Robustness May be More Brittle than We Think under Different Degrees of Distribution Shifts

Kaican Li[1], **Yifan Zhang**[2], **Lanqing Hong**[3], **Zhenguo Li**[3], **Nevin L. Zhang**[1]
[1] Hong Kong University of Science and Technology
[2] National University of Singapore [3] Huawei Noah's Ark Lab
klibf@connect.ust.hk, yifan.zhang@u.nus.edu,
{honglanqing, li.zhenguo}@huawei.com, lzhang@cse.ust.hk

Abstract

Out-of-distribution (OOD) generalization is a complicated problem due to the idiosyncrasies of possible distribution shifts between training and test domains. Most benchmarks employ diverse datasets to address the issue; however, the degree of the distribution shift between the training domains and the test domains of each dataset remains largely fixed. Our study delves into a more nuanced evaluation setting that covers a broad range of shift degrees. We show that the robustness of neural networks can be quite brittle and inconsistent under different shift degrees, and therefore one should be more cautious in drawing conclusions from evaluations under a limited set of degrees. In addition, we find that CLIP, a representative of vision-language foundation models, can be sensitive to even minute distribution shifts of novel downstream tasks. This suggests that while pre-training may improve downstream in-distribution performance, it could have minimal or even adverse effects on generalization in certain OOD scenarios of the downstream task. A longer version of this paper can be found at https://arxiv.org/abs/2310.06622.

1 Introduction

Out-of-distribution (OOD) generalization is vital to the safety and reliability of machine learning applications in the real world. However, the complexities of distribution shifts between the training domains and the real test domains make OOD generalization difficult. Numerous empirical studies [11, 42, 41] have suggested that most algorithms only offer very little improvement in OOD performance over empirical risk minimization (ERM) [40]. Furthermore, algorithms performing better than ERM against one type of distribution shift often perform poorly against another [46]. The inconsistency suggests that it is important to consider various possible types of distribution shifts of a task when evaluating the OOD performance of a model; otherwise, the evaluation might be biased.

To address the issue, most OOD benchmarks [18, 14, 11, 46] incorporate multiple datasets exhibiting a diverse range of distribution shifts. However, another potential source of evaluation bias is often overlooked: the test domains of these datasets only capture a largely fixed degree of each distribution shift. For example, in [22, 18, 13, 49], each test domain represents a different "direction" of the possible distribution shifts of a task but there is no distinction between different degrees of shift on the same direction. Similar problems can also arise when only the aggregate performance across multiple degrees is examined [15]. To see the implications of such evaluations, consider the situation (that we observed in this work) illustrated in Fig. 1, where the performance of a model is evaluated in only two domains, one for in-distribution (ID) performance in the training domain $\mathcal{D}_{\text{train}}$, and the other for OOD performance in the test domain $\mathcal{D}_{\text{test}}$. In this case, the observed performance, which can be explained by at least two distinct generalization patterns as shown in the figure, presents an oversimplified summary of the OOD generalization ability of the model. This simplification may lead to misconceptions about model robustness under various degrees.

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

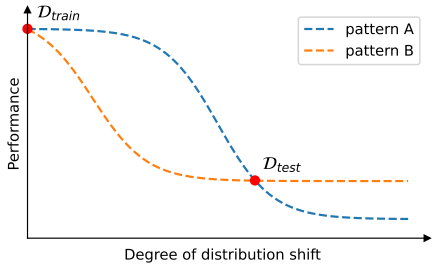

Figure 1: A typical situation where an evaluation under a limited set of shift degrees cannot tell any difference between two distinct OOD generalization patterns (labeled as A and B).

In this study, we take a closer look at OOD generalization under different degrees of distribution shifts and make several observations about the generalization behavior of neural networks under the setting. First, we highlight that the advantage of a model under some distribution shift may not apply to stronger shifts of the same type, even if the shifts are just slightly stronger[1]. Second, we find that training a model with strongly shifted data can sometimes guarantee robustness to all milder shifts, but at other times it only has a limited impact on robustness and may even harm the OOD performance under milder shifts. Lastly, the brittleness of robustness to different degrees of distribution shift is also observed in *foundation models* [4]. We find that while CLIP [30] models are able to adapt to many novel tasks, achieving great (sometimes near-perfect) downstream ID performance, they can be very sensitive to downstream distribution shifts—even an extremely mild shift that has no impact on models trained from scratch can cause a disproportionate performance drop in CLIP models.

2 Related work

OOD generalization under different degrees of distribution shifts. [15] proposed a benchmark based on images under different severity levels of common image corruption. However, the work did not provide any analysis at each individual severity level. [25] considered three severity levels of spurious correlation but did not discuss the connection between the model performance at each level. Similarly but in a different context, [36] conducted evaluations against three different degrees of spurious correlation and found that unsupervised methods are generally more robust than supervised learning and the advantage grows as the degree of the distribution shift increases. [33] showed that models regardless of supervision signal and architectural bias could not learn the underlying mechanism causing the distribution shifts on several datasets.

Robustness of CLIP. Foundation models such as CLIP [30] leverage a massive scale of training data to generalize to a great variety of downstream tasks. Zero-shot CLIP models are able to attain much higher OOD accuracy on several ImageNet variants than other models trained with a much smaller scale of data. Later, it is shown that the main source of the remarkable robustness of CLIP is the diversity of its training data distribution [9]. While CLIP can be made even more robust in some tasks after proper adaptation [43], what remains unclear in the literature is to what extent the robustness of CLIP and other foundation models can transfer to downstream tasks and how the models would behave as the degree of the downstream distribution shifts increases.

