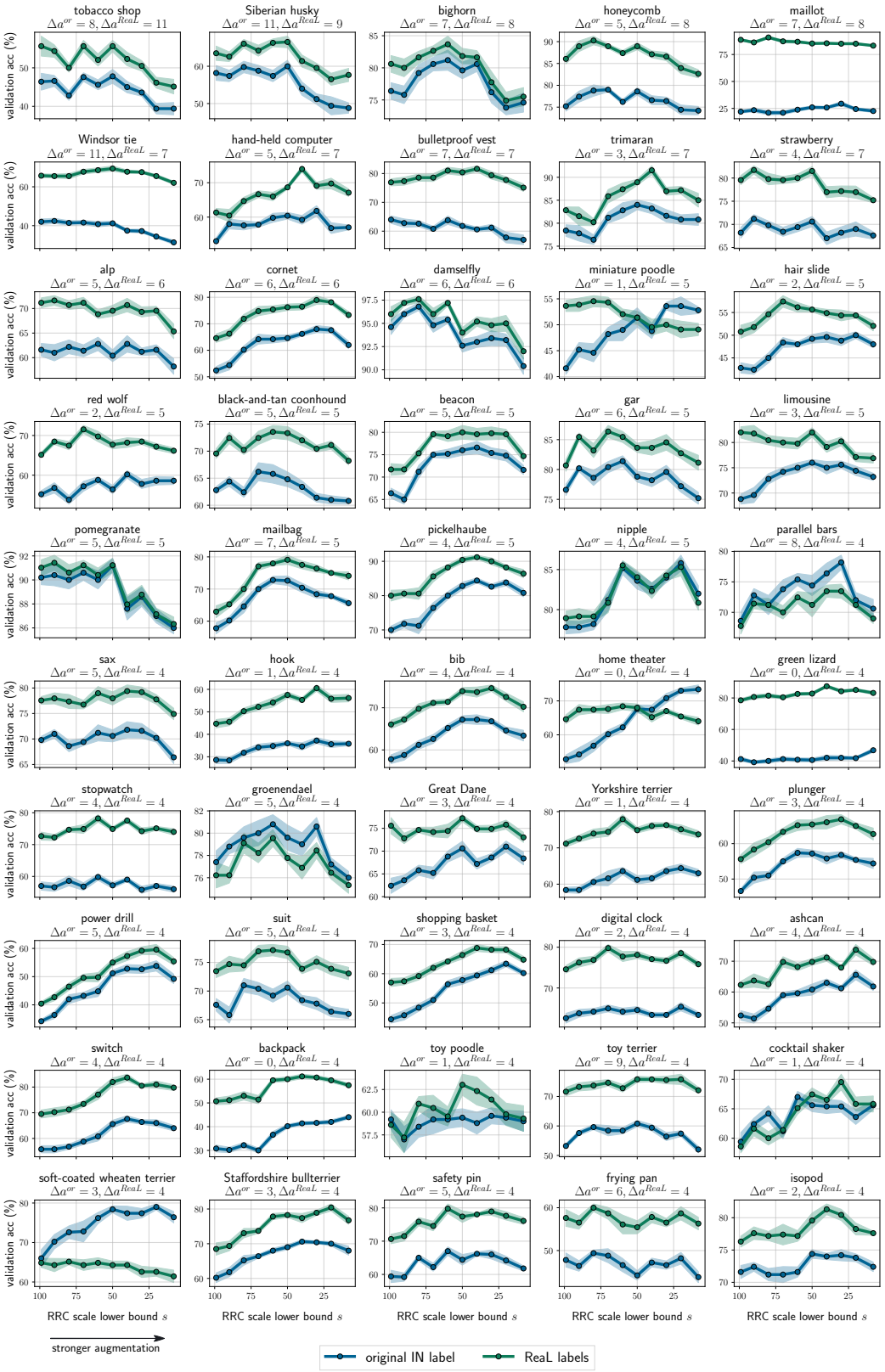


*Figure 6.* Per-class class validation accuracies of ResNet-50 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound $s$. We show the accuracy trends for the classes with the highest difference between the maximum ReaL accuracy on that class across augmentation levels $\max_s a_k^{ReaL}(s)$ and the ReaL accuracy of the model trained with $s = 8\%$. On each subplot below the name of the class we show the accuracy drops with respect to original and ReaL labels: $\Delta a_k^{or}$ and $\Delta a_k^{ReaL}$. We report the mean and standard error over 10 independent runs of the network.

