Leaving Reality to Imagination: Robust Classification via Generated Datasets

Hritik Bansal [1] **Aditya Grover** [1]

Abstract

Recent research on robustness has revealed significant performance gaps between neural image classifiers trained on datasets that are similar to the test set, and those that are from a naturally *shifted* distribution, such as sketches, and animations of the object categories observed during training. However, the notion of a dataset is also undergoing a paradigm shift in recent years. With drastic improvements in the quality, and ease-of-use to modern generative models, generated data is pervading the web. In this light, we study the question: How do these generated datasets influence the natural robustness of image classifiers? We find that Imagenet classifiers trained on real data augmented with generated data achieve higher accuracy and effective robustness than standard training and popular augmentation strategies in the presence of natural distribution shifts. Additionally, we find that the standard ImageNet classifiers suffer a performance degradation of upto 20% on the generated data, indicating their fragility at accurately classifying the objects under novel variations. Lastly, we demonstrate that the image classifiers trained on real data augmented with generated data from the base generative model, exhibit greater resilience to natural distribution shifts compared to the classifiers trained on real data augmented with generated data from the finetuned generative model on the real data. The code is available at https://github.com/Hritikbansal/generative-robustness.

[1]Department of Computer Science, UCLA. Correspondence to: Hritik Bansal <hbansal@g.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

One effective strategy to improve robustness is to enlarge the amount of training data by designing intricate augmentations (Hendrycks et al., 2019; 2022; 2021) of the training data that aid the generalization of classifier to novel domains. Similarly, datasets can also be enlarged by scraping multimodal paired datasets, such as image-caption pairs on the Internet (Radford et al., 2021; Jia et al., 2021; Pham et al., 2021). However, the notion of a dataset is also experiencing a paradigm shift in recent years. With the emergence of modern 'in the wild' generative models (Ramesh et al., 2022; Nichol et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Chang et al., 2023), generated data is pervading the web (Wang et al., 2022; Kirstain et al., 2023). These models are trained on large diverse datasets (Schuhmann et al., 2022) with open vocabulary annotations, such that post-training, they can synthesize high-fidelity images for a wide range of concepts in a *zero-shot* manner. Notably, these models are not limited to generate a fixed, finite set of hand-engineered augmentations and can be repeatedly queried to generate diverse data through various conditioning mechanisms such as text prompts, and guidance strategies.

In this work, we study the question: How do datasets generated from modern in-the-wild generative models influence the natural robustness of image classifiers? Specifically, we focus on the classification accuracy (Ravuri & Vinyals, 2019), and the effective robustness (Taori et al., 2020) of the standard classifiers trained from scratch. We present an overview of our setup in Figure 1. For generating data, we utilize Stable Diffusion (Rombach et al., 2022), an in-the-wild, open-source conditional generative model and create a synthetic dataset conditioned on objects from two source datasets ImageNet-1K (Deng et al., 2009) and ImageNet-100 (Tian et al., 2020). By repeatedly sampling from Stable Diffusion by prompting it with diverse captions for the class labels, we generate a large and diverse synthetic dataset. Specifically, we generate 1.3M synthetic images for training and 50K images for validation, which is the same size as the real ImageNet-1K training and validation data. This complements concurrent works on using synthetic data for augmentating and improving the accuracy of contrastive methods (He et al., 2022; Radford et al., 2021) on image classification and other works (Trabucco et al., 2023; Azizi et al., 2023) that study generative augmentations post-

finetuning of the part or whole of the generative model on the real data distribution. Our work focusses on the more challenging setting of transfer to image classifiers without any finetuning of the base generative model on the real images. We provide further comparison with the change in the data generation paradigm in App. §O.

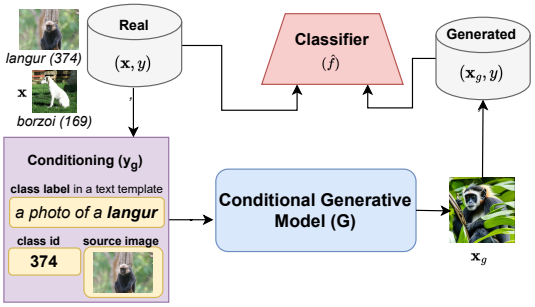


Figure 1. Overview of our approach. Our method creates generated dataset using a conditional generative model. The real dataset is then augmented with the generated dataset to train a classifier.

Our main takeaway is that training a classifier on a combination of real and generated data can achieve high absolute performance and high effective robustness (§3.1) on natural distribution shift datasets. Removing either real or generated data results in a corresponding reduction in accuracy and effective robustness respectively, thus necessitating the use of a mixture. Previous work (Yuan et al., 2022) shows that we can manipulate the generative models to adapt the images from a source domain to a single target domain which results in accurate classifiers on the target domain. Here, we create a single generated dataset from a diverse set of templates without customizing it to a single target domain.

To further explain our results, we find that the 'in-the-wild' aspects of modern generative indeed plays a role and substituting these generations with hand-crafted augmentation strategies or outputs of traditional class-conditional generative models is less effective (§3.2). We supplement this analysis with additional results on the impact of proportion sizes of real and generated data (App. §F.1), different multimodal conditioning strategies for data generation (App. §G.1), and a human and automatic evaluation study to assess and compare the class consistency, image quality, and diversity of the real and generated images (App. §I). Having studied the utility of the generated datasets for training, we study their use case for benchmarking the standard ImageNet classifiers. In §3.3, we find that the classifiers such as ResNet-101 (He et al., 2016), finetuned CLIP (Radford et al., 2021; Wortsman et al., 2022) and ViTs (Dosovitskiy et al., 2020; Tu et al., 2022) suffer an absolute degradation of up to 20% on the generated data created using text prompts with the class labels, suggesting their fragility to newly generated natural variations.

Finally, we study the impact of varying the data generation paradigm and evaluate the quality of the image classifiers trained on the generated data that is closer in distribution to the real data as compared to the generated data collected in a zero-shot way. In §3.4, we find that training the image classifier on the real data augmented with the generated data from the base generative model achieves high accuracy on the natural distribution shift datasets than training it on the real data augmented with the generated data synthesized from the finetuned generative model on the real ImageNet data. Our base generated and finetuned generated datasets will be made publicly available allowing for reproducible benchmarking of utility and critique of the generated datasets.

2. Background

Here, we provide a brief background on the robustness and data generation methods. A detailed background is present in the Appendix §D.

Robustness: For any classifier $\hat{f}$, we can quantify the *accuracy gap* (AG) between the accuracy on a test set $\mathcal{D}_{test}$ that follows the same distribution as the training set, and a test set that varies naturally from the training distribution $\mathcal{D}'$.

$$AG(\hat{f}, \mathcal{D}_{test}, \mathcal{D}') = A(\hat{f}, \mathcal{D}') - A(\hat{f}, \mathcal{D}_{test}) \quad (1)$$

However, a classifier that closes the accuracy gap might decrease the individual accuracies. Additionally, given a robust classifier $\hat{f}$ that offers high accuracy on the shifted datasets, we can assess it relative to the expected accuracy on the shifted dataset with a standard classifier that is trained on the source training set without any specific robustness intervention. This notion is formalized as *effective robustness* (ER) (Recht et al., 2019; 2018).

$$ER(\hat{f}, \mathcal{D}') = A(\hat{f}, \mathcal{D}') - \beta(A(\hat{f}, \mathcal{D}_{test}), \mathcal{D}', \mathcal{D}_{test}) \quad (2)$$

where $\beta(z, \mathcal{D}', \mathcal{D}_{test})$ is the accuracy on the shifted test set $\mathcal{D}'$ for a given accuracy z on the source test set $\mathcal{D}_{test}$. We calculate β by fitting a linear function on the collection of standard classifiers. Positive ER indicates that the robustness intervention improves over standard training.

