LEAVING REALITY TO IMAGINATION: ROBUST CLASSIFICATION VIA GENERATED DATASETS

Hritik Bansal
UCLA
hbansal@ucla.edu

Aditya Grover
UCLA
adityag@cs.ucla.edu

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, paintings, and animations of the object categories observed during training. Prior work focuses on reducing this gap by designing engineered augmentations of training data or through unsupervised pretraining of a single large model on massive in-the-wild training datasets scraped from the Internet. However, the notion of a dataset is also undergoing a paradigm shift in recent years. With drastic improvements in the quality, ease-of-use, and access 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. Further, we introduce and analyze an evolving generated dataset, ImageNet-G-v1, to better benchmark the design, utility, and critique of standalone generated datasets for robust and trustworthy machine learning.

1 INTRODUCTION

The ultimate goal of machine learning is to create models that can generalize beyond their training data. However, recent studies Recht et al. (2019); Hendrycks et al. (2021); Wang et al. (2019); Barbu et al. (2019); Taori et al. (2020) have shown a gap between the performance of deep neural classifiers on test data that is independent and identically distributed (i.i.d.) as the training data, and *shifted* datasets containing natural variations of the images in the training distribution. 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 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). 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. Notably, these models can be repeatedly queried to generate diverse data through various conditioning mechanisms such as text prompts, images, and guidance strategies. Given this emerging paradigm, we study the question: *How do these generated datasets influence the natural robustness of image classifiers?*

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 2 source datasets ImageNet-1K Deng et al. (2009) and ImageNet-100 Tian et al. (2020). We utilize the flexibility of the model's design to generate diverse data conditioned on the proxy prompts for classes. By repeatedly sampling from Stable Diffusion using a specific conditioning strategy, we generate enough samples to create a large, diverse synthetic training dataset. For instance, we generate 1.3M

images using proxy prompts for the classes in ImageNet-1K, which is the same size as the real ImageNet-1K training dataset.

Our main takeaway is that training a classifier on a combination of real and generated data can achieve high absolute performance on natural distribution shift datasets and high effective robustness (§3.1). Removing either of real or generated data results in a corresponding reduction in accuracy and effective robustness respectively, thus necessitating the use of a mixture. To further explain these 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). Previous works such as He et al. (2022) generate synthetic data using the GLIDE Nichol et al. (2021) and find that it improves the accuracy of the CLIP model Radford et al. (2021) on image classification. However, we train neural classifiers from scratch, since CLIP can leverage its pretrained knowledge from 400M data for classification, which makes it hard to disentangle the effect of generated data for improved accuracy. Additionally, we perform a detailed analysis of the effectiveness of generated data for robust classification, specifically focusing on natural distribution shifts.

With the consistent rise of generative AI, we also expect massive amounts of generated data to permeate the Internet Wang et al. (2022) and directly be used for training ML models, such as in our work. While this can result in both positive and negative outcomes for trustworthy machine learning Jahanian et al. (2021); Saha et al. (2022b); Zhang et al. (2022); Jain et al. (2022); Cooper (2022); Cho et al. (2022); Bansal et al. (2022), it is worth noting that there is a lack of any benchmarking and critique process for such generated datasets. As a final contribution, we introduce ImageNet-G-v1 (or ImageNet-G) as a evolving benchmark dataset (§4) to advance research in trustworthy ML.

2 BACKGROUND

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

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 $\mathcal{D}'$.

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

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).

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

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 §2. 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*. Results for the other generation methods are relegated to the Appendix §C.4.

3 EXPERIMENTS

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 experimental setup are provided in the Appendix §D.

3.1 Classification Accuracy and Robustness

We train on 3 kinds of datasets: 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.

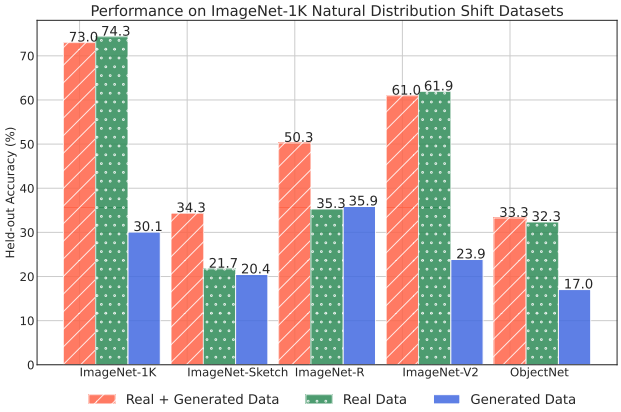


Figure 1: Accuracy of ImageNet-1K classifiers on ImageNet-1K validation set and its natural distribution shift datasets.