*Table 2.* Confusions on the classes most affected by data augmentation.

| Affected class $k$ | Confused class $l$ | Δ conf. rate (%) | | Label co-occur. | | Semantic sim. | | Confusion type |
|---|---|---|---|---|---|---|---|---|
| | | $\Delta CR_{k\to l}$ | $\Delta CR^*_{l\to k}$ | $C_{lk}$ | IoU | WN | spacy | |
| overskirt | hoopskirt | 5.80 | 3.60 | 0.31 | 0.17 | 0.91 | – | fine-gr. (ambig.) |
| | bonnet | 4.20 | 0.00 | 0.03 | 0.02 | 0.73 | 0.32 | fine-gr. |
| | gown | 4.00 | 2.40 | 0.50 | 0.21 | 0.73 | 0.37 | fine-gr. (ambig.) |
| | trench coat | 3.60 | 0.40 | 0.00 | 0.00 | 0.75 | 0.42 | fine-gr. |
| academic gown | mortarboard | 18.40 | 7.00 | 0.72 | 0.50 | 0.73 | 0.10 | co-occur. |
| sunglass | sunglasses | 13.00 | 22.40 | 0.87 | 0.81 | 0.64 | 0.84 | ambig. |
| maillot | maillot | 15.00 | 7.20 | 0.73 | 0.63 | 0.70 | 1.00 | ambig. |
| Windsor tie | suit | 7.20 | 4.00 | 0.61 | 0.32 | 0.82 | 0.24 | co-occur. |
| screen | desktop computer | 7.80 | 7.00 | 0.59 | 0.29 | 0.64 | 0.62 | ambig. |
| | monitor | 3.20 | 6.40 | 0.87 | 0.37 | 0.63 | 0.44 | ambig. |
| tobacco shop | barbershop | 5.20 | 2.80 | 0.00 | 0.00 | 0.91 | 0.56 | fine-gr. |
| | bookshop | 6.80 | 6.40 | 0.00 | 0.00 | 0.91 | 0.53 | fine-gr. |
| monastery | church | 2.80 | 6.80 | 0.11 | 0.03 | 0.70 | 0.71 | fine-gr. |
| | castle | 2.80 | 11.20 | 0.00 | 0.00 | 0.60 | 0.69 | fine-gr. |
| thresher | harvester | 6.60 | 16.40 | 0.04 | 0.01 | 0.90 | 0.49 | fine-gr. |
| parallel bars | horizontal bar | 3.20 | 2.80 | 0.00 | 0.00 | 0.90 | 0.75 | fine-gr. |
| | balance beam | 3.00 | 4.00 | 0.02 | 0.01 | 0.90 | 0.45 | fine-gr. |
| mailbag | purse | 12.80 | 2.00 | 0.10 | 0.06 | 0.89 | 0.19 | fine-gr. |
| | backpack | 4.00 | 5.60 | 0.00 | 0.00 | 0.89 | 0.16 | fine-gr. |
| chain | necklace | 9.40 | 4.40 | 0.15 | 0.09 | 0.53 | 0.31 | ambig. |
| bulletproof vest | military uniform | 5.60 | 3.40 | 0.31 | 0.13 | 0.76 | 0.38 | co-occur. (ambig.) |
| | assault rifle | 3.20 | 0.40 | 0.32 | 0.17 | 0.40 | 0.35 | co-occur. |
| sombrero | cowboy hat | 7.40 | 4.80 | 0.15 | 0.05 | 0.91 | 0.51 | fine-gr. |
| velvet | purse | 3.60 | 2.60 | 0.00 | 0.00 | 0.62 | 0.29 | unrelated |
| | necklace | 3.00 | 0.00 | 0.00 | 0.00 | 0.62 | 0.51 | unrelated |
| tape player | radio | 3.20 | 4.60 | 0.00 | 0.00 | 0.67 | 0.27 | fine-gr. |
| | cassette player | 3.00 | 0.20 | 0.08 | 0.01 | 0.89 | 0.85 | fine-gr. |