Data Generation: We describe the methods that we use to generate data from Stable Diffusion in Appendix §E. Throughout the main text, we will focus on generating images by conditioning on the natural language prompts for the class labels. For example, we can condition the model with a prompt 'a photo of a [*dog*]' to generate images for the class label *dog*.

3. Experiments

3.1. Classification Accuracy and Robustness

In our experiments, we choose ImageNet-1K as the source real dataset, and ImageNet-Sketch, ImageNet-R, ImageNet-V2, and ObjectNet as the source of natural distribution shift (NDS) datasets. We train a wide variety of classifiers e.g., ResNext-101, on the real dataset and the generated dataset. More details of the setup are provided in the App. §F.

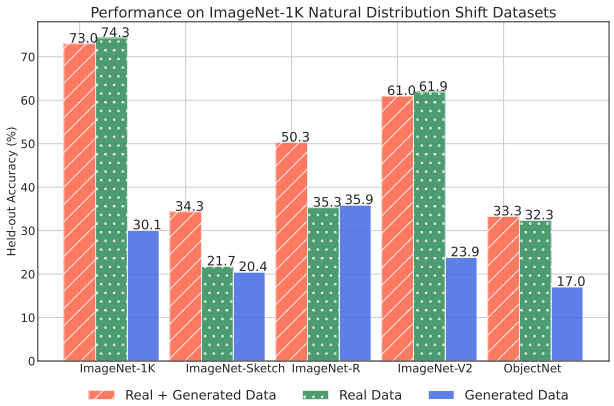


Figure 2. Accuracy of ImageNet-1K classifiers on ImageNet-1K validation set and its natural distribution shift datasets.

We train classifiers on the **real** ImageNet-1K dataset with 1.3M images, a **generated** dataset of 1.3M images created using Stable Diffusion conditioned on proxy captions for the class labels in ImageNet-1K, and a combination of all images from the **real and generated** training datasets.

Table 1. Comparison of the effective robustness of the classifiers trained solely on the generated dataset and on the real data augmented with the generated dataset.

	Im-Sketch	Im-R	Im-V2	ObjectNet	Average
Generated Data	**37.83**	**45.34**	**9.12**	**49.91**	**35.55**
Real + Generated Data	14.88	16.68	0.47	2.28	8.55

The average accuracy of five classifiers over three random seeds is shown in Figure 2. We find that models trained on the real ImageNet-1K (Im-1K) dataset (Green bar) perform well on its validation set but experience a significant drop in performance under natural shifts. Interestingly, we find that training on generated images using the same training dataset size leads to poor absolute performance on Im-1K as well as its NDS datasets. The low absolute performance may be due to the large distribution gap between the source and generated training datasets. However, we observe that the accuracy gaps performance on the real validation dataset and its NDS datasets are low, which might be attributed to the benefits of training on diverse generated data. Finally, we train the classifiers on an equal-sized combination of real and generated datasets to understand the effectiveness of generative augmentations.

Table 2. FID score averaged over the ImageNet classes between the real/generated data and the NDS datasets.

FID	Im-Sketch	Im-R	Im-V2	ObjectNet
Real Dataset	248	225	**179**	**224**
Generated Dataset	**210**	**190**	223	255

As seen in Figure 2, we find that absolute performance across the majority of the NDS datasets is higher than training solely on the real or generated dataset. Notably, training on the combination of the real and generated dataset does not affect performance on the ImageNet1K validation dataset compared to standard training (Orange and Green bar). We see a similar effect for the natural distribution dataset, ImageNet-V2, which is closest in distribution to ImageNet-1K since both the datasets are derived from Flickr30K (Recht et al., 2019). On ObjectNet, the gain is $\sim 1\%$, indicating the difficulty of this dataset. Surprisingly, we find that training with the combination of the real and generated data leads to an absolute improvement of $\sim 15\%$ on ImageNet-Sketch and ImageNet-R over standard training. Additionally, we find that the effective robustness (ER) of the classifier is higher (Table 1) than standard training (= 0) but lower than classifiers trained on the generated data (Row 1 and Row 2). We further compare the average FID scores (Table 2) between the real/generated data and the NDS datasets, and find that ImageNet-R/Sketch are closer to the generated data than real data, which might be attributed to the presence of rendition and sketch images in the generated data (App. §L), that eventually gets reflected as larger improvements on classification accuracy and ER on these datasets. For broader evaluation, we also show that training a classifier on the real data augmented with the generated data achieves high accuracy and ER on corruption-based datasets such as ImageNet-C (Hendrycks & Dietterich, 2019) (App. §G).

3.2. Comparison with Standard Augmentations

We examine the average performance of three classifiers (ResNet-18, ResNeXt-50, and ResNeXt-101) trained on the real ImageNet-100 dataset with 130K images, augmented with an equal number of generated images from Stable Diffusion, DeepAugment, PixMix, and class-conditional LDM on the set of overlapping classes with 4 NDS datasets in Table 3. We observe that augmenting with the diverse in-the-wild generated datasets yields the highest performance on Im-R, Im-Sketch, and ObjectNet, followed by DeepAugment, highlighting the utility of modern generative models

Table 3. Comparison of the models trained on real data and an equal mix of real data and generated data (100:100 ratio) using different augmentation strategies on ImageNet-100 validation set and its natural distribution shift (NDS) datasets. We report results over the classes that overlap with ImageNet-100. The results are averaged over three runs of ResNet-18, ResNeXt-50/101.

	Im-100	Im-Sketch-100	Im-R-100	Im-V2-100	Obj-100	Average
Real Data	85.7	28.4	49.8	74.8	42.3	56.2
+ DeepAugment (Hendrycks et al., 2021)	86.7	45.2	67.2	76.5	44.9	64.1
+ PixMix (Hendrycks et al., 2022)	85.3	32.7	56.6	73.7	43.9	58.5
+ Class Conditioned LDM (Rombach et al., 2022)	86.7	27.9	55.0	75.6	46.1	58.3
+ Stable Diffusion (Rombach et al., 2022) (Ours)	86.8	48.4	71.2	76.0	47.5	**66.0**

Table 4. Comparison of the classifiers on the original and filtered real and generated data. The accuracy gap between the performance is reported inside the gray brackets. We abbreviate Stable Diffusion as SD, Labels as L, Images as I, Pretraining as PT, & Finetuning as FT.

Models	**Original**		**Filtered**	
	Real	Generated	Real	Generated
ResNeXt-101 (Real ImageNet-1K)	79.3	55.9 (-23.4)	90.8	73.2 (-17.6)
ViT-L/14-336 (PT-Im12K-FT-Im1K) (Dosovitskiy et al., 2020)	**88.5**	66.2 (-22.3)	94.4	82.3 (-12.1)
Zero-shot CLIP-B/32 (Radford et al., 2021)	68.3	71.9 (+3.6)	83.1	85.6 (+2.5)
ResNeXt-101 (Real + Generated ImageNet-1K)	80.4	**89.0** (+8.6)	91.0	**97.0** (+6.0)

that are trained on larger multimodal datasets and allow for more flexible conditioning. We perform experiments to understand the effect of real and generated data size in App. §F.1 and the choice of generation strategy in App. §G.1.

3.3. Evaluating Classifiers on Generated Datasets

In our previous experiments, we showed that training a classifier on the **real** ImageNet data augmented with the **generated** data strikes a good balance between robustness and accuracy. Here, we evaluate the performance of a variety of supervised, zero-shot, and fine-tuned ImageNet classifiers on generated dataset, containing 50K generated images, similar to ImageNet-1K validation dataset. In Table 4, we find that all the classifiers, except for zero-shot CLIP, underperform on generated data while performing well on the ImageNet-1K validation dataset. The performance of the classifier trained on the mix of real and generated data (Row 6) highlights the potential for further improvements in the existing models on generated data. We perform detailed analysis in App. §N.1.

