Understanding the Detrimental Class-level Effects of Data Augmentation: Supplementary Material

A Training details

Following [1], we train ResNet-50 models for 88 epochs with SGD with momentum 0.9, using batch size 1024, weight decay 10^{-4}, and label smoothing 0.1. 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` [45], automatic mixed precision training with `torch.amp` package[5], `ffcv` package [34] for fast data loading. We use image resolution 176 during training, and resolution 224 during evaluation, following Touvron et al. [61] and `torchvision` training recipe[6]. Balestriero et al. [1] 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.

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 [4]. 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 5 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 6, 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:

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

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

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

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:

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

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 6 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 Δa_k following Balestriero et al. [1] and measure the average accuracy on this set of classes as a function of s and after interventions in Section 6.

In Section 6, 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 $FP_k(s_k^*) = \min_s FP_k(s)$.

C Additional related work

Adaptive and learnable data augmentation. Xu et al. [66] 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. [47] 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. [49] and Ratner et al. [48] learn DA transformation using GAN framework, while Hu and Li [28] study the bias of GAN-learned data augmentation. Fujii et al. [16] take into account the distances between classes to adapt mixed-sample DA. Hauberg et al. [20] learn class-specific DA on MNIST. Numerous works, e.g. Cubuk et al. [12], Lim et al. [36], Ho et al. [24], Hataya et al. [19], Li et al. [35], Cubuk et al. [13], Tang et al. [58], Müller and Hutter [42] and Zheng et al. [69] find dataset-dependent augmentation strategies. Benton et al. [3] proposed Augerino framework to learn augmentation form training data. Zhou et al. [70], Cheung and Yeung [11], Mahan et al. [39] and Miao et al. [40] learn class- or input-dependent augmentation policies. Yao et al. [67] propose to modify mixed-sample augmentation to improve out-of-domain generalization.

Robustness and model evaluation beyond average accuracy. While Miller et al. [41] 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. [60] showed counter-examples for "accuracy-on-the-line" phenomenon. Kaplun et al. [32] show that while model's average accuracy improves during training, it may decrease on a subset of examples. Sagawa et al. [51] show that training with Empirical Risk Minimization may lead to suboptimal performance in the worst case. Bitterwolf et al. [6] evaluated ImageNet models' performance in terms of a number of metrics beyond average accuracy, including worst-class accuracy and precision.

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 Δa_k^{or} in Figure 5, and (2) 50 classes most negatively affected in ReaL accuracy In Figure 6.

E Class confusion types

In Table 1 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:

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

We use WordNet class similarity and similarity of word embeddings from spacy [25] 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 5.

F Additional details for the 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 6, 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 6 we show results for adapting augmentation level for classes using original labels to evaluate False Positive and False Negative mistakes, in Table 2 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).

We also experimented with fine-tuning the model from the checkpoint trained with the strongest augmentation $s = 8\%$ using either regular augmentation policy which was used during training or class-conditional policy with augmentation strength changed for $k = 10$ classes as in Section 6: we fine-tuned the model for 5 epochs with linearly decaying learning rate starting from the value 10^{-4}. However, both regular and class-conditional DA lead to slight drop in average accuracy on all classes (from 76.79% to 76.73% for either DA) and in particular the accuracy dropped more significantly for negatively affected classes: from 53.93% to 53.4%. We hypothesize that this is due to model memorizing train examples so even class-conditional augmentation policy is not able to recover performance on the affected classes if we re-use the same data for fine-tuning. In the future analysis, we will explore whether it is possible to alleviate DA bias if we fine-tune the model from an earlier checkpoint as opposed to fully trained model or if we use additional held-out data for fine-tuning.

G Multi-label annotations

In this work we use ReaL labels released in Beyer et al. [4] to account for the label noise in evaluation of per-class accuracy effects of data augmentation. A more recent work Vasudevan et al. [64] released re-assessed multi-label annotations for a half of the ImageNet validation set. Since they did not release the annotations for the entire validation set, we decided to use older and more commonly used ReaL labels. However, one could merge the two multi-label annotation sets from Beyer et al. [4] and Vasudevan et al. [64] for more accurate evaluation. In particular, Vasudevan et al. [64] discussed the class mappings that they collapsed, and among those classes are the ones negatively affected in ReaL accuracy by data augmentation, e.g. "siberian huskies are also eskimo dogs", "coffee mug is also a cup", "maillot and maillot, tanksuit are the same class" "monitor and screen are the same class", "cassette player is also a tape player" [64].