3 Robustness may not even extrapolate to slightly higher degrees

Data under strong distribution shifts are usually very rare in the real world. We often face situations where we only have access to a reasonable amount of data under relatively mild distribution shifts. In these situations, an important question is: how much can the performance of a model under some distribution shift tell us about its performance under stronger shifts? To approach the question, we constructed a dataset, NOISYMNIST, by adding Gaussian noise to MNIST [21]. As shown in Fig. 5, NOISYMNIST consists of a clean subset $\mathcal{D}_0$ of MNIST and 10 subsets $\{\mathcal{D}_i\}_{i=1}^{10}$ under different

[1]We use "mild/strong" and "low/high-degree" interchangeably when describing a distribution shift.

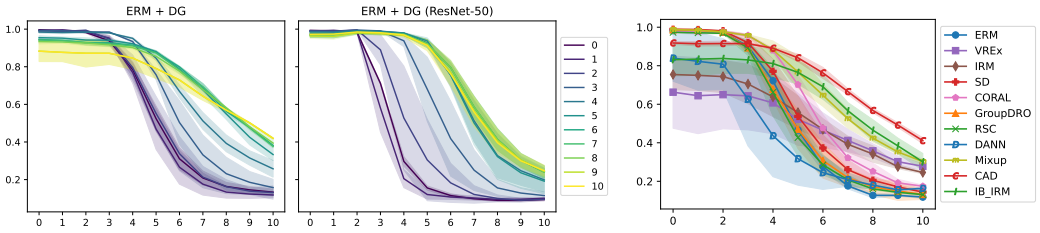

Figure 2: (Left) Average accuracy of the top-3 (among 400+) models at each degree of NOISYMNIST. The label of the curves denotes the domain on which the models perform best. (Right) Average accuracy of top-3 (among 20 for each algorithm) models of ERM and representative domain generalization algorithms on NOISYMNIST. The models are selected by worst-domain accuracy.

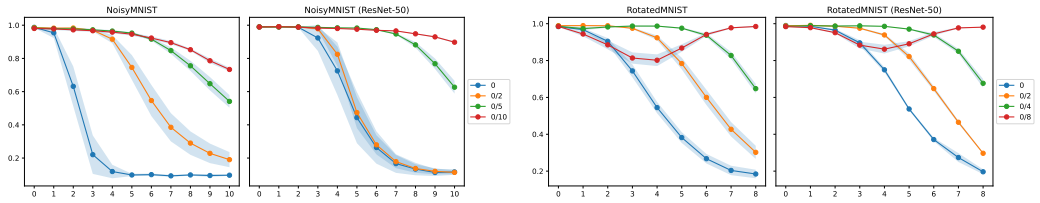

Figure 3: Average accuracy of ERM models trained on domains under different shift degrees. The label of the curves denotes the training-domain indices, e.g., "0/2" means that the models are trained on $\mathcal{D}_0$ and $\mathcal{D}_2$ of the dataset. The results are averaged over 20 models with different initialization.

degrees of noise. While the construction of NOISYMNIST is simple, it is nonetheless representative of a wide range of distribution shifts that gradually corrupt predictive features in an image.

We trained a pool of models on $\mathcal{D}_0$ and $\mathcal{D}_1$ of NOISYMNIST with ERM and more than 20 domain generalization (DG) algorithms (see the full list in Appendix A.2). The performance of the best-performing models in each domain is shown in Fig. 2 (left, ERM+DG). The result indicates that models that are better under milder shifts are often significantly *worse* than the other models under stronger shifts. In particular, the average accuracy of the best models in $\mathcal{D}_4$ has dropped by more than 10% in $\mathcal{D}_5$ which is only under a slightly more intense noise than $\mathcal{D}_4$. We further experimented with ResNet-50 [12] as shown in Fig. 2 (left, ERM+DG (ResNet-50)). While this shows that larger networks help, the gap may never be closed by merely increasing the capacity of the network. More importantly, *the robustness of a model may be more brittle than we think: even under the same type of shift, a slight increase in the degree of the shift may severely harm the performance of the model*.

The brittleness of robustness under different shift degrees also has implications in evaluating different learning algorithms. The performance of ERM and representative DG algorithms on NOISYMNIST are shown in Fig. 2 (right), where the algorithms exhibit very different generalization patterns that cannot be accurately captured under a limited set of shift degrees. In Appendix B.1, we further show that the performance drop of individual algorithms can be even more drastic than that is shown in Fig. 2, and then provide some analysis about the brittleness we observed in this section.

4 Robustness at higher degrees does not imply robustness at lower degrees

In this section, we shed some light on the reverse question of the previous section: do models being more robust to stronger distribution shifts imply them being more robust to milder distribution shifts? To start with, we obtained models that are robust to strong distribution shifts by training the models on strongly shifted data together with clean data. In Fig. 3, we compare these models with (i) models trained on much more mildly shifted data (also in addition to clean data) and (ii) models trained on clean data alone. For NOISYMNIST, the answer to our question is affirmative. However, on another dataset, ROTATEDMNIST (see Fig. 5 for examples), robustness against higher shift degrees does *not* guarantee robustness to lower degrees. In fact, it may even harm the performance at lower degrees. Larger networks and DG algorithms are helpful but only to a limited extent (see Fig. 8 for the DG results). Again, this demonstrates the brittleness of the robustness of neural networks.