3.4. Data from Finetuned Stable Diffusion

Here, we aim to study the impact of varying the data generation paradigm and evaluate the quality of the image classifiers trained on the generated data that is closer in distribution to the real data as compared to the generated data collected in a zero-shot way. We observe that the accuracy gains over standard training on the natural distribution datasets are higher for the classifier trained on the real data augmented with the base-generated data as compared to the one trained on the real data augmented with the finetuned generated data. For example, the classifier trained on the real and base-generated data achieves an accuracy of 40.1% and 56.2% whereas the classifier trained on the real and finetuned-generated data achieves an accuracy of 56.2% and 41.5% on ImageNet-Sketch and ImageNet-R, respectively. Our finding further highlights the usefulness of training the classifiers on the diverse data, from the base generative model, over the generated data that is closer to the real data distribution, on natural distribution shift datasets. More details for the experiment are present in App. O.

Table 5. Comparison of the performance of a ResNext-50 classifier on the ImageNet-1K validation dataset, and its natural distribution shift datasets. The training data contains 1.3M examples for the Real, Base-Generated, and Finetune-Generated data. Here, Real + Base-Generated or Finetune-Generated indicates that the generated data is used to augment the real data.

Data	**Im**	**Im-Sketch**	**Im-R**	**Im-V2**	**ObjectNet**	**Average**
Real	78.4	25.0	42.2	68.5	40.6	51.0
Base-Generated	32.4	21.6	37.4	26.2	19.4	27.4
Finetune-Generated	38.1	9.4	18.4	28.0	16.7	22.1
Real + Base-Generated	78.4	40.1	56.2	66.5	39.4	**56.1**
Real + Finetune-Generated	78.0	28.2	41.5	66.0	37.5	50.2

4. Conclusion

We developed a framework to improve performance of image classifiers by augmenting real datasets with a diverse dataset generated by a modern 'in-the-wild' generative models. Our results show that classifiers trained with this method exhibit high performance on test and natural distribution shift datasets. This is due to the increased robustness obtained from training on generated data compared to standard training methods. Additionally, we used the synthetic data as an evaluation dataset and highlighted the brittleness of state-of-the-art models to natural variations in generated

images. Finally, we showed that the generated data from the base generative model has more practical usefulness for training robust classifiers as compared to the generated data from a finetuned generative model on the real data.

References

Antoniou, A., Storkey, A., and Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Bansal, H., Yin, D., Monajatipoor, M., and Chang, K.-W. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chambon, P., Bluethgen, C., Langlotz, C. P., and Chaudhari, A. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Cho, J., Zala, A., and Bansal, M. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

Cooper, D. Is dall-e's art borrowed or stolen? *https://www.engadget.com/dall-e-generative-ai-tracking-data-privacy-160034656.html*, 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.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Goel, S., Bansal, H., Bhatia, S., Rossi, R. A., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., and Steinhardt, J. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16783–16792, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. Matryoshka representations for adaptive deployment. *arXiv preprint arXiv:2205.13147*, 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 b444f0fa8c66bb5132af3ad6ec8db70fb94a3825.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

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.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 529–544. Springer, 2022.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ravuri, S. and Vinyals, O. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saha, A., Tejankar, A., Koohpayegani, S. A., and Pirsiavash, H. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13337–13346, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Scheuerman, M. K., Hanna, A., and Denton, E. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 464–472. IEEE, 2017.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Tomczak, J. M. *Deep generative modeling*. Springer, 2022.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

Udandarao, V., Gupta, A., and Albanie, S. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*, 2022.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

West, P., Bhagavatula, C., Hessel, J., Hwang, J. D., Jiang, L., Bras, R. L., Lu, X., Welleck, S., and Choi, Y. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Xu, X., Wang, Z., Zhang, E., Wang, K., and Shi, H. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022.

Yuan, J., Pinto, F., Davies, A., Gupta, A., and Torr, P. Not just pretty pictures: Text-to-image generators enable interpretable interventions for robust representations. *arXiv preprint arXiv:2212.11237*, 2022.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

A. Limitations

While the ability to train robust classifiers using generated data from modern text-to-image generative models represents a significant advancement in generative AI for trustworthy machine learning (ML), there are other equally important aspects, such as fairness and privacy, that have not been explored in this work. In this study, our focus is on highlighting the benefits of generated data for objects in the ImageNet-1K dataset. However, this raises intriguing questions about the generalizability of these results to larger datasets like ImageNet-21K (Ridnik et al., 2021).

Our approach primarily concentrates on generating data from the base generative model in a zero-shot manner for objects that are well-represented within its distribution. Nevertheless, it is crucial to fine-tune the base model for domains that are not adequately captured in its training distribution, such as medical images (Chambon et al., 2022). Despite these limitations, the core contributions of this paper remain highly valuable and provide crucial insights for promoting positive impact in trustworthy ML.

B. Ethics Statement

In our work, we utilize modern 'in the wild' generative models to create generated data, that is further employed for training Image classifiers. Since these generative models are trained on large, diverse, and uncurated web-scraped datasets, there are several privacy concerns surrounding the suitable use of public data (Scheuerman et al., 2021), and their harmful biases and stereotypes (Birhane et al., 2021; Bender et al., 2021). Once trained, these generative models can amplify these biases during generation (Saharia et al., 2022; Cho et al., 2022; Bansal et al., 2022). With the generative model's ability to create and combine different concepts in realistic ways, there are harms associated with changing the predictions based on the natural language descriptions of the concepts as it is much easier to generate objectionable content with these. It necessitates further research into closely curating the generated data as well as building fairer multimodal representations of the real world.

As generated data pervades the Internet, it is inevitable that they will be explicitly used or automatically scraped as training data for building new data-driven models, such as our work. These scenarios present a difficult challenge for researchers to better understand and track the source of harmful biases introduced in the dataset. Additionally, there are equally relevant privacy concerns as we train on the model generations, which in recent times, have been shown to replicate styles of real artists (Cooper, 2022). Hence, making the generated dataset publicly available is a step in the direction towards future benchmarking and critique of the design and use of generated datasets for trustworthy ML.

C. Related Work

Training Robust Classifiers: Many works propose hand-engineered augmentations to increase the training data and improve generalizability of the classifiers, e.g., (Hendrycks et al., 2019; 2022; Zhang et al., 2017). (Cubuk et al., 2018; 2020) learn augmentation policies directly from the data and have been shown to improve classification accuracy. DeepAugment (Hendrycks et al., 2021) was one of the first augmentation strategies to perform well on natural distribution shifts. Additionally, studies on CLIP-verse (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Goel et al., 2022; Mu et al., 2022) have shown natural robustness. In our work, we take the best of both paradigms by leveraging the strengths of modern generative models to augment real datasets. We find that classifiers trained with generated datasets are effectively robust and outperform current data augmentation strategies in eliciting robustness.

Robustness via Generated Data: (Gowal et al., 2021; Sehwag et al., 2021) studied the effectiveness of synthetic data from these models for creating adversarially robust classifiers, but did not examine the robustness in the regime of natural distribution shifts (NDS) and modern in-the-wild generative models (Rombach et al., 2022; Ramesh et al., 2021; Xu et al., 2022; Saharia et al., 2022; Balaji et al., 2022; Chang et al., 2023). (He et al., 2022) generates synthetic data using the GLIDE (Nichol et al., 2021) and finds that it improves the accuracy of the CLIP model (Radford et al., 2021), indicating the usefulness of synthetic data for pre-training image models. Our work focuses on the use-case of the generated data, created in a zero-shot manner, for training robust image classifiers against natural distribution shifts, and benchmarking the existing image classifiers.