The average accuracy of five classifiers over three random seeds is shown in Figure 1. 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 (32%) 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.

As seen in Figure 1, 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 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. §I), 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. §E). In summary, generated data alone increases the effective robustness at the cost of accuracy, whereas an augmented mixture of real and generated data strikes a good balance for robust and accurate training.

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.

	ImageNet-Sketch	ImageNet-R	ImageNet-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

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

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

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.

	Im-100	Im-Sketch-100	Im-R-100	Im-V2-100	Obj-100	Average
Real Data	85.67	28.64	49.76	74.83	42.26	56.23
+ DeepAugment Hendrycks et al. (2021)	86.73	45.15	67.2	**76.50**	44.94	64.10
+ PixMix Hendrycks et al. (2022)	85.33	32.71	56.60	73.69	43.92	58.45
+ Class Conditioned LDM Rombach et al. (2022)	86.69	27.88	54.99	75.57	46.12	58.25
+ Stable Diffusion Rombach et al. (2022)	86.79	**48.40**	**71.23**	75.96	**47.49**	**65.97**

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 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. §E.1, choice of conditioning in App. §E.2.1, and perform human evaluation of generated data in App. §E.2.2.

4 Benchmarking Generated Datasets

In our previous experiments §3, we showed that training a classifier on the **real** ImageNet data augmented with the **generated** ImageNet-G 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 ImageNet-G 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, underperforms on ImageNet-G while performing well on 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 ImageNet-G. We perform detailed analysis in App. §E.3.

Table 4: Comparison of different classifiers on the ImageNet-G variants. The accuracy gap between Im-1K and the Im-G variant is reported inside the gray brackets.

Models	ImageNet-1K	ImangeNet-1K-G
ResNeXt-101 (Trained on ImageNet1K)	79.28	55.93 (-23.3)
ViT-L/14-336 (PT-Im12K-FT-Im1K) Dosovitskiy et al. (2020)	**88.54**	66.24 (-22.30)
MaxViT-XL-512 (PT-Im21K-FT-Im1K) Tu et al. (2022)	88.26	68.61 (-19.65)
Zero-shot CLIP (ViT-B/32) Radford et al. (2021)	68.33	71.89 (+3.56)
Finetuned CLIP (ViT-B/32) Wortsman et al. (2022)	81.31	64.05 (-17.26)
ResNeXt-101 (Trained on real and generated data)	80.39	**89.04** (+8.65)

5 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. We introduced ImageNet-G-v1, an evaluation dataset that highlights the brittleness of state-of-the-art models to natural variations in images. An important future direction and a current limitation is evaluating the trustworthiness of generated data based solely on robustness. Future research should incorporate a multi-dimensional analysis, including factors such as privacy and the presence of harmful stereotypes.

REFERENCES

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

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

Daniel Cooper. Is dall-e's art borrowed or stolen? *https://www.engadget.com/dall-e-generative-ai-tracking-data-privacy-160034656.html*, 2022.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

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

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16783–16792, 2022.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

Anubhav Jain, Nasir Memon, and Julian Togelius. A dataless faceswap detection approach using synthetic images. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–7. IEEE, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Tero Karras, Samuli Laine, and Timo Aila. 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.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representations for adaptive deployment. *arXiv preprint arXiv:2205.13147*, 2022.

Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. FFCV: Accelerating training by removing data bottlenecks. `https://github.com/libffcv/ffcv/`, 2022. commit b444f0fa8c66bb5132af3ad6ec8db70fb94a3825.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 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.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 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.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13337–13346, 2022a.

Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Ganorcon: Are generative models useful for few-shot segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9991–10000, 2022b.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

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

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

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

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

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

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

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, 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.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022.

Jianhao Yuan, Francesco Pinto, Adam Davies, Aarushi Gupta, and Philip Torr. Not just pretty pictures: Text-to-image generators enable interpretable interventions for robust representations. *arXiv preprint arXiv:2212.11237*, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *arXiv preprint arXiv:2211.13976*, 2022.