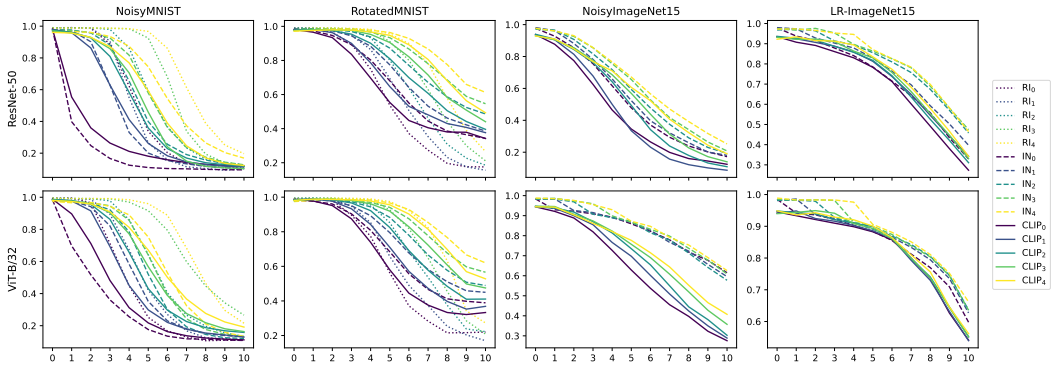

Figure 4: Performance of randomly initialized (**RI**) models, ImageNet (**IN**) pre-trained models, and **CLIP** models on different downstream tasks, evaluated over a broad range of shift degrees. The color of the curves indicates the domains used to train/adapt the models, e.g., RI_d stands for models trained on $\{\mathcal{D}_0, \ldots, \mathcal{D}_d\}$ from scratch. The pre-trained models are adapted to the downstream tasks through linear probing. The results are averaged over three runs (see Appendix B.3 for more details).

An important practical implication of our finding in this section is that, *even for the same type of distribution shift, obtaining data under high degrees of the shift may not be sufficient to learn a model that is robust to the lower degrees*. Combined with our finding in Sec. 3, we arrive at the conclusion that, for some tasks, the corresponding training data may be necessary to guarantee robustness at a certain degree of distribution shift. Meanwhile, we should also note that there are scenarios (e.g., NOISYMNIST) where training on a dataset under a sufficiently strong shift is able to guarantee robustness to all milder shifts. These kinds of distribution shifts therefore require much less data to induce general robustness across degrees.

5 Pre-trained representations are sensitive to novel downstream shifts

Pre-training on large-scale datasets is one of the most effective ways that are known to consistently improve the generalization of neural networks across a wide range of tasks [39, 26]. In particular, foundation models like CLIP [30] have demonstrated remarkable generalization capability. In Fig. 4, we compare CLIP models with ImageNet (IN [6]) pre-trained models and randomly initialized (RI) models trained from scratch on the downstream tasks to investigate the robustness of pre-trained representations. For NOISYMNIST, although the pre-trained models are able to perform equally well on clean domains as the RI models, they are surprisingly much more brittle to the shift induced by the noise. Notably, the gap of accuracy between CLIP_0 and RI_0 increased by more than 40% from $\mathcal{D}_0$ to $\mathcal{D}_1$ on ResNet-50. On ViT-B/32 [7], a similar pattern is observed albeit slightly improved. We hypothesize that the sensitiveness is largely because Gaussian noise is very rare in the training data of CLIP and also in ImageNet. Evaluation on ROTATEDMNIST, which exhibits a more common type of shift, provided some support for our hypothesis. The pre-trained models are only slightly worse than the RI models under mild to moderate shifts while being much better under strong shifts.

We also compare CLIP models with IN pre-trained models on harder problems: NOISYIMAGENET15 and LR-IMAGENET15 (see Appendix A.1 for more details). We observe that IN pre-trained models are generally more robust than CLIP models under both distribution shifts, often with *enlarging accuracy gaps* as the shift gets stronger. This suggests that not only the nature of the downstream shift (e.g., noise) but also the difference between the pre-training data and the downstream task itself plays a role in determining the robustness of the pre-trained representations against the shifts.

Lastly, we note that further adapting the pre-trained models to downstream shifts can sometimes significantly improve their robustness. On one hand, this corroborates existing findings that large-scale pre-trained representations are highly versatile. On the other hand, this also suggests that unleashing the power of pre-training may still require sufficiently diverse downstream task data that covers the potential distribution shifts. Nevertheless, there are still inherent limits to the power of pre-training under novel downstream distribution shifts as demonstrated in the case of NOISYMNIST where further adaptations are not nearly as effective as training from scratch (e.g., CLIP_4 vs. RI_4).

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, 2011.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

[6] 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*, pages 248–255. Ieee, 2009.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

[9] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[11] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[12] 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*, pages 770–778, 2016.

[13] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[14] 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*, pages 8340–8349, 2021.

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[16] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020.

[17] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, 2021.

[18] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[19] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv:2008.01883*, 2020.

[20] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv:2003.00688*, 2020.

[21] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[22] D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018.

[24] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[25] Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

[26] 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 *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[27] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[28] Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *ICLR*, 2021.

[29] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[31] Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

[32] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

[33] Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[35] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

[36] Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip Torr, and Amartya Sanyal. How robust is unsupervised representation learning to distribution shift? In *The Eleventh International Conference on Learning Representations*, 2022.

[37] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv:2104.09937*, 2021.

[38] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[39] 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.