| assault rifle | military uniform | 8.40 | 0.40 | 0.47 | 0.24 | 0.42 | 0.42 | co-occur. |
| cornet | trombone | 4.80 | 2.40 | 0.23 | 0.14 | 0.91 | 0.41 | fine-gr. |
| pole | traffic light | 4.00 | 0.40 | 0.05 | 0.03 | 0.12 | 0.21 | unrelated |
| muzzle | sandal | 3.20 | 0.00 | 0.00 | 0.00 | 0.56 | 0.23 | unrelated |
| ear | corn | 5.40 | 4.40 | 0.81 | 0.52 | 0.78 | 0.23 | ambig. |
| vault | altar | 6.40 | 4.40 | 0.21 | 0.12 | 0.62 | 0.41 | fine-gr. (ambig.) |
| frying pan | Dutch oven | 6.00 | 3.00 | 0.00 | 0.00 | 0.40 | 0.59 | fine-gr. |
| | wok | 3.40 | 2.60 | 0.09 | 0.05 | 0.92 | 0.72 | fine-gr. |
| French loaf | bakery | 4.40 | 1.80 | 0.10 | 0.06 | 0.24 | 0.42 | co-occur. |
| barrel | rain barrel | 7.60 | 2.20 | 0.16 | 0.07 | 0.76 | 0.70 | fine-gr. (ambig.) |
| spatula | wooden spoon | 4.40 | 2.80 | 0.24 | 0.12 | 0.57 | 0.62 | fine-gr. |
| sax | flute | 3.20 | 0.40 | 0.00 | 0.00 | 0.83 | 0.65 | fine-gr. |
| seashore | sandbar | 3.80 | 2.80 | 0.64 | 0.47 | 0.57 | 0.69 | co-occur. |
| coffee mug | cup | 7.80 | 0.80 | 0.61 | 0.34 | 0.19 | 0.63 | ambig. |
| | espresso | 3.00 | 2.60 | 0.18 | 0.13 | 0.21 | 0.72 | co-occur. |
| breastplate | cuirass | 6.00 | 6.40 | 0.71 | 0.50 | 0.67 | 0.48 | ambig. |
| | shield | 3.20 | 1.20 | 0.07 | 0.05 | 0.70 | 0.59 | |
| beacon | breakwater | 7.80 | 0.60 | 0.07 | 0.04 | 0.71 | 0.33 | co-occur. |
| suit | miniskirt | 3.20 | 1.60 | 0.02 | 0.01 | 0.86 | 0.32 | fine-gr. |
| hand-held computer | cellular telephone | 8.80 | 5.60 | 0.22 | 0.06 | 0.50 | 0.42 | ambig. |
| | notebook | 4.60 | 0.40 | 0.03 | 0.01 | 0.92 | 0.32 | fine-gr. |
| stopwatch | digital watch | 4.80 | 0.60 | 0.00 | 0.00 | 0.83 | 0.62 | fine-gr. |
| strawberry | trifle | 4.40 | 1.40 | 0.06 | 0.03 | 0.32 | 0.40 | co-occur. |
| trimaran | catamaran | 4.80 | 1.40 | 0.18 | 0.09 | 0.92 | 0.60 | fine-gr. |
| digital clock | digital watch | 3.00 | 7.00 | 0.02 | 0.01 | 0.83 | 0.71 | fine-gr. |
| hair slide | necklace | 5.60 | 0.60 | 0.00 | 0.00 | 0.50 | 0.42 | fine-gr. |
| hook | necklace | 3.60 | 0.00 | 0.00 | 0.00 | 0.53 | 0.33 | unrelated |
| backpack | purse | 3.00 | 0.00 | 0.02 | 0.01 | 0.89 | 0.56 | fine-gr. |
| home theater | monitor | 2.80 | 0.00 | 0.03 | 0.00 | 0.56 | 0.18 | co-occur. |
| bath towel | pillow | 4.40 | 0.60 | 0.00 | 0.00 | 0.59 | 0.56 | unrelated |