A 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, introducing an evolving dataset, ImageNet-G, and making it publicly available is a step in the direction towards future benchmarking and critique of the design and use of generated datasets for trustworthy ML.

B Related Work

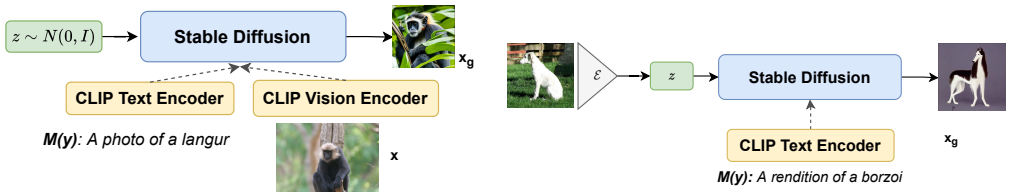
Training Robust Classifiers: Many works propose hand-engineered augmentations to increase the training data and improve generalizability of the classifiers, e.g., Hendrycks et al. (2019; 2022); Zhang et al. (2017). Cubuk et al. (2018; 2020) learn augmentation policies directly from the data and have 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) 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. However, we perform a detailed analysis of the effectiveness of generated data for robust classification, specifically focusing on NDS. Yuan et al. (2022) adapt to the target domain by training on a generated dataset using the variations of the images in the source domain. In contrast, our work does not make assumptions about the target domain and does not require access to source images to train robust classifiers. Our study further includes experiments on the effectiveness of various generation strategies to elicit robustness to NDS for larger datasets such as ImageNet.

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 propose a dataset, ImageNet-G-v1, containing new realizations of the objects in the ImageNet-1K dataset that may be difficult to acquire in the real-world. With advancements in generative modeling, we can synthesize more novel, consistent, and high-quality images that can be integrated into ImageNet-G as an evolving benchmark dataset.

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.



(a) Data Generation using either the proxy captions for class labels (*SD-Labels*) or the source image from the dataset (*SD-Images*).

(b) Data Generation using both the proxy captions and the source image from the dataset (*SD-Labels&Images*).

Figure 2: 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, which we refer as *SD-Labels*, or conditioning on the images from the source dataset which is referred as *SD-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. We refer to this strategy as *SD-Labels&Images*.

C Detailed Background

C.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 ith 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.

C.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).

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

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

C.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.

C.4 Data Generation using Stable Diffusion

In this work, we employ Stable Diffusion (SD) Rombach et al. (2022), a conditional probabilistic latent space diffusion model that learns the data distribution by denoising the normally distributed variable $z_T \sim N(0, I)$ for a finite number of denoising steps T. Specifically, it is a text-to-image generative model that can create novel images conditioned on their natural language descriptions.

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. Under a noise schedule $p(\epsilon)$, a small amount of normally distributed gaussian noise ϵ_t is added to z_0 for finite steps T, also referred to as *forward diffusion*, until the representation converges to a normal distribution $z_T \sim N(0, I)$. The objective of the denoising model $R(z_t, t, y_g)$ is to predict z_0 from every intermediate representation z_t where $t := 1, \ldots, T$ such that the conditioning variable y_g guides the training process. In practice, Stable Diffusion uses CLIP's Radford et al. (2021) text encoder $y_g = h_{text}(c)$ for conditioning during the training process. We can sample from $z_T \sim N(0, I)$ and use the trained model $R(.)$ to reconstruct z_0 that is eventually decoded using a pretrained decoder $\mathbf{x}_g = \mathcal{D}(z_0)$.

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 Figure 2.

Generation w/ Class Labels (SD-Labels): We can 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 $y_g = h_{text}(M(y))$.

Generation w/ Source Images (SD-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 w/ Labels and Images (SD-Labels and Images): 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 can generate a new image by denoising $\hat{z}_T(\mathbf{x})$ conditioned on the natural description of the class label $y_g = h_{text}(M(y))$.

D 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. (2022a); 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. We provide more further description about these datasets in Appendix §H.

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). We provide further details on training them in Appendix §G.

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 (Appendix Table 11). It took us $\sim$10 days to generate the complete dataset on 5 Nvidia RTX A5000 GPUs with a batch size of 12 per GPU. Additionally, we generate a separate training dataset of 130K images and a validation dataset of 5K images for every generation strategy described in §C.4. We present some sample generations in Appendix Figure 6.