Model Evaluation: Studies by (Recht et al., 2018; 2019; Hendrycks et al., 2021; Wang et al., 2019; Barbu et al., 2019) assess the model's ability to generalize to natural variations in images containing objects from the source dataset, showing severe performance dips and questioning their usefulness for real-world applications. In our work, we create a generated validation set from a modern generative model, containing new realizations of the objects in the ImageNet-1K dataset that

may be difficult to acquire in the real-world. We find that the state-of-the-art ImageNet classifiers experience a performance degradation on the generated validation data, highlighting at a gap that the robustness research should aim to bridge.

Augmenting with Generated Data: (Antoniou et al., 2017) used generated data to enhance the diversity of training data, leading to improved classification results, via an image-conditional GAN (Goodfellow et al., 2020). Since then, numerous studies have applied generated data in various domains. (West et al., 2021) generated a massive commonsense knowledge corpus using GPT-3 (Brown et al., 2020) to train commonsense models. Brooks et al. (Brooks et al., 2022) fine-tuned a stable diffusion model with a set of creative image-text pairs generated from a combination of GPT-3 and Stable diffusion for image editing. Our work demonstrates a practical application of using generated data for improved robustness in model training.

D. Detailed Background

D.1. Supervised Classification

Given a labelled dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \sim P(\mathbf{x}, y)$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathcal{R}^d$ represents the i^{th} input, and $y_i \in \mathcal{Y} \subset \{1, 2, \ldots, \mathcal{K}\}$ represents its corresponding target label, we train a classifier $\hat{f}(\mathbf{x})$ on $\mathcal{D}_{train} \subset \mathcal{D}$ such that it models $P(y|\mathbf{x})$, i.e., conditional distribution of y given the input $\mathbf{x}$. The classification model is usually trained via empirical risk minimization, $L(\hat{f}, \mathcal{D}_{train}) = \mathbb{E}\left[l(\hat{f}(\mathbf{x}_{train}), y_{train})\right]$, where l is the training objective, under the assumption that samples in the training data are identically and independently distributed (i.i.d.). Eventually, we evaluate the performance of the classifier on a held test set $D_{test} \subset \mathcal{D} \sim P$ with $\mathcal{D}_{test} \cap \mathcal{D}_{train} = \phi$) using *accuracy* $A(\hat{f}, \mathcal{D}_{test}) = \mathbb{E}\left[\mathbb{I}(\hat{f}(\mathbf{x}_{test}) = y_{test})\right]$.

If a classifier achieves high accuracy on the examples from the test set, we hope that it will perform well on the other examples that come from P as well as semantically related data distributions. However, in practice, we encounter test sets $\mathcal{D}'$ sampled from a data distribution P' that contains the samples resembling the ones in $\mathcal{D}$ with slight variations e.g., images in $\mathcal{D}'$ may vary from the images in the $\mathcal{D}$ in terms of differences in camera settings, and captured views.

D.2. Robustness

For any classifier, we can quantify the *accuracy gap* (AG) between the accuracy on a test set that follows the same distribution as the training set, and a test set that varies naturally from the training distribution.

$$AG(\hat{f}, \mathcal{D}_{test}, \mathcal{D}') = A(\hat{f}, \mathcal{D}') - A(\hat{f}, \mathcal{D}_{test}) \tag{3}$$

For a robust classifier, the accuracy gap should be low up to random sampling error. However, a classifier that closes the accuracy gap might decrease the individual accuracies. Additionally, given a robust classifier $\hat{f}$ that offers high accuracy on the shifted datasets, we can assess it relative to the expected accuracy on the shifted dataset with a standard classifier that is trained on the source training set without any specific robustness intervention. This notion is formalized as *effective robustness* (ER) (Recht et al., 2019; 2018).

$$ER(\hat{f}, \mathcal{D}') = A(\hat{f}, \mathcal{D}') - \beta(A(\hat{f}, \mathcal{D}_{test}), \mathcal{D}', \mathcal{D}_{test}) \tag{4}$$

where $\beta(z, \mathcal{D}', \mathcal{D}_{test})$ is the accuracy on the shifted test set $\mathcal{D}'$ for a given accuracy z on the source test set $\mathcal{D}_{test}$. We calculate β by fitting a linear function on the collection of standard classifiers. Positive ER indicates that the robustness intervention improves over standard training.

D.3. Generative Modeling

Generative models $p_\theta(\mathbf{x})$ are probabilistic models that are trained to learn the data distribution $p_{data}(\mathbf{x})$ (Tomczak, 2022). Due to their flexible design, we can further train their class-conditional versions (Brock et al., 2018; Karras et al., 2019) to model the class-conditional distributions $p(\mathbf{x}|y_g)$ where is y_g is the conditioning variable, that can take various forms, which we describe in next section. Post-training, we can generate a new sample $\mathbf{x}_g$ by sampling from the class-conditional model distribution $\mathbf{x}_g \sim p_\theta(\mathbf{x}|y_g)$. In Figure 1, this stochastic mapping $p_\theta(\mathbf{x}|y_g)$ is referred to as G. Thus, we can create a

generated dataset $\mathcal{D}_g = \{(\mathbf{x}_g, y_g)\}$ by repeatedly querying the conditional generative model.

E. Background - Data Generation using Stable Diffusion

In this work, we employ Stable Diffusion (SD) (Rombach et al., 2022), an 'in the wild' generative model is one that can generate images from the natural language description of a wide range of concepts, combine unrelated concepts in a realistic manner, and apply novel transformations to existing images. Such abilities are exhibited by Stable Diffusion through training on a large, diverse dataset LAION (Schuhmann et al., 2022) on matched image-text pairs $(\mathcal{X}, \mathcal{C})$ scraped from the web where $\mathbf{x} \subset \mathcal{X}$ denotes a raw image and $c \subset \mathcal{C}$ denotes its corresponding caption in natural language.

During training, the image $\mathbf{x}$ is passed through a pre-trained encoder $z_0 = \mathcal{E}(\mathbf{x})$ where z_0 is the latent representation of x. The objective of the denoising model $R(z_t, t, y_g)$ is to predict z_0 from every intermediate representation z_t where z_t is sampled from $t := 1, \ldots, T$ while the conditioning variable y_g guides the training process. For image generation, we sample from $z \sim N(0, I)$, and use the trained model $R(.)$ with a predefined sampling scheme (DDPM (Ho et al., 2020), DDIM (Song et al., 2020)) to reconstruct z_0 iteratively. Finally, the latent representation z_0 is decoded using the pretrained decoder $\mathbf{x}_g = \mathcal{D}(z_0)$ to generate the synthetic image $\mathbf{x}_g$.

Given a single data point $(\mathbf{x}, y)$ from the source dataset, we have various ways to generate a new data point $\mathbf{x}_g$ with a trained Stable Diffusion, as summarized in Appendix Figure 3.

Generation via Class Labels: In practice, Stable Diffusion uses CLIP's (Radford et al., 2021) text encoder $y_g = h_{text}(c)$ for conditioning during the training process. Here, we synthesize images by denoising $z_T \sim N(0, I)$ conditioned on the natural language templates $\mathcal{M}$ for the class labels y. An example template $M \subset \mathcal{M}$ includes 'A photo of a *dog*' where *dog* is the class label y. This generation strategy involves using a pretrained CLIP text encoder $y_g = h_{text}(M(y))$. Since generating data conditioned on the natural text descriptions is the default setting for data generation using Stable Diffusion, our primary focus is on the natural robustness elicited by this data generation strategy.

In addition to the traditional zero-shot data generation approach, we study the following other ways to generate images without any training or finetuning of the generative model on the images from the source dataset. We specifically study the effect of these data generation procedures in §G.1.

Generation via Real (Source) Images: Here, we use CLIP's vision encoder $y_g = h_{image}(\mathbf{x})$ for conditioning. In this case, we generate variations of the images from the source dataset by denoising the latent variable z_T conditioned on their representations.