H 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 [12, 42]. 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 and study per-class accuracies of EfficientNet-B0 [57] in Section I, however, Balestriero et al. [1] showed that per-class biases persist in other architectures like Vision Transformers [14] and DenseNets [29] 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 (also see Section G) and imprecise word similarity metrics.

Broader impact. A potential negative outcome that can result from misinterpretation of our analysis in Section 4 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.

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 4). The experiments were run on GPU clusters on Nvidia Tesla V100, Titan RTX, RTX8000, 3080 and 1080Ti GPUs.

I Additional architecture results

On Figures 10 and 11 we show the per-class accuracy trends for classes most affected in original and ReaL accuracy of EfficientNet-B0 [57] model, trained using a similar setup to the main ResNet-50 model (see Section A). We can see that many affected classes are the same for ResNet-50 and EfficientNet models.

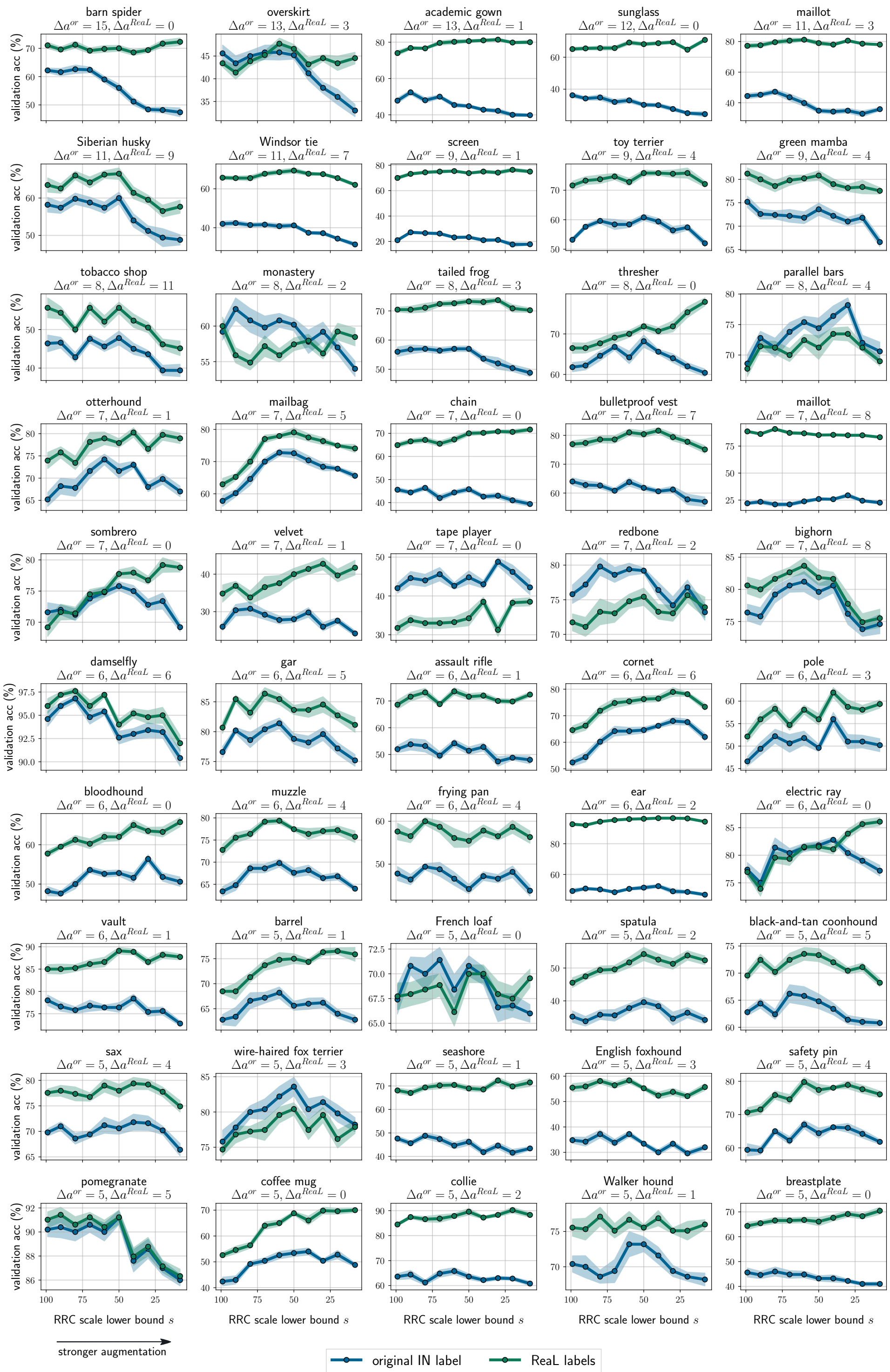

Figure 5: Per-class class validation accuracies of ResNet-50 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest difference between the maximum accuracy on that 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: Δa_k^{or} and Δa_k^{ReaL}. We report the mean and standard error over 10 independent runs of the network.