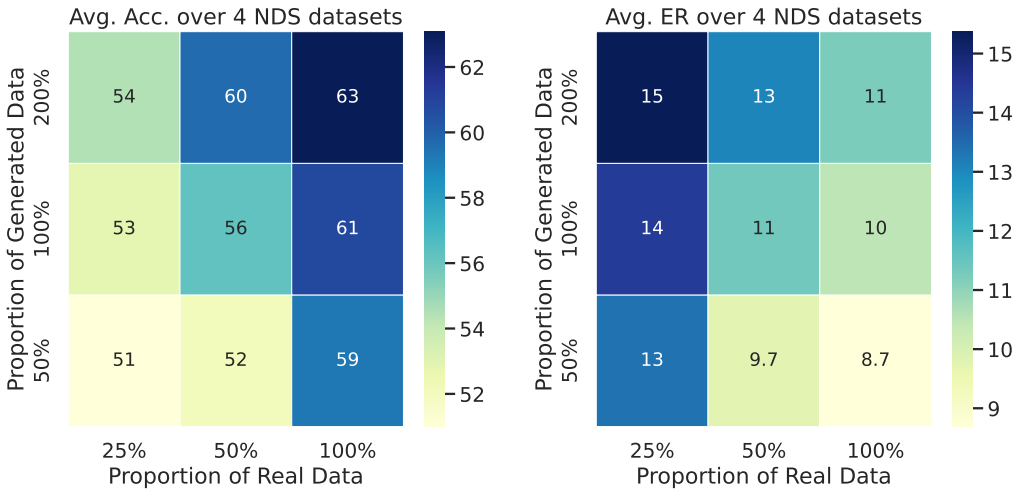
E IMAGENET-C

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 5.

Table 5: 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.



(a) Average Accuracy

(b) Average Effective Robustness

Figure 3: 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.

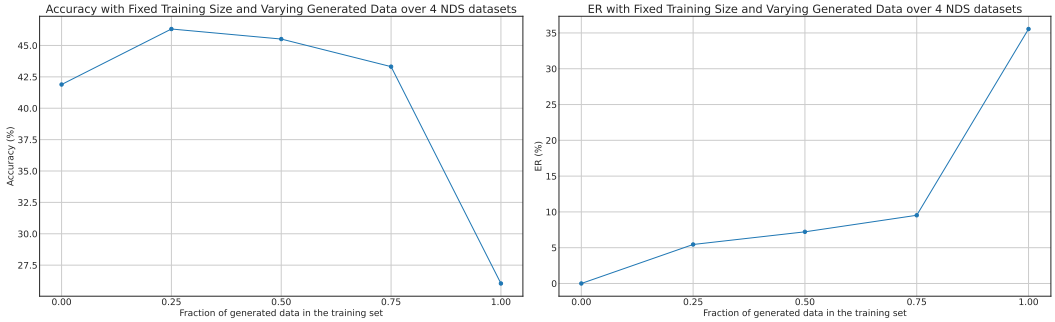
E.1 Effect of Real and Generated Dataset Size

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 so, 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 3a, 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 3b, 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. These findings hold across majority of dataset-specific results, as shown in Appendix §K.

We conducted an experiment to examine the impact of varying the amount of generated data with a fixed 1.3M training sample budget on ImageNet1K. Results are shown in Figures 4 and 4a. Figure 4 shows the accuracy and robustness of ResNeXt-50 over four natural distribution shifts, with an average of the results. In Figure 4a, accuracy increases initially with increasing generated data but drops by 15% when the fraction of generated data increases from 0.75 to 1. Conversely, the effective robustness increases monotonically with the increase in the proportion of generated data in the training mixture.



(a) Average Accuracy

(b) Average Effective Robustness

Figure 4: Variation in the accuracy and the effective robustness as we vary the proportion of the generated ImageNet1K data by fixing the number of training samples to 1.3M. While calculating effective robustness, standard training is performed on 1.3M real data. We report the results for the ResNeXt-50 classifier over three random seeds.

Table 6: 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 10 human annotators.

	Real	SD-Labels	SD-Images	SD-Labels&Images
Consistency	**0.96**	0.86	0.54	0.85
Quality	**4.52**	4.2	2.96	3.8

E.2 Effectiveness of Generation Strategies

E.2.1 Choice of Conditioning

Previously, we trained classifiers on data generated using 80 diverse templates with class label information from the ImageNet datasets. We chose this generation strategy as it does not require access to source images and thus generates diverse images that complement the images in the real dataset. As an alternative, we could have used just a single template, such as '*a photo of a* ***class label***', to create the complete dataset. We describe the other data generation strategies, SD-Images and SD-Labels&Images in §C.4.