[40] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[41] Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning. In *Advances in Neural Information Processing Systems*, 2022.

[42] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *ICLR*, 2022.

[43] 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*, pages 7959–7971, 2022.

[44] Yilun Xu and Tommi Jaakkola. Learning representations that support robust transfer of predictors. *arXiv preprint arXiv:2110.09940*, 2021.

[45] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*, 2020.

[46] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7947–7958, 2022.

[47] Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and improving transferability in domain generalization. *Advances in Neural Information Processing Systems*, 34:10957–10970, 2021.

[48] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[49] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European Conference on Computer Vision*, pages 163–180. Springer, 2022.

A Additional information about experiment setup

A.1 Datasets

Our study employs two altered versions of the MNIST dataset [21], herein referred to as NOISYMNIST and ROTATEDMNIST. To introduce varying degrees of distribution shifts, the NOISYMNIST dataset is generated by introducing Gaussian noise to the original images, resulting in 10 shifted domains under different degrees. More specifically, The standard deviation of the noise is linearly spaced between 0 and 0.8, in increments of 0.08, at the pixel level, normalized to the pixel value range of 0 to 1. Any pixel value beyond this range is clipped to fit within the 0-1 boundary. The ROTATEDMNIST dataset is created by rotating the original images, with degrees linearly spaced from 0 to 80, at intervals of 10 degrees, resulting in 8 shifted domains. Note that our ROTATEDMNIST is different from the ones in other papers, e.g., [11] which covers a smaller set of rotation degrees. We extend ROTATEDMNIST to span from 0 to 100 degrees in the experiments of Sec. 5.

In addition to NOISYMNIST and ROTATEDMNIST, we consider two more complicated datasets, NOISYIMAGENET15 and LR-IMAGENET15, which are modifications of a 15-category subset of ImageNet on bird species. NOISYIMAGENET15 follows a similar construction to NOISYMNIST, introducing Gaussian noise on the pixel level, linearly spaced between 0 and 0.8, with values clipped to the 0-1 range. Meanwhile, LR-IMAGENET15 involves altering image resolution, first downsampling via bilinear interpolation and subsequently upsampling to 256×256, with the downsampled resolution in each domain corresponding to a factor of $0.8^d \cdot 256$, where d represents the degree of distribution shift.

Random examples drawn from each domain of the datasets we used are shown in Fig. 5 and Fig. 6. The order of the examples is arranged according to the degree of the distribution shift from low to high.

Figure 5: Examples of NOISYMNIST and ROTATEDMNIST.

Figure 6: Examples of NOISYIMAGENET15 and LR-IMAGENET15.

For NOISYMNIST and ROTATEDMNIST, 60,000 images were divided into distinct training domains. For instance, in scenarios involving two training domains, each domain would encompass 30,000 images. Within the training domains, 20% of the data is allocated for in-distribution validation, aiding model calibration and selection. Every test domain of each altered dataset consists of 10,000 images, constructed using the same set of original images.

For NOISYIMAGENET15 and LR-IMAGENET15, we use the images in the training split of ImageNet to construct the training domains and the images in the validation split of ImageNet to construct the test domains. Similarly, the training domains divide the total 15,000 images in the training split. The test domains are constructed using the same set of original images, which consist of 750 images in total.

The 15 categories of birds we used in NOISYIMAGENET15 and LR-IMAGENET15, which correspond to indices 10 to 24 of the 1,000 categories of ImageNet, are “brambling, Fringilla montifringilla”,

"goldfinch, Carduelis carduelis", "house finch, linnet, Carpodacus mexicanus", "junco, snowbird", "indigo bunting, indigo finch, indigo bird, Passerina cyanea", "robin, American robin, Turdus migratorius", "bulbul", "jay", "magpie", "chickadee", "water ouzel, dipper", "kite", "bald eagle, American eagle, Haliaeetus leucocephalus", "vulture", and "great grey owl, great gray owl, Strix nebulosa".

A.2 Algorithms

Here is the full list of domain generalization algorithms we used in this study:

- Invariant Risk Minimization (**IRM**, [2])
- Group Distributionally Robust Optimization (**GroupDRO**, [32])
- Interdomain Mixup (**Mixup**, [45])
- Marginal Transfer Learning (**MTL**, [3])
- Maximum Mean Discrepancy (**MMD**, [23])
- Deep CORAL (**CORAL**, [38])
- Domain Adversarial Neural Network (**DANN**, [10])
- Conditional Domain Adversarial Neural Network (**CDANN**, [24])
- Style Agnostic Networks (**SagNet**, [27])
- Adaptive Risk Minimization (**ARM**, [48])
- Variance Risk Extrapolation (**VREx**, [20])
- Representation Self-Challenging (**RSC**, [16])
- Spectral Decoupling (**SD**, [29])
- Learning Explanations that are Hard to Vary (**AND-Mask**, [28])
- Smoothed-AND mask (**SAND-mask**, [35])
- Out-of-Distribution Generalization with Maximal Invariant Predictor (**IGA**, [19])
- Gradient Matching for Domain Generalization (**Fish**, [37])
- Self-supervised Contrastive Regularization (**SelfReg**, [17])
- Learning Representations that Support Robust Transfer of Predictors (**TRM**, [44])
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (**IB-ERM** & **IB-IRM**, [1])
- Optimal Representations for Covariate Shift (**CAD** & **CondCAD**, [31])
- Quantifying and Improving Transferability in Domain Generalization (**Transfer**, [47])
- Invariant Causal Mechanisms through Distribution Matching (**CausIRL** with CORAL or MMD, [5])
- Empirical Quantile Risk Minimization (**EQRM**, [8])

We use the DomainBed [11] implementation for all the above algorithms.