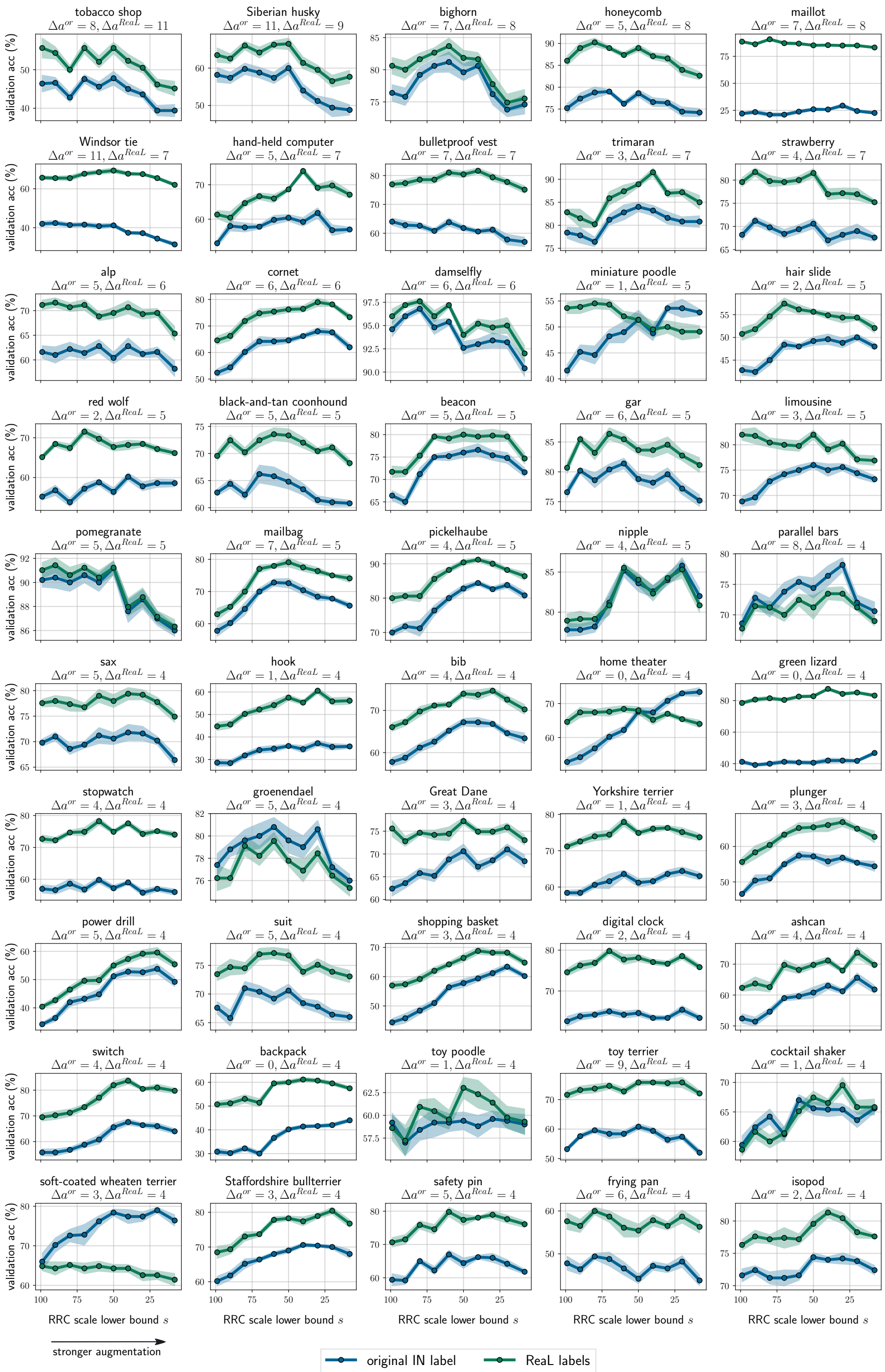

Figure 6: Per-class class validation accuracies of ResNet-50 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest difference between the maximum ReaL accuracy 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: Δa_k^{or} and Δa_k^{ReaL}. We report the mean and standard error over 10 independent runs of the network.