To compare the diversity of images in the generated datasets, we calculated the diversity scores 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). The diversity scores are presented in Table 7. We find that the generated dataset created with a diverse set of templates ranked higher on the diversity score than the dataset created with a single template. Furthermore, we observed that data generation using only source image information (SD-Images) led to the most diverse creations within each class.

We compare the performance of classifiers trained on equal mixes of real ImageNet-100 data and synthetic data using various generation strategies results in Table 8. We find that the performance on training with synthetic dataset generated using diverse templates for class labels (SD-Labels), or the one generated using both class labels and source images (SD-Labels&Images), were closely tied at $\sim$66%. 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. The lowest performances were achieved by training on real data and the mix of real data and generated data using real data images (SD-Images), even though they ranked highly on the diversity metric (Table 7). This implies that the objects created by SD-Images may not be represented meaningfully or the quality of images is low.

Table 7: Comparison of diversity for various generation strategies.

Data	**Diversity**
Real	0.30
SD-Labels (Diverse Templates)	0.26
SD-Labels (a photo of a {class label})	0.15
SD-Labels (a rendition of a {class label})	0.16
SD-Images	**0.32**
SD-Labels and Images (Diverse Templates)	0.23

Table 8: 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 of ResNet-18, ResNeXt-50/101. We abbreviate ImageNet as Im.

	Im-100	Im-Sketch-100	Im-R-100	Im-V2-100	Obj-100	Average
Real Data	85.67	28.64	49.76	74.83	42.26	56.23
+ SD-Labels (Diverse Templates)	86.79	**48.40**	71.23	75.96	47.49	**65.97**
+ SD-Labels (a photo of a {class label})	87.45	35.66	59.51	75.63	44.87	60.62
+ SD-Labels (a rendition of a {class label})	87.37	46.27	67.75	75.95	46.48	64.83
+ SD-Images	85.89	32.20	50.00	74.90	45.13	59.51
+ SD-Labels and Images	**87.37**	46.73	**71.42**	**76.50**	**47.87**	**65.98**

E.2.2 Human Evaluations

Since the classifiers are trained using the generated data, 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, and whether the generated images are of poor quality i.e., they lack sharpness or contain perceptible noise. To that end, we collect 1600 annotations from 20 human surveyors for 40 images that are sampled from different datasets (one real and three generated) belonging to 10 ImageNet classes. Further details on the data collection process are presented in Appendix §F.

We find that images belonging to the real ImageNet dataset are the most consistent and of highest quality, which is expected given the extensive data curation and cleaning process that went into creating ImageNet. Additionally, we observe that the consistency and quality scores of images generated via the SD-Labels or SD-Labels&Images strategies are close, providing further evidence for the effectiveness of these strategies for training robust classifiers. However, we also find that synthetic data generated using the SD-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.

Table 9: Comparison of different classifiers on the ImageNet-G (Im-G) variants. The accuracy gap between Im-1K and the Im-G variant 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	Im1K	Im1K-G (SD-L)	Im1K-G (SD-I)	Im1K-G (SD-LI)
ResNeXt-101 (Trained on ImageNet1K)	79.28	55.93 (-23.3)	33.38 (-45.90)	60.9 (-18.38)
ViT-L/14-336 (PT-Im12K-FT-Im1K) Dosovitskiy et al. (2020)	**88.54**	66.24 (-22.30)	35.67 (-52.87)	66.78 (-21.76)
MaxViT-XL-512 (PT-Im21K-FT-Im1K) Tu et al. (2022)	88.26	68.61 (-19.65)	**39.29** (-48.97)	70.61(-17.65)
Zero-shot CLIP (ViT-B/32) Radford et al. (2021)	68.33	71.89 (+3.56)	34.45 (-33.88)	73.09 (+4.76)
Finetuned CLIP (ViT-B/32) Wortsman et al. (2022)	81.31	64.05 (-17.26)	34.67 (-46.64)	65.17 (-16.14)
ResNeXt-101 (Trained on ImageNet1K&SD-Labels)	80.39	**89.04** (+8.65)	34.40 (-45.99)	**83.91** (+3.52)

E.3 Evaluating Classifiers on Generated Datasets

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

Table 10: Comparison of different classifiers on the filtered images from the ImageNet-G variants. We abbreviate Filtered-ImageNet as F-Im, Stable Diffusion as SD, Labels as L, and Images as I. The performance gap between ImageNet-1K validation set and the respective generated evaluation set is reported inside the grey brackets.