Generation via Real (Source) Images and Class Labels: We can create realistic variations of the source image $\mathbf{x}$ by first encoding it using the pretrained encoder $\mathcal{E}(\mathbf{x})$ followed by forward diffusion for T steps to approximate a normal distribution $\hat{z}_T(\mathbf{x})$. Consequently, we generate a new image by denoising $\hat{z}_T(\mathbf{x})$ conditioned on the natural description of the class label $y_g = h_{text}(M(y))$.

F. Setup

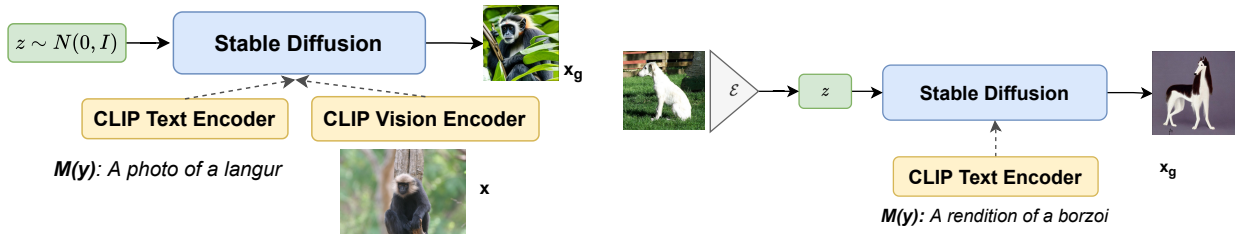
Real Dataset: The ImageNet-1K dataset is widely used as a benchmark for building robust classifiers for image recognition. It contains 1.3 million labeled training images and 50,000 validation images across 1000 categories. To evaluate the effectiveness of generated data in this task, we use ImageNet-1K as our benchmark. However, due to the limitations of compute and storage, we also utilize ImageNet-100, a subset of 100 classes randomly sampled from ImageNet-1K, for many of our analysis and ablation studies. In line with previous studies (Saha et al., 2022; Tian et al., 2020), we find that the trends observed in ImageNet-100 are similar to those in ImageNet-1K.

Natural Distribution Shift Datasets: Similar to the previous studies (Miller et al., 2021; Radford et al., 2021; Nguyen et al., 2022), we consider ImageNet as the reference dataset where ImageNet-Sketch (Wang et al., 2019), ImageNet-R (Hendrycks et al., 2021), ImageNet-V2 (Taori et al., 2020), and ObjectNet (Barbu et al., 2019) are natural distribution shift datasets.

Classifiers: We consider models with varying architectures and model capacities as classifiers. This includes ResNet-18 (He et al., 2016), ResNeXt-50, ResNeXt-101 (Xie et al., 2017), EfficientNet-B0 (Tan & Le, 2019) and MobileNet-V2 (Howard et al., 2017).

Data Generation: We utilize Stable Diffusion (Rombach et al., 2022) to generate synthetic data conditioned on the natural

descriptions of the objects in the dataset, and/or the training images. Specifically, we use the Stable Diffusion-V1-5 implementation and inference settings detailed in the diffusers (von Platen et al., 2022) library. For ImageNet-1K, we construct a 1.3M generated training dataset and 50K validation dataset from Stable Diffusion by conditioning on the proxy captions for the class labels. The proxy captions are a set of 80 diverse templates given by Radford et al. (2021) to evaluate their CLIP model.

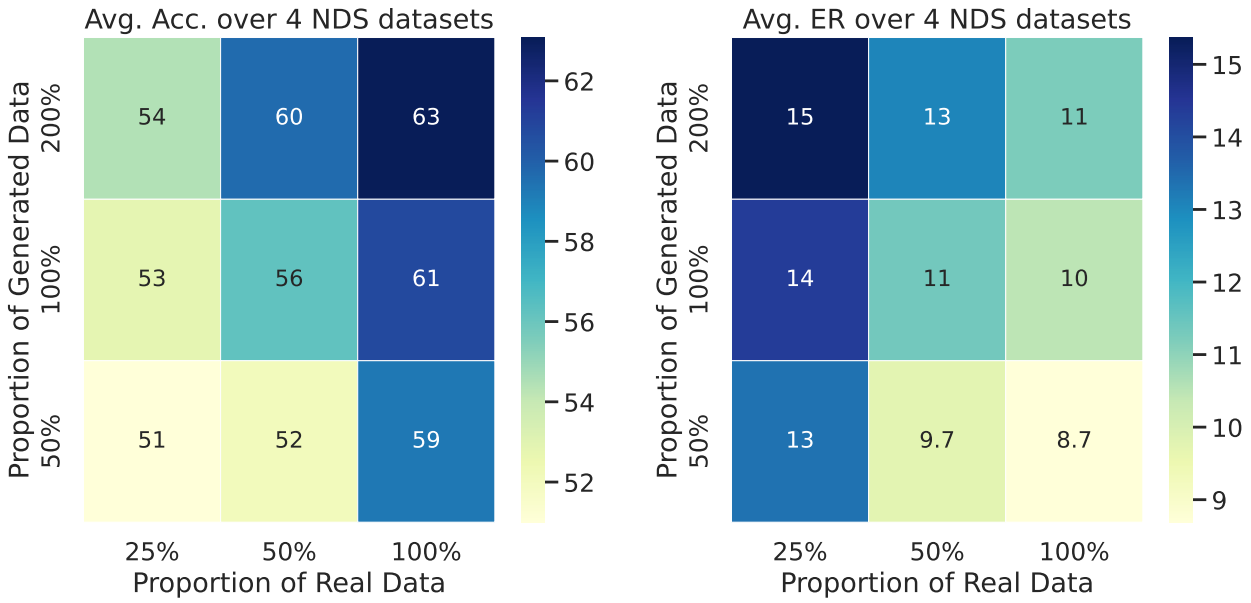


(a) Data Generation using either the proxy captions for class labels or the source image from the dataset.

(b) Data Generation using both the proxy captions and the source image from the dataset.

Figure 3. Overview of our generation strategies. We use Stable Diffusion (SD) to create the generated dataset. (a) We can create images by conditioning on either the proxy caption for the class label (Generation via Class Labels), or conditioning on the images from the source dataset (Generation via Real Images). (b) We can also generate data by first encoding the source images to get the latent representation, which is then denoised conditioned on the text prompt for the class label (Generation via Real Images and Class Labels).

F.1. Effect of Real and Generated Dataset Size



(a) Average Accuracy

(b) Average Effective Robustness

Figure 4. Variation in the accuracy and the effective robustness as we vary the proportion of the real ImageNet-100 data and the generated data created using its class labels in the training set. Here 100% refers to 130K training size. While calculating effective robustness, standard training is performed on 100% real data.

Here, we investigate how different combinations of the real dataset and the generated one can help the classifiers take advantage of the complementary strengths of the two data sources. To do this, we assessed the average performance of classifiers (ResNet-18, ResNext-50, and ResNext-101) trained with six different input mixing combinations created by using $25\%, 50\%, 100\%$ of the real data for ImageNet-100 and $50\%, 100\%, 200\%$ of the generated dataset using the class labels

from ImageNet-100.

As shown in Figure 4a, we observed an increase in accuracy on shifted datasets as the size of the real data increases while keeping the amount of generated data fixed. Similarly, when the proportion of the generated data increases while keeping the proportion of the real data fixed, we observed similar results. Overall, we found that increasing the amount of training data from either distribution leads to an improvement in performance on the shifted test beds.

In Figure 4b, we present the average effective robustness of the classifiers across NDS datasets. Interestingly, we observe that as the proportion of real data increases while keeping the amount of generated data fixed, the effective robustness of the classifier decreases. Conversely, as the proportion of generated data increases while keeping the amount of real data fixed, the effective robustness of the classifier increases.

Table 6. Comparison of the models trained on real data and an equal mix of real data and generated data (100:100 ratio) using different generation strategies on ImageNet-100 validation set and its natural distribution shift (NDS) datasets. We report results over the classes that overlap with ImageNet-100. The results are averaged over three runs. We abbreviate ImageNet as Im, and Class Label as CL.