Table 1: 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

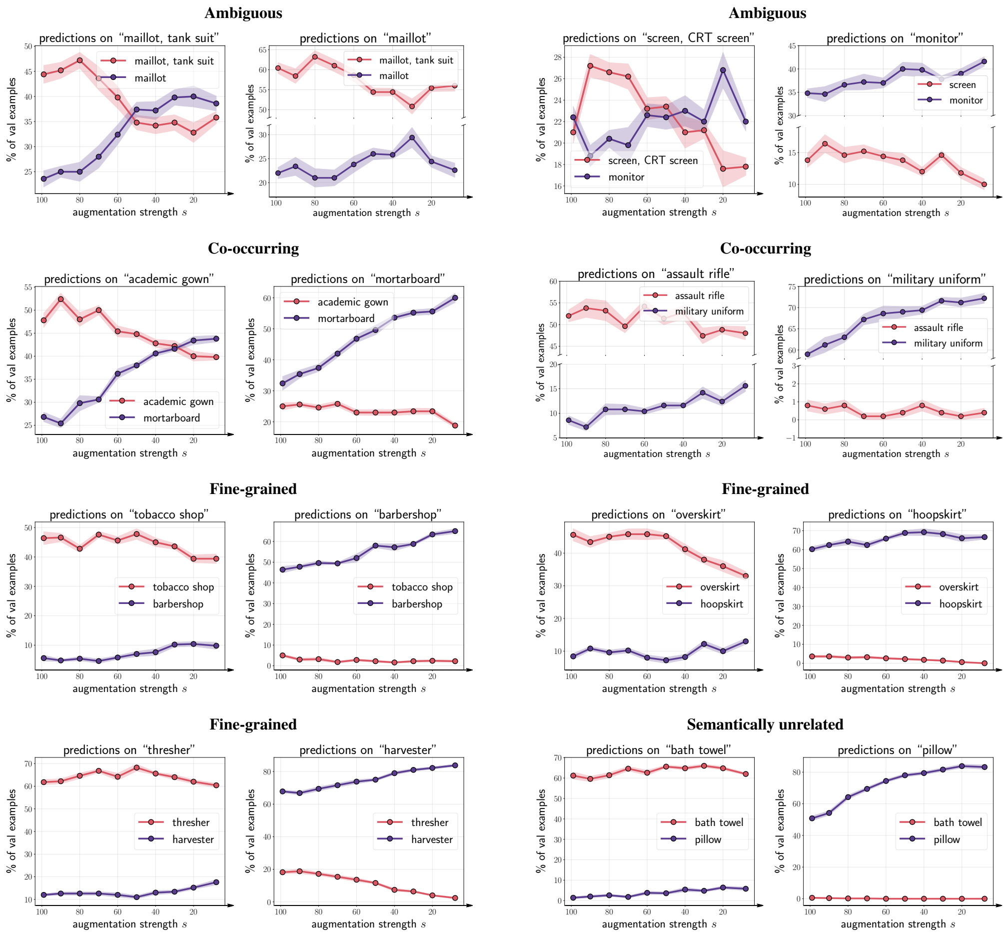

Figure 7: Confusion rate for classes most negatively affected by strong data augmentation and the corresponding classes they get confused with. We categorize confusions into ambiguous, co-occurring, fine-grained and unrelated.