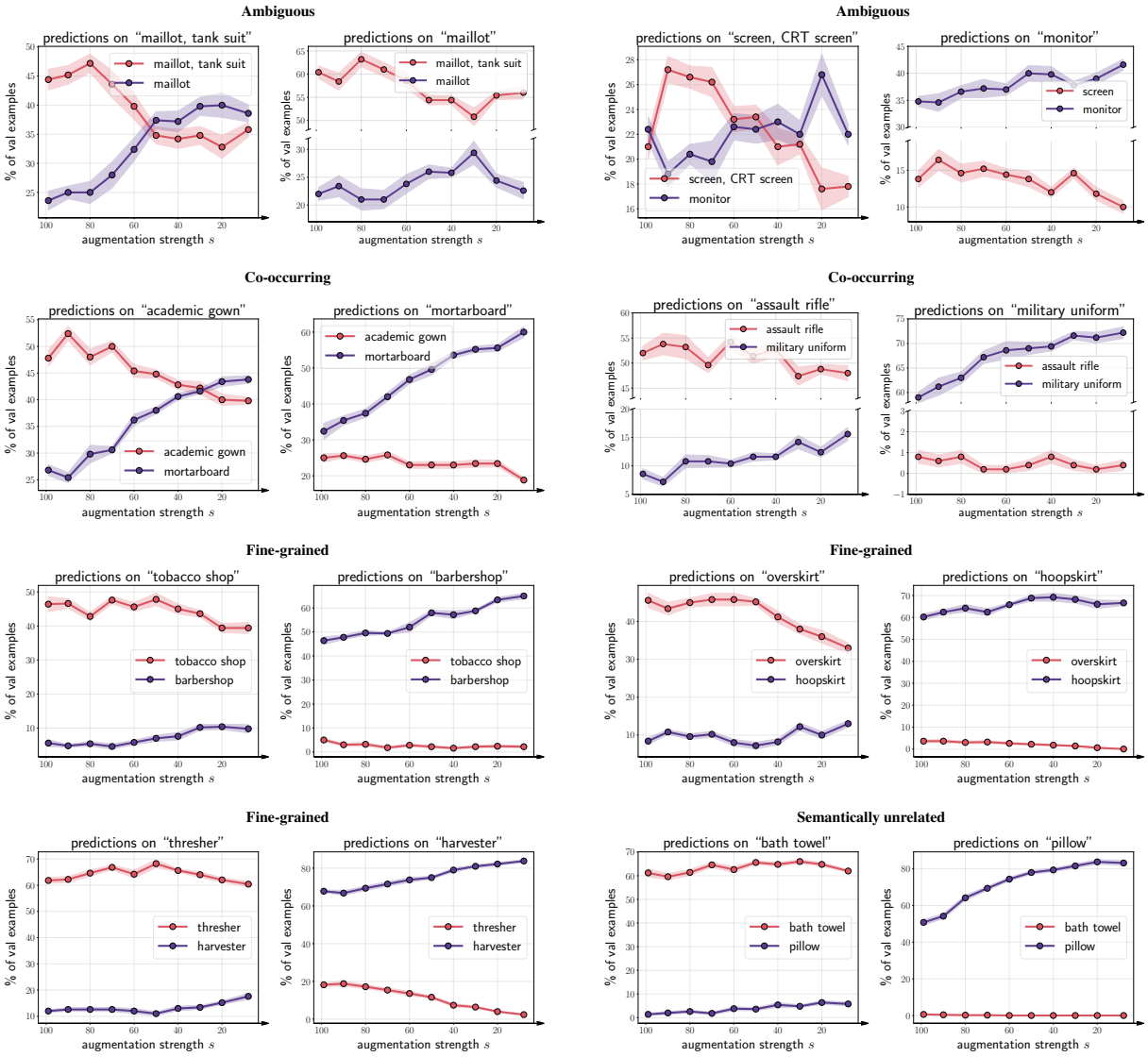


*Figure 7.* Confusion rate for classes most negatively affected by strong data augmentation and the corresponding classes they get confused with. We categorize confusions into ambiguous, co-occurring, fine-grained and unrelated.