	Im-100	Im-Sketch-100	Im-R-100	Im-V2-100	Obj-100	Average
Real data	85.7	28.6	49.8	74.8	42.3	56.2
+ Generated data via Class labels ('a photo of a [CL]' template)	87.4	35.7	59.5	75.6	44.9	60.6
+ Generated data via Class labels ('a rendition of a [CL]' template)	87.4	46.3	67.8	76.0	46.5	64.8
+ Generated data via Class labels (80 diverse templates)	86.8	48.4	71.2	76.0	47.5	**66.0**
+ Generated data via Real images	85.9	32.2	50.0	74.9	45.1	59.5
+ Generated data via Real images and Class labels	87.4	46.7	71.4	76.5	47.9	**66.0**

G. ImageNet-C

The evaluation datasets such as ImageNet-C intend to perturb the real images and distort their quality, such that the representations of the perturbed images are pushed outside the decision boundary of their true class ids. This differs from natural distribution shift datasets such as ImageNet-V2, ObjectNet, ImageNet-R, and ImageNet-Sketch, since these datasets are acquired under different environments in the real-world rather than formed by perturbing the original datasets themselves. To understand the usefulness of the generated data for ImageNet-C, we provide the results for the absolute accuracy and effective robustness of the models on ImageNet-C (Severity-5). We report the average accuracy over all the sub-datasets in the ImageNet-C, in Table 7.

Table 7. Comparison of training ImageNet-1K classifiers on the real data, generated data, and the equal mix of real and generated data, on ImageNet-C (Severity = 5) validation datasets.

Method	Accuracy (%)	Effective Robustness (%)
Real Data	20.5	-
Generated Data	3.3	**25.5**
Real Data + Generated Data	**21.75**	1.3

We find that the classifiers trained with solely the generated data as well as the mix of real and generated achieve high effectiveness robustness over standard training on the real data (Column 2). The absolute accuracy increases by 1.25% on the validation set of the ImageNet-1K using our augmentation.

G.1. Effectiveness of Generation Strategies

In the previous sections, we focused on generated data using 80 diverse templates with class label information from the ImageNet datasets. Here, we compare the performance of the classifiers that are trained on the real data augmented with the generated data created through mechanisms i.e., (a) diverse templates for class labels, (b) single template for class labels such as '*a photo of a **class label***', (c) real (source) images used for conditioning the generative model, and (d) real (source) images are first encoded and then denoised conditioned on the class labels.

We report the results for ImageNet-100 in Table 6. We find that the performance on training with synthetic dataset generated

using diverse templates for class labels, or the one generated using both class labels and source images, are closely tied at $\sim 66\%$. We observe that there is no additional benefit of using source domain information over just using the class labels information for zero-shot data generation from the modern generative models. This is different from previous works (Trabucco et al., 2023) which learns an optimized conditioning embedding from the source data to reduce the domain gap.

Further, we observe that training on the generated datasets created solely with single templates while utilizing class information results in lower robustness than training on images created via diverse templates. Interestingly, we find that the classifiers trained with images generated via a single template 'a photo of a [class label]', which does not prompt the model to generate either sketches or renditions explicitly, significantly outperform the classifiers trained solely on the real data (Row 1 and Row 2). This indicates that in some cases the classifiers augmented with the generated data can be robust to specific domains without any customization during data generation. Though we lack the resources for this type of study, future work should perform large-scale human evaluations for the generated datasets along these dimensions.

H. Generated Data Analysis

Table 8. Comparison of consistency (0-1) and quality (1-5) between the real images and the synthetic images created using various generation images. The numbers are averaged over the individual scores of the 20 human annotators.

	Real	Generated (Class Labels)	Generated (Real Images)	Generated (Real and Class Labels)
Consistency (Humans)	**0.96**	0.86	0.54	0.85
Quality (Humans)	**4.52**	4.2	2.96	3.8
Diversity (CLIP)	**0.30**	0.26	0.32	0.23

Since the generative model is prompted in a *zero-shot* manner, it is important to compare the consistency, quality, and diversity of the generated data with the real data. To do so, we perform a human evaluation study to assess whether there is a lack of useful information in the generated datasets that might be relevant to classify an object (Consistency), and whether the generated images are of poor quality i.e., they lack sharpness or contain perceptible noise (Quality). We collect 1600 annotations from 20 human surveyors for 40 images that are sampled from different real/generated datasets from 10 ImageNet classes. Further details on the data collection process are presented in Appendix §I. In addition, we compare the diversity in the real and generated dataset by subtracting the average of 1 - mean cosine pair-wise similarities between the CLIP representations of the images within each class of ImageNet-100, as done in (Udandarao et al., 2022).

We find that images belonging to the real ImageNet dataset are more consistent, of higher quality and more diverse than generated data created by conditioning a modern generative model on the natural descriptions of the class labels. This is expected since the real ImageNet went through extensive data curation and cleaning process during its creation. Since the scores for the generated data via class labels are not that far off, it provides further evidence for its effectiveness and potential training robust classifiers. In addition, we observe that the consistency and quality scores of images generated via class labels and the ones generated via source images and class labels are close. In terms of the diversity, we observe that data generation using only source image information led to the most diverse creations within each class. However, we also find that synthetic data generated using just the source images had low consistency and quality scores, suggesting at the poor object representations and image quality, which do not aid in robustness to natural distribution shifts.

I. Setup for Human Evaluation

We randomly selected images from 10 classes of the ImageNet1K dataset and used them to synthesize generated images using three different strategies: generated data via class labels, via real (source) images, and a combination of both, as described in §G.1. This resulted in a total of 40 images for our study, including the real images. We then recruited a pool of 20 human annotators to independently complete a survey in which they were shown each image without any information about its source. [1] They were asked two questions for each image: 1) whether the image contained the intended class label, and 2) to rate the image's quality on a scale of 1-5. The screenshot of the survey for one image is provided for reference in Figure 5.

[1] More details will be made public in the camera ready version.

Figure 5. Survey screenshot

J. Setup for Training Image Classifiers

As suggested in previous studies (Kusupati et al., 2022), we train all the models using the efficient dataloaders of FFCV (Leclerc et al., 2022). We train the models for 40 epochs with the batch size of 512 on ImageNet-1K, and for 88 epochs with the batch size of 512 on ImageNet-100. All the models are trained with a learning rate of 0.5 with a cyclic learning rate schedule (Smith, 2017). All the models are trained with SGD optimizer with a weight decay of 5e-5.

K. More Details on Natural Distribution Shift Datasets

ImageNet-Sketch contains the sketches of ImageNet-1K objects. ImageNet-R contains the renditions (paintings, sculptures) for 200 ImageNet-1K classes, 19 of which overlap with ImageNet-100. ImageNet-V2 is a reproduction of ImageNet-1K validation dataset, and we consider its matched frequency variant that closely follows the ImageNet-1K data distribution. Finally, ObjectNet contains a objects in novel backgrounds and rotations with 113 overlapping classes with ImageNet-1K, and 13 classes overlapping with ImageNet-100.

L. Templates used for Data Generation

We present the list of 80 diverse templates that were used to generate the new images in Table 9.

M. Visualization of Image Generations

We present a sample visualizations of the images generated via different generated strategies in Figure 6.

N. Effect of changing the training size

We present the effect of variation in the training size along the dimensions of the training data and the generated data in Figure 7, 8, 9, 10.