Table 2: 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	$83.57_{\pm 0.01}$	$72.20_{\pm 0.03}$	$83.78_{\pm 0.01}$

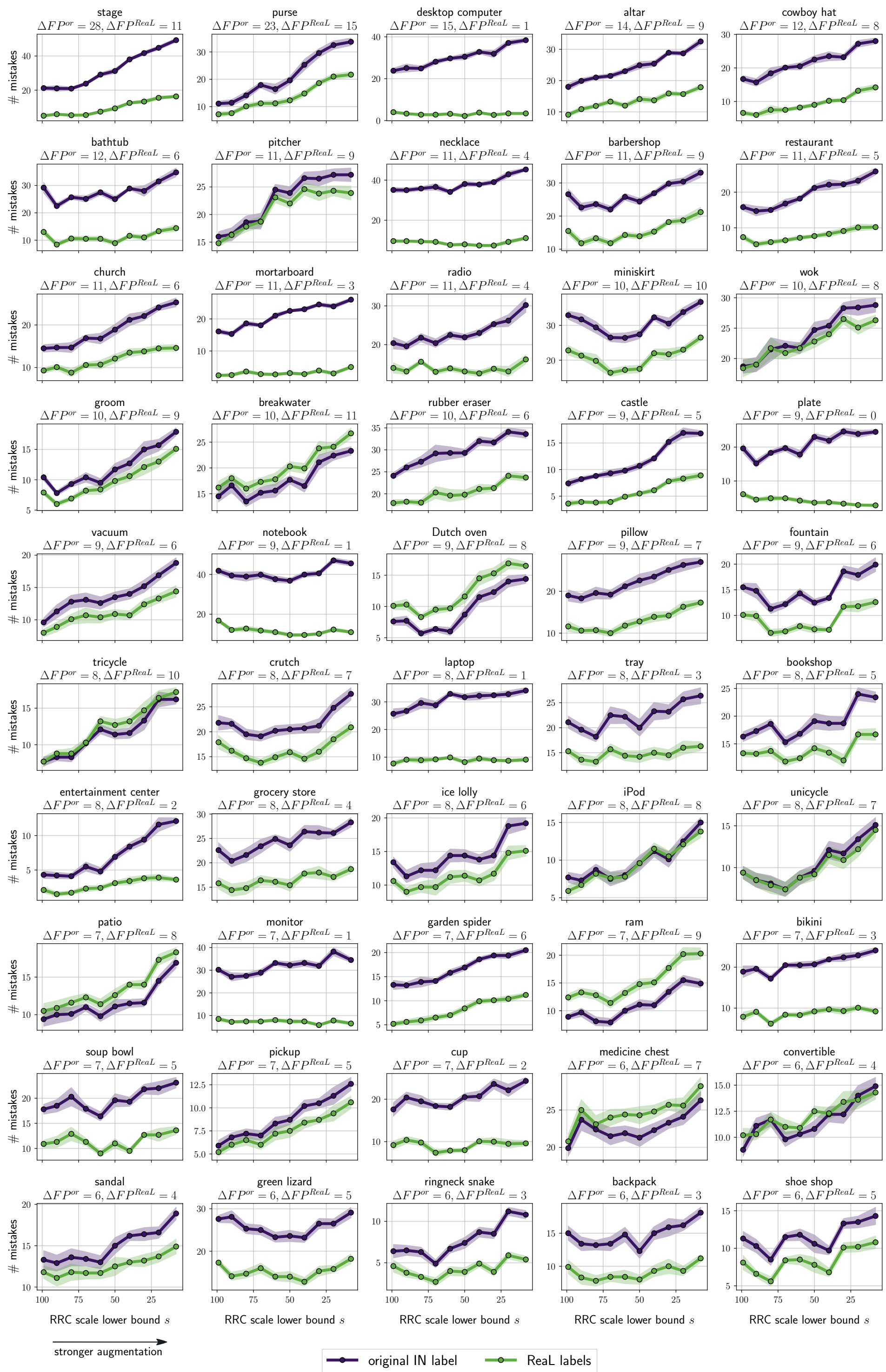

Figure 8: The number of per-class False Positive (FP) mistakes for the set of classes where FP computed with original labels increases the most when using strong data augmentation. We show the trends using both original and ReaL labels.