Models	F-Im1K	F-Im1K-G (SD-L)	F-Im1K-G (SD-I)	F-Im1K-G (SD-LI)
ResNeXt-101 (Trained on ImageNet1K)	90.75	73.18 (-17.57)	57.08 (-33.67)	74.54 (-16.21)
ViT-L/14-336 (pt-12K-ft-1K)	94.41	82.29 (-12.12)	64.29 (-30.12)	81.44 (-12.97)
MaxViT-XL-512 (pt-21K-ft-1K)	**94.49**	79.93 (-14.93)	58.87 (-35.62)	76.87 (-17.62)
Zero-shot CLIP (ViT-B/32)	83.13	85.63 (+2.50)	**65.50** (-17.63)	85.01 (+1.88)
Finetuned CLIP (ViT-B/32)	90.73	78.35 (-12.38)	59.61 (-31.12)	76.86 (-13.87)
ResNeXt-101 (Trained on ImageNet1K&SD-Labels)	90.99	**96.95** (+5.96)	60.07 (-30.92)	**91.75** (+0.76)

(ImageNet-21K/12K) and finetuned on ImageNet-1K, (c) Zero-shot CLIP, (d) CLIP finetuned on the real ImageNet-1K dataset, on the three ImageNet-G variants, in Table 9.

We find that the ResNeXt-101 model trained solely on the real ImageNet-1K dataset experiences a large accuracy gap between the performance on ImageNet-1K validation dataset, and ImageNet-G variants. Despite performing the best on ImageNet-1K validation datasets, ViTs underperform on ImageNet-G variants. The performance of this classifier on ImageNet-G (SD-Labels) highlights the potential for further improvements in the existing models on ImageNet-G benchmark. Interestingly, we find that zero-shot CLIP does not undergo a distribution shift on two of the three ImageNet-G variants. 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. However, we find that finetuned CLIP to large accuracy gaps on ImageNet-G variants. An identical behavior of the finetuned CLIP was observed in Wortsman et al. (2022) for established NDS datasets, which suggests at the worthiness of ImageNet-G to be also considered as distribution shift test bed. We also train a classifier on the mix of real and generated data to assess the best achievable performance on ImageNet-G (SD-Labels). The performance of this classifier on ImageNet-G highlights the potential for further improvements in the existing models on Im-G.

In Table 9, we find that all the models suffer the most on the dataset created using the SD-Image strategy. Hence, to assess the performance of the models on a cleaner generated dataset using zero-shot CLIP, we filter out 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 used in Schuhmann et al. (2022). We present the results of different classifiers on the filtered versions of the real Im-1K validation set and Im-G variants in Table 10. We find that the accuracy gap, ranging from $12\% - 30\%$ between the performance on Im-1K and Im-G variants still persists for majority of the classifiers. As before, we observe large gaps between the performance of a classifier that is trained with the generated dataset (SD-Labels) and the other classifiers, trained with real data, on ImageNet-G.

F 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: SD-Label, SD-Image, and a combination of both, as described in §E.2. 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] Human annotators are graduate students from the department of CS at UCLA.

Figure 5: Survey screenshot

G 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.

H 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.

I TEMPLATES USED FOR DATA GENERATION

We present the list of 80 diverse templates that were used to generate the new images in Table 11.

J VISUALIZATION OF IMAGE GENERATIONS

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

K 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.

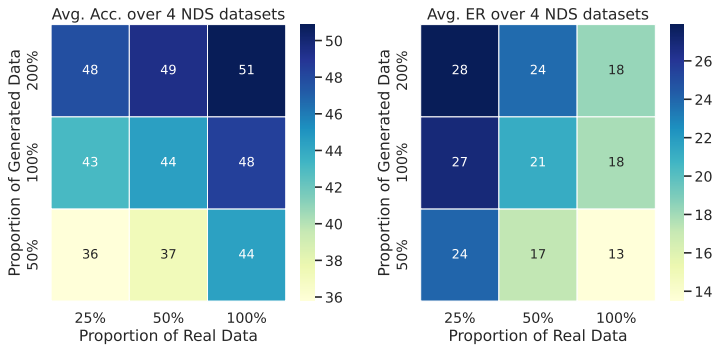
	'a bad photo of a {class label}.', 'a photo of many {class label}.', 'a sculpture of a {class label}.', 'a photo of the hard to see {class label}.', 'a low resolution photo of the {class label}.', 'a rendering of a {class label}.', 'graffiti of a {class label}.', 'a bad photo of the {class label}.', '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 11: List of diverse templates used for generating data.