A.3 Implementation details

To conduct our experiments in Sec. 3 and Sec. 4, we employed two neural network architectures: a simple 4-layer Convolutional Neural Network (CNN) specialized for MNIST (see [11]) and a more complex ResNet-50 [12] model. Both models were implemented without any form of pretraining. For optimization purposes, we utilized the Adam optimizer with a static learning rate of 0.001. The total batch size was fixed at 64 and was evenly divided across each training domain. No weight decay was applied during the training process. Training iterations were set to a maximum of 5,000 for the 4-layer CNN and 10,000 for the ResNet-50 to ensure convergence. No form of data augmentation was used throughout the training process, preserving the inherent distribution and characteristics of the datasets.

To ensure the reliability of our results, we conducted a thorough random search for hyperparameters, repeated 20 times, for all algorithms. Except for learning rate, batch size, weight decay, and

dropout which are fixed, the search of other hyperparameters follows that of DomainBed [11]. For experiments utilizing the ResNet-50 architecture, the original MNIST digits were resized to a resolution of 224×224 pixels. Subsequent normalization was performed using the mean and standard deviation inherent to the MNIST dataset.

ImageNet pre-trained ResNet-50 and ViT-B/32 from torchvision, along with CLIP checkpoints of these models released by OpenAI, serve as our primary models in Sec. 5. These models are adapted to downstream tasks through linear probing aligned with [30]. For both pre-trained and randomly initialized models, we use training-domain validation to select the best models among different iterations. All MNIST-based datasets were resized to 224×224 for uniformity across models. Pre-trained models normalized all datasets based on the statistics of their respective pre-training datasets, while randomly initialized models normalized based on MNIST statistics. For the randomly initialized ResNet-50, no data augmentation was implemented during training on either NOISYMNIST or ROTATEDMNIST; we trained for 10,000 maximum iterations under a fixed learning rate 0.001. For the ViT-B/32 model, random affine transformations were applied, with rotation and shearing disabled for ROTATEDMNIST; we trained for 200,000 maximum iterations under a fixed learning rate 0.00003. For both the randomly initialized ResNet-50 and ViT-B/32, we used batch size 64 evenly divided for each training domain, and no weight decay or dropout was applied. We do not use any data augmentation for the experiments on NOISYIMAGENET15 and LR-IMAGENET15.

B Supplementary experiment results and analyses

B.1 Robustness may not even extrapolate to slightly higher degrees

Table 1: Performance of the best models in $\mathcal{D}_4$ of ERM and representative DG algorithms. The relative performance drops (%) with respect to the performance in $\mathcal{D}_4$ are shown in the parentheses. All results are averaged over the top 3 models among 20 models with different initialization and hyperparameters for training.

Algorithm	CNN			ResNet-50		
	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$
ERM	77.8±2.8 (0.0)	47.7±5.2 (38.7)	26.5±5.0 (66.0)	97.4±0.3 (0.0)	84.0±5.5 (13.8)	54.8±14.6 (43.8)
VREx	90.1±1.7 (0.0)	74.3±5.6 (17.6)	53.4±6.4 (40.8)	64.6±1.7 (0.0)	32.1±2.3 (50.3)	17.4±0.9 (73.1)
IRM	78.7±2.4 (0.0)	57.6±8.2 (26.8)	38.0±11.3 (51.7)	95.7±0.8 (0.0)	82.4±1.3 (13.9)	56.6±6.8 (40.9)
SD	81.7±1.2 (0.0)	57.7±2.3 (29.4)	35.7±2.1 (56.4)	97.8±0.4 (0.0)	92.4±2.1 (5.5)	76.5±6.0 (21.7)
GroupDRO	74.0±1.7 (0.0)	50.3±5.7 (32.1)	29.9±8.4 (59.6)	82.4±9.0 (0.0)	53.1±17.7 (35.5)	30.5±10.4 (63.0)
RSC	84.3±4.6 (0.0)	61.4±7.2 (27.2)	39.6±7.1 (53.0)	88.4±4.0 (0.0)	64.6±8.6 (26.9)	39.2±8.0 (55.6)
Mixup	93.2±0.4 (0.0)	84.1±2.3 (9.7)	69.2±1.9 (25.7)	85.4±3.7 (0.0)	49.2±16.2 (42.4)	26.6±13.2 (68.8)
CAD	94.1±1.0 (0.0)	78.7±3.1 (16.3)	58.6±4.0 (37.7)	78.8±19.6 (0.0)	50.6±22.0 (35.8)	30.5±13.1 (61.4)
IB-IRM	86.1±5.6 (0.0)	68.9±9.7 (19.9)	54.6±13.3 (36.5)	91.0±2.9 (0.0)	59.5±12.0 (34.6)	30.1±12.3 (66.9)

When looking at the best-performing models in $\mathcal{D}_4$, the same brittleness can be generally observed for all the algorithms in Tab. 1, where the performance drop can be even more drastic than that is shown in Fig. 2 (left). Astoundingly, the relative performance drop can go up to 50.3% from $\mathcal{D}_4$ to $\mathcal{D}_5$ and 73.1% from $\mathcal{D}_4$ to $\mathcal{D}_6$.