	'a bad photo of a {class label}.', 'a photo of many {class label}.', 'a sculpture of a {class label}.', 'a photo of the hard to see {class label}.', 'a low resolution photo of the {class label}.', 'a rendering of a {class label}.', 'graffiti of a {class label}.', 'a bad photo of the {class label}.', 'a cropped photo of the {class label}.', 'a tattoo of a {class label}.', 'the embroidered {class label}.', 'a photo of a hard to see {class label}.', 'a bright photo of a {class label}.', 'a photo of a clean {class label}.', 'a photo of a dirty {class label}.', 'a dark photo of the {class label}.', 'a drawing of a {class label}.', 'a photo of my {class label}.', 'the plastic {class label}.', 'a photo of the cool {class label}.', 'a close-up photo of a {class label}.', 'a black and white photo of the {class label}.', 'a painting of the {class label}.', 'a painting of a {class label}.', 'a pixelated photo of the {class label}.', 'a sculpture of the {class label}.', 'a bright photo of the {class label}.', 'a cropped photo of a {class label}.', 'a plastic {class label}.', 'a photo of the dirty {class label}.', 'a jpeg corrupted photo of a {class label}.', 'a blurry photo of the {class label}.', 'a photo of the {class label}.', 'a good photo of the {class label}.', 'a rendering of the {class label}.', 'a {class label} in a video game.', 'a photo of one {class label}.', 'a doodle of a {class label}.', 'a close-up photo of the {class label}.', 'a photo of a {class label}.', 'the origami {class label}.', 'the {class label} in a video game.', 'a sketch of a {class label}.', 'a doodle of the {class label}.', 'a origami {class label}.', 'a low resolution photo of a {class label}.', 'the toy {class label}.', 'a rendition of the {class label}.', 'a photo of the clean {class label}.', 'a photo of a large {class label}.', 'a rendition of a {class label}.', 'a photo of a nice {class label}.', 'a photo of a weird {class label}.', 'a blurry photo of a {class label}.', 'a cartoon {class label}.', 'art of a {class label}.', 'a sketch of the {class label}.', 'a embroidered {class label}.', 'a pixelated photo of a {class label}.', 'itap of the {class label}.', 'a jpeg corrupted photo of the {class label}.', 'a good photo of a {class label}.', 'a plushie {class label}.', 'a photo of the nice {class label}.', 'a photo of the small {class label}.', 'a photo of the weird {class label}.', 'the cartoon {class label}.', 'art of the {class label}.', 'a drawing of the {class label}.', 'a photo of the large {class label}.', 'a black and white photo of a {class label}.', 'the plushie {class label}.', 'a dark photo of a {class label}.', 'itap of a {class label}.', 'graffiti of the {class label}.', 'a toy {class label}.', 'itap of my {class label}.', 'a photo of a cool {class label}.', 'a photo of a small {class label}.', 'a tattoo of the {class label}.'

Table 9. List of diverse templates used for generating data.

N.1. Evaluating Classifiers on Generated Datasets

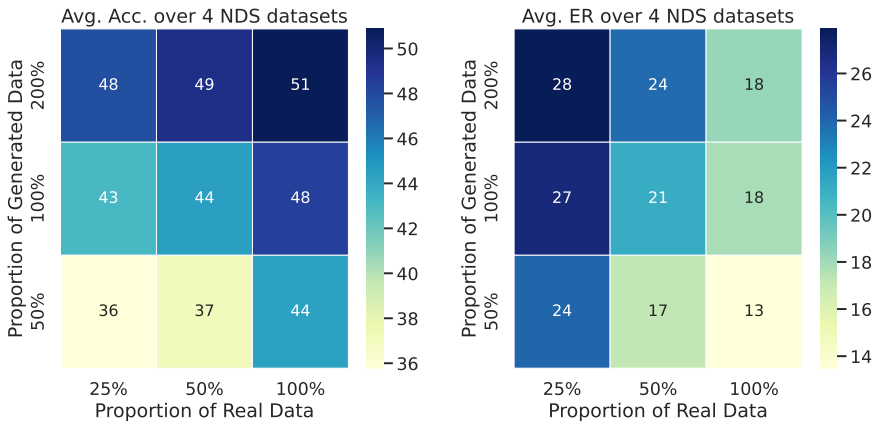
In the past sections, we established a case for using the generated data for training robust classifiers. However, the generated data can also be utilized for guiding the creation of robust image classifiers. To that end, we compare the performance of a diverse set of classifiers, (a) ResNeXt-101 trained solely on the real ImageNet-1K (ImageNet-1K), (b) ViTs pretrained on a larger set of ImageNet categories (ImageNet-21K/12K) and finetuned on ImageNet-1K, (c) Zero-shot CLIP, (d) CLIP finetuned on the real ImageNet-1K dataset, in Table 10. We report the results of the classifiers on the original real/generated datasets, and their filtered versions that are constructed by removing all the images whose cosine similarity score with their class label's proxy caption ('a photo of a {class label}') is less than 0.3, as done in (Schuhmann et al., 2022).

Table 10. Comparison of different classifiers on the original and filtered real and generated data. The accuracy gap between the performance is reported inside the gray brackets. We abbreviate Stable Diffusion as SD, Labels as L, Images as I, Pretraining as PT, & Finetuning as FT.

Models	**Original**		**Filtered**	
	Real	Generated	Real	Generated
ResNeXt-101 (Real ImageNet-1K)	79.3	55.9 (-23.4)	90.8	73.2 (-17.6)
ViT-L/14-336 (PT-Im12K-FT-Im1K) (Dosovitskiy et al., 2020)	**88.5**	66.2 (-22.3)	94.4	82.3 (-12.1)
MaxViT-XL-512 (PT-Im21K-FT-Im1K) (Tu et al., 2022)	88.3	68.6 (-19.7)	**94.5**	79.9 (-14.6)
Finetuned CLIP-B/32 (Real ImageNet-1K) (Wortsman et al., 2022)	81.3	64.1 (-17.2)	90.7	78.4 (-12.3)
Zero-shot CLIP-B/32 (Radford et al., 2021)	68.3	71.9 (+3.6)	83.1	85.6 (+2.5)
ResNeXt-101 (Real + Generated ImageNet-1K)	80.4	**89.0** (+8.6)	91.0	**97.0** (+6.0)

Despite performing the best on ImageNet-1K validation datasets, ViTs underperform on the generated data. We further find that the CLIP finetuned on ImageNet-1K experiences a performance degradation of upto 17%, 12% absolute accuracy on the original and filtered datasets respectively. However, we find that zero-shot CLIP does not undergo a distribution shift on the generated data. Since the zero-shot CLIP encoders are used as module in our data generator Stable Diffusion, the good performance of CLIP on the generated dataset underscores a "cyclic consistent" nature where the conditional generations of an encoder-decoder generative model (Stable Diffusion) agree with the standalone encoders in CLIP. To better quantify the

Figure 6. Visualization of samples from the real dataset and various generation strategies using Stable Diffusion (SD).



(a) Accuracy (b) Effective Robustness

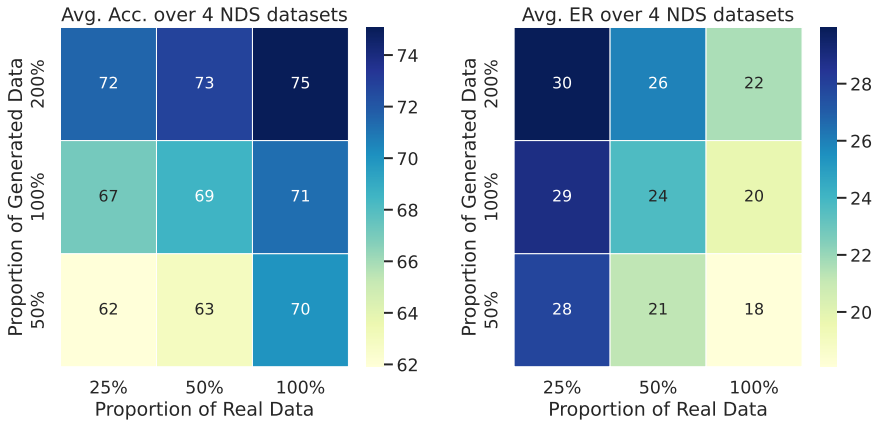
Figure 7. Variation in the accuracy and the effective robustness on ImageNet-Sketch as we vary the proportion of the real ImageNet-100 data and the generated data created using its class labels in the training set. Here 100% refers to 130K training size. While calculating effective robustness, standard training is performed on 100% real data.

performance gap on the generated data, we evaluate the performance of a classifier trained on the combination of the real and generated data. We observe that the classifier achieves upto 89%, 97% on the real and generated data, respectively, which highlights the potential for further improvements of the existing models on the novel realizations of the ImageNet objects.