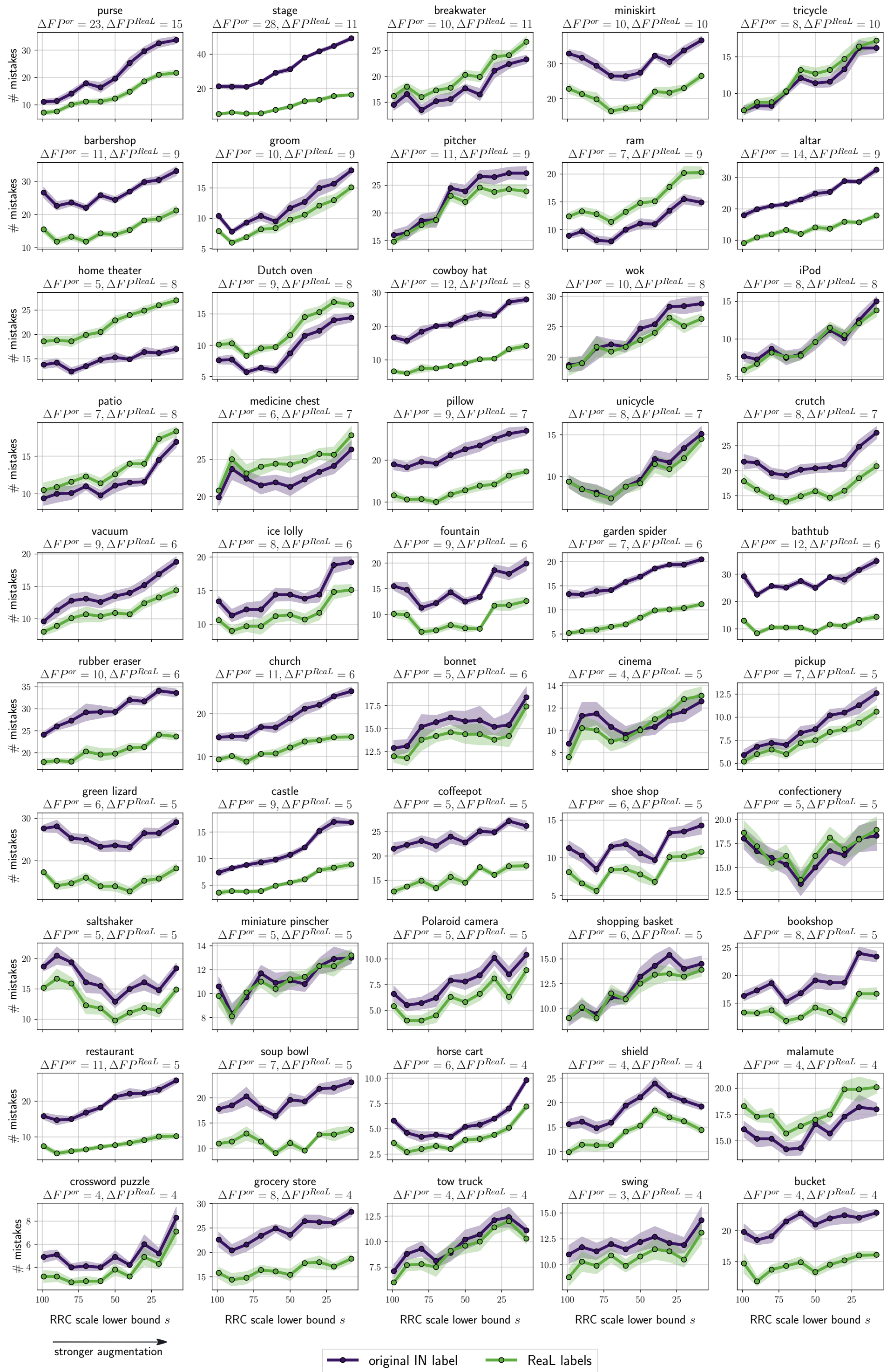

Figure 9: The number of per-class False Positive (FP) mistakes for the set of classes where FP computed with ReaL labels increases the most when using strong data augmentation. We show the trends using both original and ReaL labels.

