Overview. We include ImageNet-R results, better models on ImageNet-C and improve our ablation studies.

We thank the reviewers for their extensive and helpful comments which contributed to improving our manuscript. The reviewers state that the "simplicity of the method is appealing, and it provides a substantial improvement for little extra effort" (R3) and agree on the importance of the considered problem for the NeurIPS community (R1,2). Below, we address main concerns and discuss updated results with more robust models (DeepAugment) and new datasets (ImageNet-R) which appeared in parallel work during the review phase. We also incorporated most of the suggestions regarding figure formatting and formal methods in the camera-ready version.

R1, R2, R4: Does the proposed method generalize to other datasets? We already showed gains across the 15 different datasets in the IN-C benchmark (of four different types). We now extend this analysis to 15 new data shifts in ImageNet-R (IN-R; 200 class IN, 30,000 images), another large image dataset with more challenging dataset shifts like art, cartoons, deviantart or graffiti. We observe consistent gains (Table 1) with a new RN50 SoTA of 48.9% when using a batch size of 2048 for adaptation. For the vanilla RN50, we observe performance improvements on IN-R when using a batch size larger than 32 (Fig. 1) almost reaching AugMix performance w/o adaptation for large batch sizes.

T1: ImageNet-R (n=2048), top-1 error.

Model, adaptation:	base	adapt
ResNet50	63.8	59.9
Fixup	61.2	—
GroupNorm	65.0	—
SIN	58.6	54.2
ANT	61.0	58.0
ANT+SIN	53.8	52.0
AugMix (AM)	59.0	55.8
DeepAug (DAug)	57.8	52.5
DAug + AM	53.2	48.9
DAug + AM (RNXt101)	**47.9**	**44.0**

Fig 1: ImageNet-R results

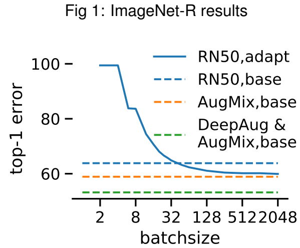

Table 2: New models on IN-C (n=2048), mCE

Model	base	adapt
DeepAug	60.36	49.44
DeepAug+AugMix	53.55	45.36
DeepAug+AugMix+RNXt101	**44.52**	**37.96**

T. 3a: ObjectNet evaluation (n = 512), acc; T. 3b: Mixed IN-C, err

ResNet50 model	T. 3a top-1	T. 3a top-5	T. 3b top-1	T. 3b top-5
BatchNorm w/o adapt	21.85	39.09	61.08	40.81
BatchNorm w/ adapt	24.04	41.15	60.87	40.31
GroupNorm	**29.18**	**50.24**	57.25	35.97
Fixup	28.52	48.56	**56.83**	**35.43**

Clarifications around novelty & central hypotheses: Adaptation of BN layers is a well-known method in domain adaptation. Our contribution is to extensively evaluate (and theoretically analyze) its performance on *systematic* dataset shifts in both large and small sample size adaptation scenarios, and to show that a domain adaptation evaluation scenario has the potential to substantially improve over the ad-hoc setting on robustness datasets, making it a strong baseline. Our main hypotheses (**H**) and tests (**T**) (asked by R3) are:

- **H**: *Systematic* dataset shifts yield a mismatch in internal statistics and result in decreased accuracy. **T**: The Wasserstein distance between source and target statistics quantifies the amount of mismatch and is predictive of degradation, especially within a corruption type.
- **H**: Correcting the statistics improves accuracy under distribution shift. **T**: We show consistent, substantial improvements due to BN adaptation across a wide range of models and 17 domains (15 IN-C + IN-R + ON).
- **H**: The observed sample size performance trade-off can be explained by statistical estimation errors (theoretical model) and can be mitigated using a Bayesian approach. **T**: We propose a theoretical model to qualitatively explain the sample size vs. performance degradation trade-off and propose an easy fix for the small sample case.

Additional Control Experiments (ObjectNet, mixed IN-C) R4 discussed our negative results on IN-V2 and ON. We want to stress that these results are control experiments, and the observed outcome matches the expectations. BatchNorm adaptation can only mitigate *systematic shifts* in the data distribution, which is unlike the shift in IN-V2 (iid data, or a more complex sampling bias) or ObjectNet (complex distributional shift by random variations in poses, etc).

To stress this point, we perform two additional controls: We evaluate GroupNorm + Fixup on ObjectNet as suggested by R4, which outperform the BN model (T3a). We also randomly sample 50,000 IN-C images across corruptions and severities (3 seeds), destroying the systematic shift. GN+Fixup now also outperforms BN w + w/o adapt (Table 3b).

Use of exponential moving average instead of a weighted average (R1) We agree that this is the correct method especially for practitioners, and added a note in the Appendix. Results are indistinguishable from the "full adaptation" results due to the large number of samples in the test set and we can add a short comparison on this to the appendix.

Manuscript edits We fixed Figs.1,2,4 according to R2's suggestions; the color code in Fig. 4, IN-V2 was indeed wrong, colors should match in the limit of many samples (adaptation converges to baseline performance). We revised § 1–2 & fixed Def. 1. We revised Fig. 3 and note linear relationships between the Wasserstein distance & accuracy *both before and after* adaptation, highlighting the usefulness to quantify domain shift; we do not observe a relationship between Wasserstein distance and the amount of correction by adaptation (R3) and will add a supplementary figure. We thoroughly revised the appendix and sectioning.