*Table 3.* Class-conditional augmentation intervention using ReaL labels.

| # classes with adapted aug. | ReaL avg acc | ReaL avg acc of 50 aff. classes | ReaL avg acc of remaining 950 classes |
|---|---|---|---|
| 0 | $83.70_{\pm 0.01}$ | $70.66_{\pm 0.08}$ | $84.00_{\pm 0.01}$ |
| 10 | $83.63_{\pm 0.01}$ | $72.01_{\pm 0.04}$ | $83.86_{\pm 0.01}$ |
| 30 | $83.64_{\pm 0.01}$ | $72.28_{\pm 0.05}$ | $83.86_{\pm 0.01}$ |
| 50 | $83.57_{\pm 0.01}$ | $72.20_{\pm 0.03}$ | $83.78_{\pm 0.01}$ |

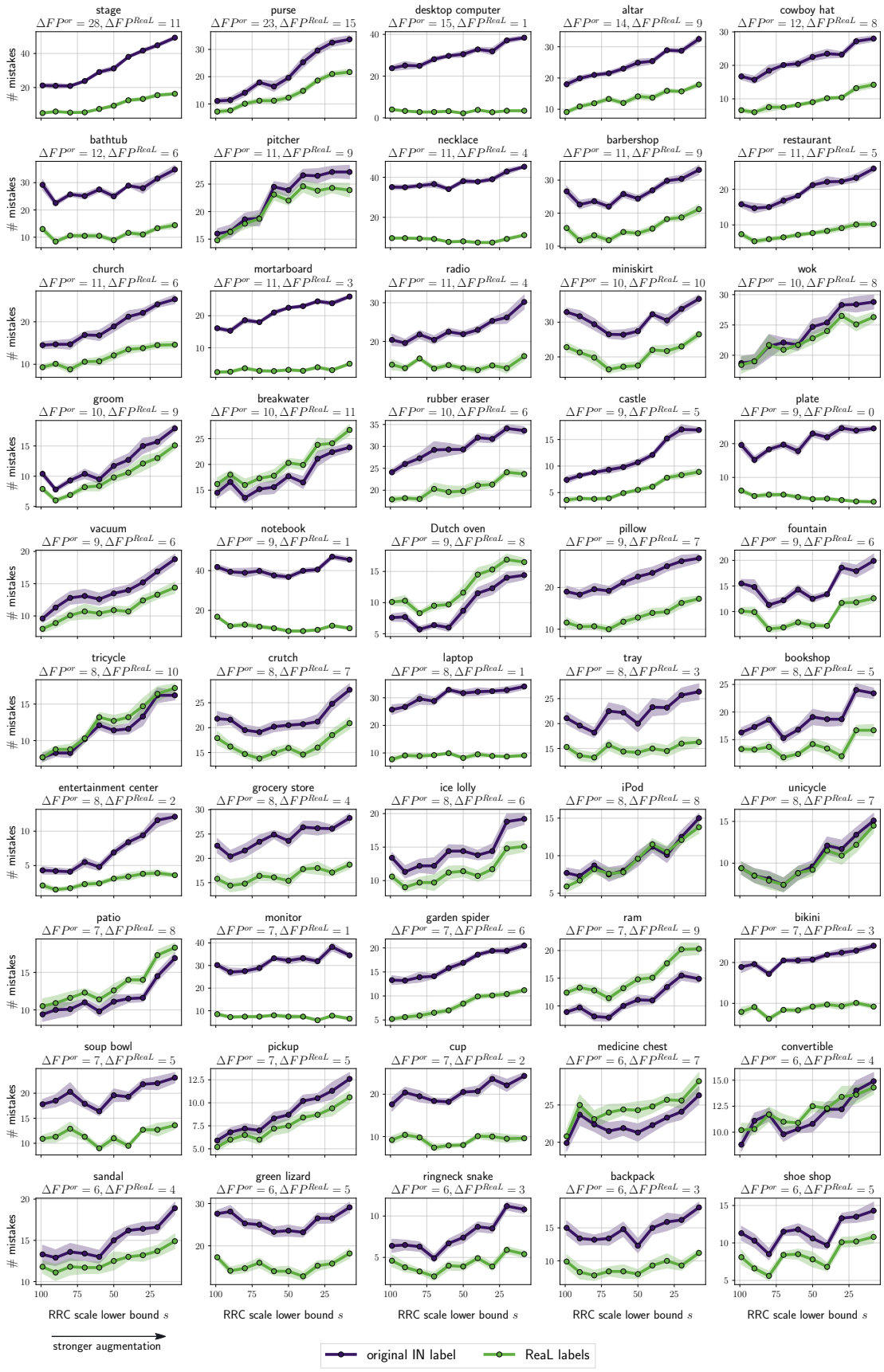


*Figure 8.* The number of per-class False Positive (FP) mistakes for the set of classes where FP computed with original labels increases the most when using strong data augmentation. We show the trends using both original and ReaL labels.