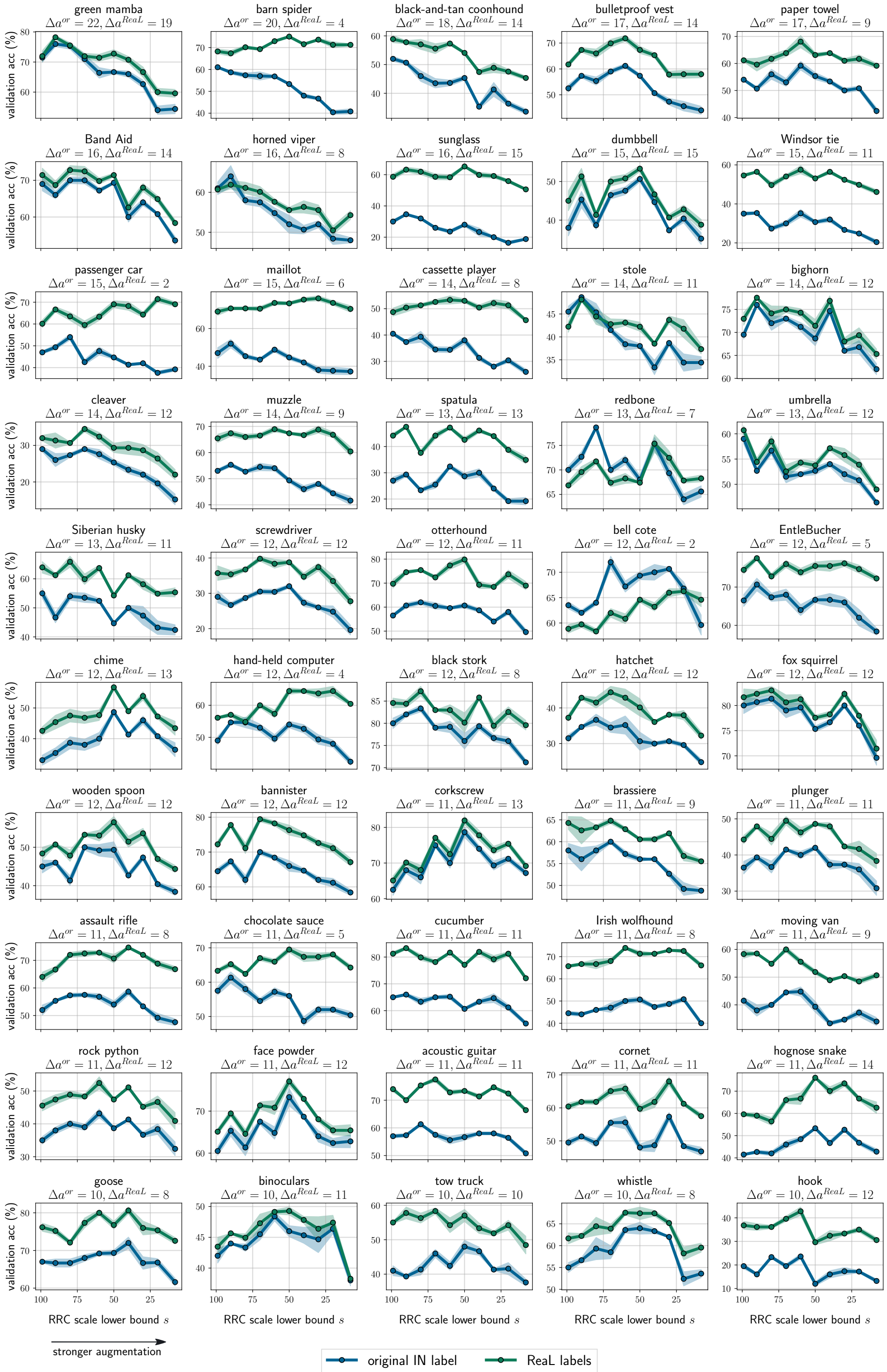

Figure 10: Per-class class validation accuracies of EfficientNet-B0 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest difference between the maximum accuracy 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: Δa_k^{or} and Δa_k^{ReaL}.