Analysis. To better understand the observed brittleness, we visualized the attention of ERM and CAD models on NOISYMNIST using GradCAM [34] (see Fig. 7). While both ERM and CAD models can make accurate predictions in the clean domain $\mathcal{D}_0$, they rely on radically different patterns to do so. ERM prefers the most predictive features regardless of whether they are robust or not. In the case of NOISYMNIST, these features turn out to be local features, which are easily corrupted by the noise, and thus no longer predictive when the noise becomes intense.

From this perspective, we can see that the brittleness manifests when the spurious correlation between the local features and the target labels reaches a breaking point. However, where this breaking point is and how rapidly the correlation breaks seem to be totally dependent on the nature of the distribution shift and the task itself. While NOISYMNIST demonstrates a simple case where the breaking point is at a moderate degree of distribution shift, there can be scenarios where the break happens at a much lower or higher degree of distribution shift and happens much more rapidly. As a consequence,

evaluations that only consider a narrow range of possible shift degrees would be highly unreliable in those scenarios.

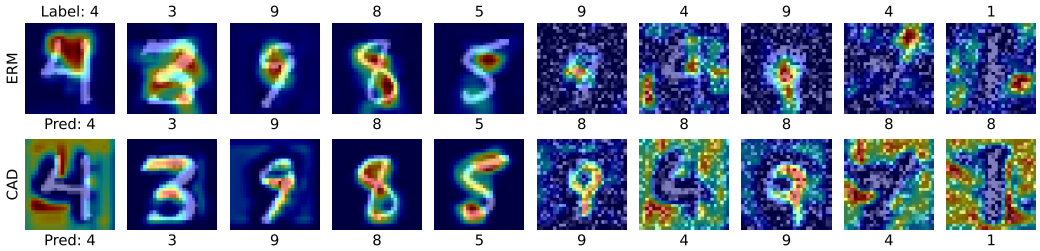

Figure 7: GradCAM visualization of model attention on random examples from $\mathcal{D}_0$ (left) and $\mathcal{D}_7$ (right) of NOISYMNIST. The two models (ERM and CAD) demonstrate distinctive generalization patterns, one relying on the local features while the other more on the global structures. The local features become unreliable as the noise becomes intense.

B.2 Robustness at higher degrees does not imply robustness at lower degrees

From Fig. 8, we can see that most DG algorithms are helpful to the generalization from higher shift degrees to lower shift degrees but only to a limited extent in the case of ROTATEDMNIST.

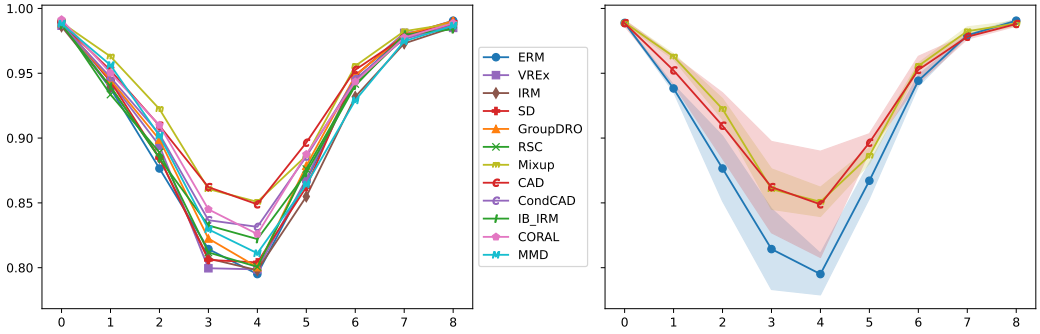

Figure 8: Average accuracy of the top-3 models of ERM and various DG algorithms trained on $\mathcal{D}_0$ and $\mathcal{D}_8$ of ROTATEDMNIST. The models are selected via training-domain validation. Error bars are omitted for clarity in the left sub-figure.

B.3 Pre-trained representations are sensitive to novel downstream shifts

Table 2: ResNet-50 results on NOISYMNIST.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.4±0.1	39.8±4.2	24.8±4.7	16.7±3.2	12.4±2.1	10.9±2.0	10.3±1.5	10.1±1.4	9.8±1.1	9.6±0.8	9.5±0.6
IN_1	97.9±0.1	97.3±0.2	90.2±0.7	62.9±2.0	33.1±1.7	20.1±1.9	15.4±1.7	13.3±1.1	11.7±1.1	10.8±0.9	10.5±1.0
IN_2	97.6±0.2	97.6±0.4	95.8±0.4	88.2±0.5	65.4±0.8	39.9±1.8	25.6±2.6	19.1±2.8	15.7±2.1	13.5±1.3	12.2±1.0
IN_3	97.6±0.2	97.2±0.1	95.9±0.1	92.6±0.8	81.8±0.4	57.9±1.3	36.2±1.6	23.5±1.7	17.4±1.9	14.3±1.6	12.7±1.1
IN_4	96.6±0.3	96.6±0.5	96.1±0.6	93.5±0.6	87.4±1.0	74.4±0.6	54.4±1.4	37.3±2.0	26.5±1.7	20.4±1.1	16.9±0.9
CLIP_0	98.1±0.1	55.4±4.1	35.9±4.4	26.4±3.3	21.1±2.7	18.1±1.4	15.8±0.4	14.1±0.4	13.0±0.4	12.3±0.5	12.1±0.2
CLIP_1	97.9±0.2	95.9±0.2	86.3±0.7	63.0±2.1	41.2±2.4	26.2±1.2	18.3±1.0	14.5±1.0	12.8±0.8	11.8±0.8	11.4±0.7
CLIP_2	97.4±0.2	96.2±0.2	92.7±0.5	81.0±0.6	58.3±1.7	35.9±1.7	22.9±0.6	16.9±0.4	14.0±0.3	12.8±0.6	11.8±0.4
CLIP_3	97.0±0.3	95.7±0.4	93.0±0.7	85.9±0.6	70.6±0.2	45.9±1.4	24.0±2.1	15.0±1.6	11.7±0.8	10.8±0.5	10.3±0.2
CLIP_4	96.1±0.4	95.7±0.5	92.9±0.1	87.1±0.7	76.9±1.0	59.3±0.5	38.9±0.6	24.6±1.0	17.0±1.0	13.6±0.6	12.2±0.4