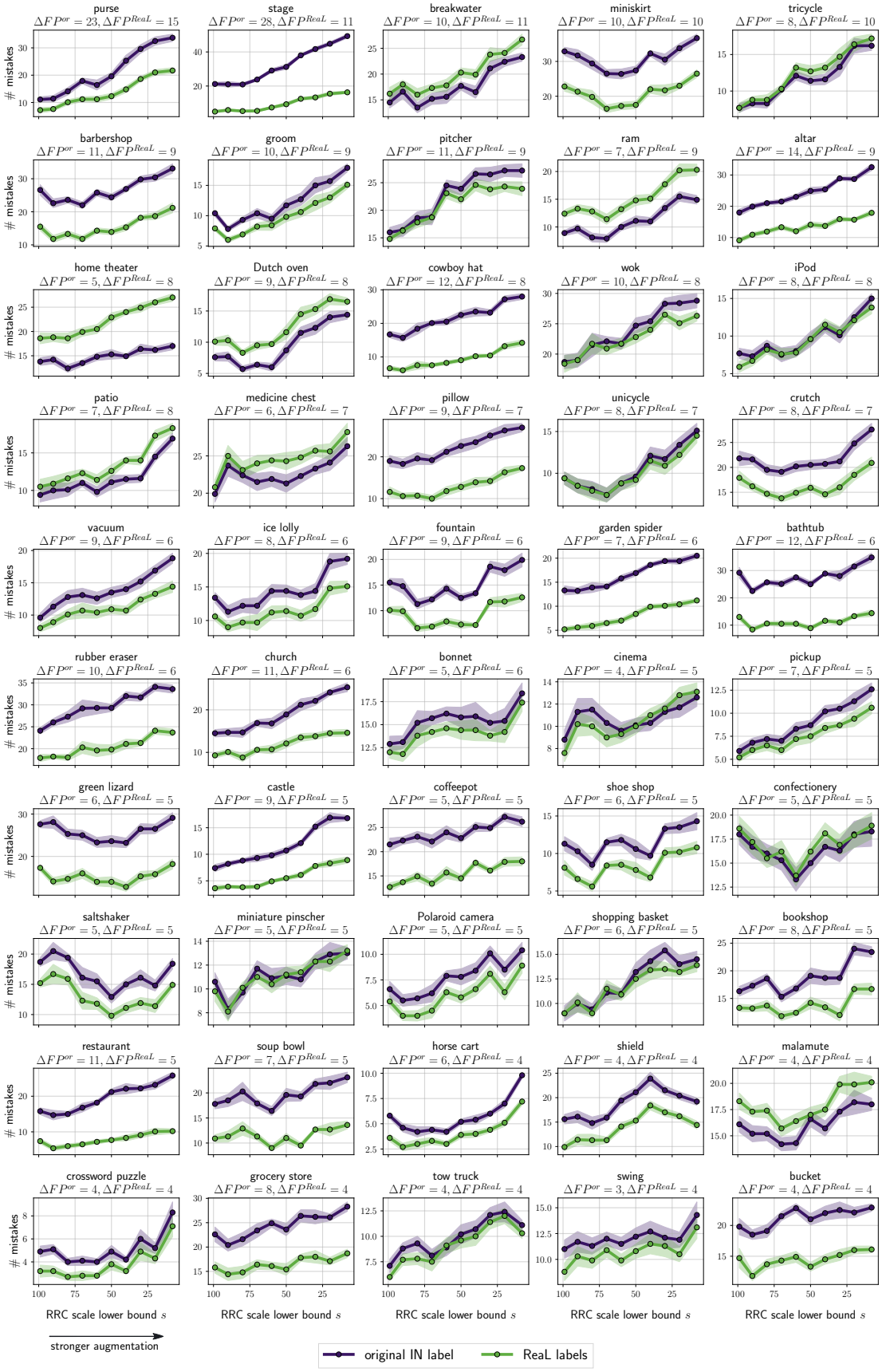


*Figure 9.* The number of per-class False Positive (FP) mistakes for the set of classes where FP computed with ReaL labels increases the most when using strong data augmentation. We show the trends using both original and ReaL labels.