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

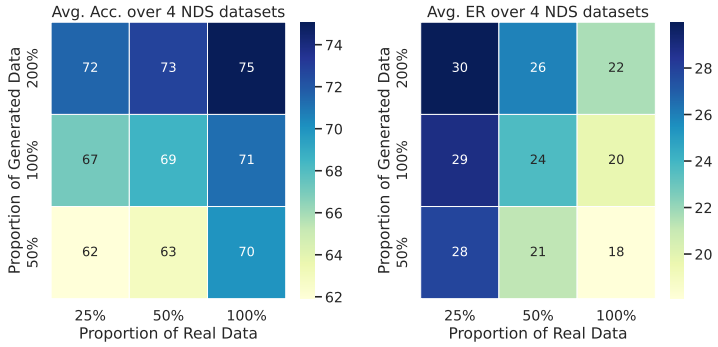
L Training Dynamics

We present the loss curve, in Figure 11, to compare the training dynamics of a classifier, ResNeXt-50, on the real ImageNet-1K data and an equal mix of real and generated ImageNet-1K data in 100:100 proportion.



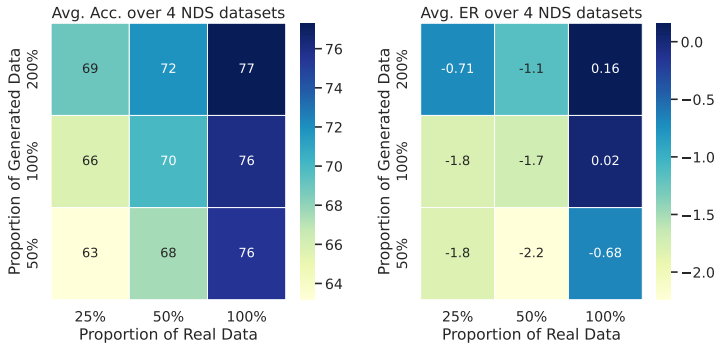
(a) Accuracy (b) Effective Robustness

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.



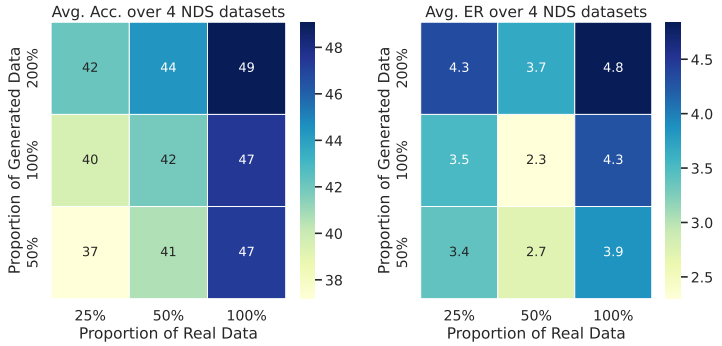
(a) Accuracy (b) Effective Robustness

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.



(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.

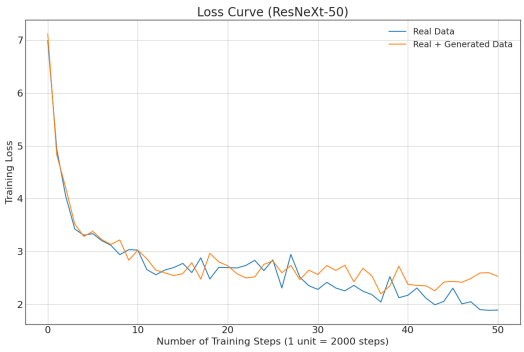


Figure 11: Comparison of the Loss Curve for ResNeXt-50 while training with the real and the generated data. The number of training samples in the real data is 1.3M whereas the number of training samples in the real and generated data scenario is 2.6M.