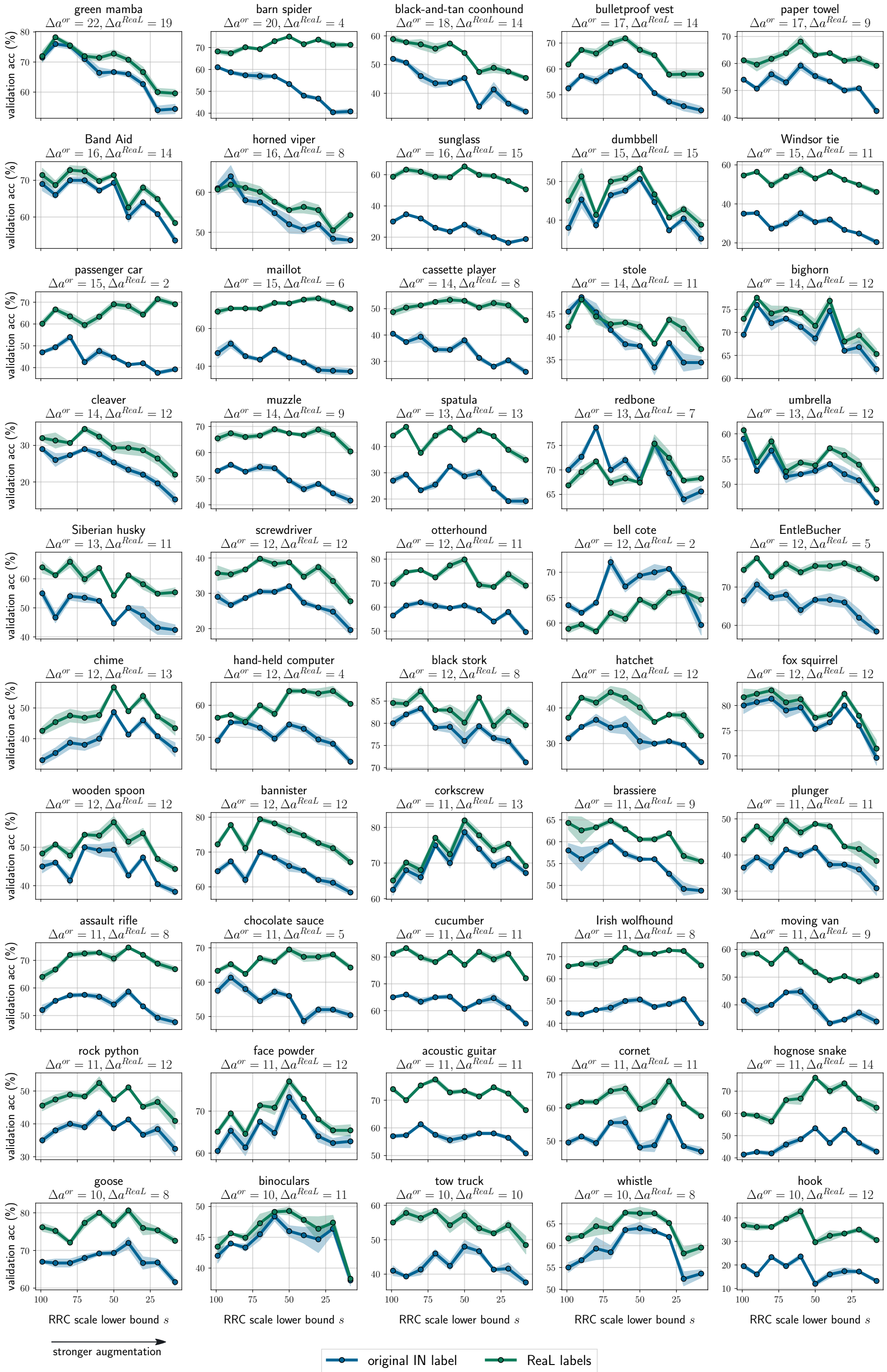

Figure 11: Per-class class validation accuracies of EfficientNet-B0 trained on ImageNet computed with original and ReaL labels as a function of Random Resized Crop data augmentation scale lower bound s. We show the accuracy trends for the classes with the highest difference between the maximum ReaL accuracy on that class across augmentation levels $\max_s a_k^{or}(s)$ 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: Δa_k^{or} and Δa_k^{ReaL}.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[15] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

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

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

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

[19] Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. (2020). 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*, pages 1–16. Springer.

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

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

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

[23] Hernández-García, A. and König, P. (2018). 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*, pages 95–103. Springer.

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

[25] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.

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

[27] Hovy, D. and Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[46] 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. (2019). 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., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

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

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

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

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

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

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

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

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

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

[57] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

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

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

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

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

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

[63] Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. (2015). 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*, pages 595–604.

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

[65] Xu, M., Yoon, S., Fuentes, A., and Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, page 109347.

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

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

[68] Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., and Chun, S. (2021). 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*, pages 2340–2350.

[69] Zheng, Y., Zhang, Z., Yan, S., and Zhang, M. (2022). Deep autoaugment. *arXiv preprint arXiv:2203.06172*.

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