Table 3: ViT-B/32 results on NOISYMNIST.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.4±0.3	69.7±1.4	51.7±1.7	36.1±1.6	25.6±1.5	17.6±0.8	13.5±0.4	11.8±0.4	11.6±0.7	11.1±0.5	10.8±0.6
IN_1	98.1±0.2	97.6±0.2	94.2±0.2	82.2±1.6	58.4±4.7	35.2±4.7	23.0±2.3	18.0±1.0	15.1±0.5	13.9±0.4	13.1±0.7
IN_2	98.1±0.1	97.7±0.4	96.3±0.3	92.2±0.2	78.5±0.6	52.0±1.1	30.0±1.5	19.3±1.7	14.5±1.0	12.5±0.4	12.1±0.5
IN_3	98.0±0.5	97.4±0.1	96.5±0.3	94.7±0.5	85.7±0.4	64.0±1.2	39.4±2.2	24.3±1.8	17.6±1.5	14.6±1.5	13.4±1.3
IN_4	97.4±0.6	96.6±0.5	96.7±0.2	94.3±0.4	89.1±0.6	75.2±0.6	53.0±2.1	32.4±3.0	21.1±2.8	15.8±2.3	13.3±1.6
CLIP_0	98.8±0.1	89.7±1.9	71.2±3.7	48.2±2.3	31.2±1.3	21.7±1.0	16.3±1.0	13.7±1.0	12.3±1.1	11.6±1.1	11.2±1.2
CLIP_1	98.5±0.2	97.8±0.2	91.5±1.0	69.3±4.2	45.0±5.6	29.9±5.5	22.1±4.1	17.8±3.1	15.4±2.0	14.2±1.9	12.9±1.3
CLIP_2	98.7±0.1	98.0±0.4	95.6±0.4	86.3±0.6	65.7±1.3	43.1±0.6	29.7±0.4	22.6±0.6	18.8±0.5	16.8±0.1	15.8±0.3
CLIP_3	98.3±0.2	97.5±0.1	96.0±0.1	90.4±0.6	77.2±0.4	56.7±0.6	38.8±0.2	27.6±0.5	21.8±0.4	18.1±0.5	16.3±0.5
CLIP_4	98.0±0.6	97.4±0.3	95.8±0.6	91.0±0.4	81.3±0.8	67.1±0.7	50.1±0.9	36.6±0.3	27.6±0.5	22.3±0.4	19.1±0.3

Table 4: ResNet-50 results on ROTATEDMNIST.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.5±0.1	97.3±0.1	94.7±0.1	89.2±0.2	80.3±0.2	68.1±0.0	55.0±0.3	44.5±0.4	38.6±0.7	36.5±0.8	34.3±1.2
IN_1	98.4±0.1	98.6±0.2	97.4±0.1	94.5±0.1	88.3±0.2	77.4±0.2	64.2±0.6	53.2±0.9	46.2±1.1	42.2±1.4	38.6±1.2
IN_2	98.2±0.2	98.5±0.3	98.3±0.1	97.0±0.1	93.8±0.3	87.2±0.5	77.1±1.1	67.1±0.9	58.6±1.4	52.4±2.0	48.6±2.1
IN_3	98.1±0.2	98.6±0.3	98.5±0.1	98.6±0.2	96.7±0.1	93.4±0.0	86.7±0.2	78.1±0.8	68.5±1.4	58.9±1.6	54.6±0.7
IN_4	97.2±0.2	98.0±0.3	98.4±0.2	98.5±0.6	97.8±0.4	96.4±0.1	92.9±0.2	86.8±0.1	77.6±0.9	65.9±1.4	61.3±1.5
CLIP_0	98.2±0.1	96.6±0.1	93.4±0.2	83.6±1.0	69.6±1.4	55.4±1.4	44.8±0.7	40.5±0.9	38.0±1.7	37.9±0.6	34.2±1.9
CLIP_1	98.0±0.1	98.2±0.1	96.3±0.2	90.0±0.3	78.6±0.6	64.7±0.9	53.1±0.6	48.1±0.8	43.0±0.8	40.7±0.7	37.6±0.6
CLIP_2	97.7±0.1	98.0±0.1	97.5±0.1	95.4±0.2	90.1±0.2	81.0±0.7	69.9±0.9	60.7±1.1	50.5±0.7	44.2±0.7	39.6±1.5
CLIP_3	97.6±0.2	98.1±0.1	97.9±0.5	97.6±0.4	95.3±0.1	90.5±0.3	82.1±0.5	71.5±0.6	58.3±0.8	50.8±0.6	44.1±1.4
CLIP_4	97.2±0.4	97.6±0.2	98.0±0.5	97.8±0.3	96.7±0.4	94.7±0.1	89.2±0.2	80.6±0.5	68.0±0.9	56.4±0.9	49.3±1.0