O. Generated Data from Finetuned Stable Diffusion

In our work, we showed that the classifiers trained on the real data augmented with the generated data, acquired in a zero-shot manner from the base generative model, are robust to natural distribution shifts. Here, we aim to study the impact of varying the data generation paradigm and evaluate the quality of the image classifiers trained on the generated data that is closer in distribution to the real data as compared to the generated data collected in a zero-shot way.

To this end, we finetune the base Stable Diffusion v1.5 for 1 epoch on the complete 1.3M (real) ImageNet-1K data and their corresponding class labels, at the default resolution of 512 x 512. Post-finetuning, we repeatedly query the generative model conditioned on the class labels to synthesize a newly generated data of the same size as ImageNet-1K training and validation datasets. Finally, we train ResNext-50 classifier (a) solely on the newly generated data, and (b) an equal mix of real data and newly generated data, from the finetuned Stable Diffusion. In Table 11, we compare the performance of the same classifier trained with the (a) real data, (b) generated data from the base generative model conditioned on the class labels, and (c) an equal mix of the real and base generated data, on the real ImageNet-1K test set and its natural distribution shift datasets.



(a) Accuracy (b) Effective Robustness

Figure 8. Variation in the accuracy and the effective robustness on ImageNet-R as we vary the proportion of the real ImageNet-100 data and the generated data created using its class labels in the training set. Here 100% refers to 130K training size. While calculating effective robustness, standard training is performed on 100% real data.

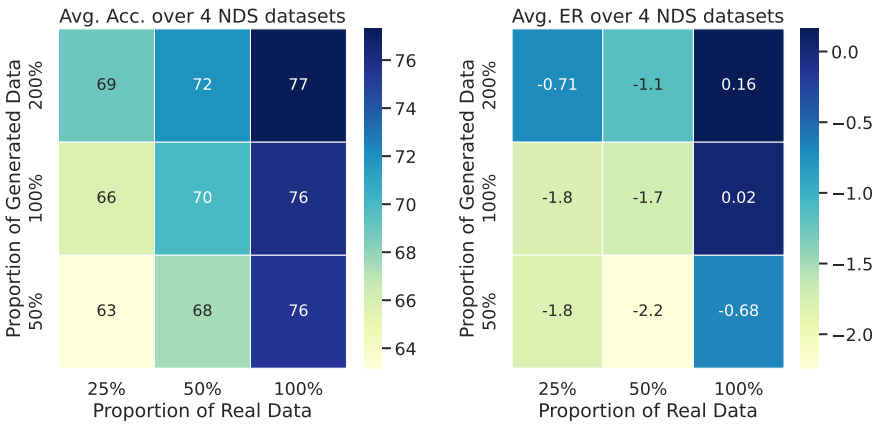
Table 11. Comparison of the performance of a ResNext-50 classifier on the ImageNet-1K validation dataset, and its natural distribution shift datasets. The training data contains 1.3M examples for the Real, Base-Generated, and Finetune-Generated data. Here, Real + Base-Generated or Finetune-Generated indicates that the generated data is used to augment the real data.

Data	ImageNet	ImageNet-Sketch	ImageNet-R	ImageNet-V2	ObjectNet	Average
Real	78.4	25.0	42.2	68.5	40.6	51.0
Base-Generated	32.4	21.6	37.4	26.2	19.4	27.4
Finetune-Generated	38.1	9.4	18.4	28.0	16.7	22.1
Real + Base-Generated	78.4	40.1	56.2	66.5	39.4	**56.1**
Real + Finetune-Generated	78.0	28.2	41.5	66.0	37.5	50.2

We find that the image classifiers trained with solely the finetuned-generated data (Row 3) outperform the one trained with the base-generated data (Row 2) on the ImageNet-1K validation dataset. This is due to the reduction in the distribution gap between the real data and the generated data from the finetuned Stable Diffusion model. We note that the accuracy achieved by the classifiers trained on the finetuned Stable Diffusion i.e., 38.1% lags behind the accuracy achieved in (Azizi et al., 2023) by training on the generated data from the finetuned ImaGen model i.e., 67%. We attribute this difference in the accuracies to the differences in the quality of the base generative models themselves.

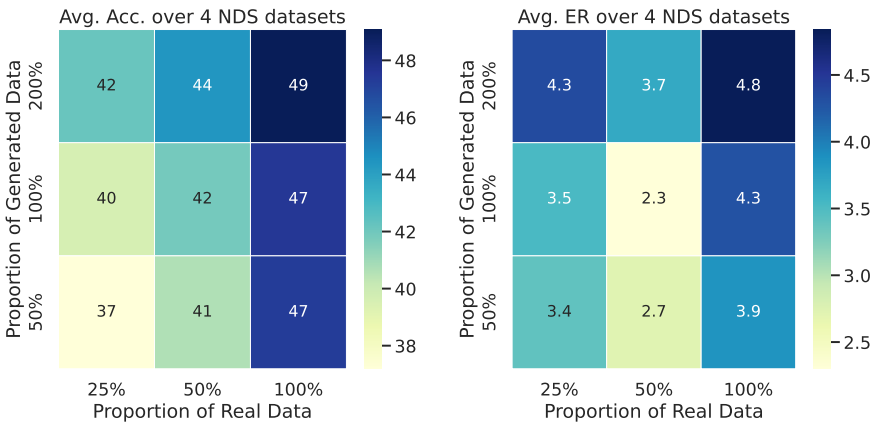
Despite the reduction in the domain gap between the real data and generated data via finetuning, we find that the ImageNet-1K validation accuracy for the classifier trained on the real data augmented with the finetuned generated data 78% (Row 5) is close to the one trained on the real data augmented with the generated data from the base model 78.4% (Row 4). Although our observation may surprising, we find that similar observations were made in Table 4 in (Azizi et al., 2023) and Figure 5 in (Ravuri & Vinyals, 2019) at high resolutions. The exact reason behind this empirical finding is still unclear, and a potential future work.

Lastly, we observe that the accuracy gains over standard training on the natural distribution datasets are higher for the classifier trained on the real data augmented with the base-generated data as compared to the one trained on the real data augmented with the finetuned generated data. For example, the classifier trained on the real and base-generated data achieves an accuracy of 40.1% and 56.2% whereas the classifier trained on the real and finetuned-generated data achieves an accuracy of 56.2% and 41.5% on ImageNet-Sketch and ImageNet-R, respectively. Our finding further highlights the usefulness of training the classifiers on the diverse data, from the base generative model, over the generated data that is closer to the real data distribution, on natural distribution shift datasets.



(a) Accuracy (b) Effective Robustness

Figure 9. Variation in the accuracy and the effective robustness on ImageNet-V2 as we vary the proportion of the real ImageNet-100 data and the generated data created using its class labels in the training set. Here 100% refers to 130K training size. While calculating effective robustness, standard training is performed on 100% real data.



(a) Accuracy (b) Effective Robustness

Figure 10. Variation in the accuracy and the effective robustness on ObjectNet as we vary the proportion of the real ImageNet-100 data and the generated data created using its class labels in the training set. Here 100% refers to 130K training size. While calculating effective robustness, standard training is performed on 100% real data.