Table 5: ViT-B/32 results on ROTATEDMNIST.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.3±0.2	97.7±0.0	96.1±0.1	90.3±0.3	80.8±0.5	68.4±0.7	55.7±0.8	46.7±0.1	41.2±0.3	39.8±0.6	38.9±0.4
IN_1	98.3±0.0	98.6±0.1	97.8±0.1	95.0±0.2	89.2±0.4	79.5±0.5	67.7±0.6	58.2±0.9	51.1±0.3	45.9±0.9	45.0±0.9
IN_2	98.2±0.1	98.7±0.3	98.7±0.2	97.4±0.1	94.4±0.2	88.3±0.5	78.0±1.6	66.6±1.9	57.7±1.6	50.8±1.5	48.9±0.7
IN_3	97.9±0.2	98.4±0.3	98.7±0.4	98.6±0.1	96.9±0.1	94.1±0.3	87.5±0.3	77.7±0.2	67.7±0.6	59.7±1.0	56.6±0.8
IN_4	97.4±0.7	98.2±0.5	99.0±0.2	98.5±0.3	98.4±0.4	96.7±0.1	92.8±0.1	85.4±0.2	76.7±0.3	66.6±0.1	62.0±0.3
CLIP_0	98.8±0.1	97.8±0.1	95.1±0.1	87.5±0.4	73.9±0.7	58.0±0.6	44.5±0.5	37.6±0.4	33.3±0.1	32.1±0.5	33.3±0.2
CLIP_1	98.6±0.2	98.9±0.1	97.7±0.1	93.3±0.3	84.2±0.6	71.0±0.7	57.1±1.0	47.3±0.9	39.7±0.5	35.3±0.4	36.9±0.5
CLIP_2	98.6±0.2	98.9±0.2	98.4±0.2	96.7±0.0	92.2±0.2	82.5±0.7	69.1±1.5	57.8±1.8	48.4±1.3	41.0±1.9	41.1±1.0
CLIP_3	98.0±0.2	98.6±0.1	98.9±0.1	98.0±0.2	96.2±0.1	91.9±0.2	82.9±0.4	72.3±0.3	60.8±0.4	49.8±0.6	47.5±1.0
CLIP_4	98.1±0.5	98.9±0.1	98.7±0.2	98.4±0.2	97.8±0.4	95.6±0.1	90.2±0.2	81.5±0.5	69.7±1.2	56.9±1.1	52.8±1.7

Table 6: ResNet-50 results on NOISYIMAGENET15.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	97.7±0.1	92.7±0.2	85.5±0.5	75.5±1.1	61.7±1.3	47.8±1.0	37.1±1.0	29.4±1.5	23.9±0.5	20.2±0.8	17.2±1.1
IN_1	98.1±0.5	96.2±0.6	88.2±0.8	78.2±0.8	65.9±0.9	52.4±0.5	39.7±1.1	32.1±1.2	25.9±0.6	20.1±0.6	18.1±1.0
IN_2	97.1±0.6	96.6±0.3	90.7±0.4	81.7±0.8	71.7±0.3	59.4±1.6	47.8±1.3	37.5±1.2	29.0±0.9	23.0±1.9	19.3±0.7
IN_3	97.4±0.5	96.0±0.7	92.6±1.4	85.5±0.7	75.5±1.5	65.6±1.5	52.8±0.6	41.7±0.4	34.1±0.8	26.4±1.6	20.8±1.0
IN_4	97.7±0.2	96.0±0.3	93.1±0.8	85.6±2.4	76.6±1.8	67.1±0.4	56.6±0.9	47.1±2.0	39.2±0.7	31.7±0.6	25.2±0.8
CLIP_0	93.8±0.3	87.7±0.3	77.4±0.9	62.5±1.5	46.5±1.4	34.5±0.4	26.5±1.6	19.9±0.9	16.0±1.4	14.5±1.1	12.4±1.0
CLIP_1	93.4±0.9	90.5±0.7	81.4±1.0	67.8±0.2	50.0±1.2	33.5±1.9	23.2±0.7	15.7±0.4	12.2±0.4	10.2±0.6	8.7±0.5
CLIP_2	93.9±0.7	90.9±0.6	84.4±0.5	76.2±1.1	64.2±1.1	49.3±1.9	34.2±1.6	23.9±1.2	18.0±1.0	13.2±0.8	10.9±1.0
CLIP_3	93.0±0.8	90.5±1.2	85.5±1.8	77.6±1.3	68.6±1.0	56.9±1.1	43.8±0.5	31.3±1.7	22.7±1.2	17.2±0.9	14.0±0.4
CLIP_4	92.9±1.3	91.0±1.6	84.8±1.5	76.8±1.8	71.0±1.4	60.6±1.8	52.3±1.3	40.7±1.1	32.3±1.7	24.4±0.7	19.0±0.6

Table 7: ViT-B/32 results on NOISYIMAGENET15.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.3±0.3	92.9±0.2	92.2±0.7	90.9±0.3	89.3±0.3	85.9±0.8	82.2±0.4	77.9±0.5	71.8±0.4	67.7±1.5	61.5±0.9
IN_1	98.4±0.3	98.4±0.3	92.6±0.4	91.5±0.3	89.0±0.7	86.6±0.6	82.6±0.8	76.7±1.4	72.3±1.3	65.7±1.0	59.6±1.2
IN_2	98.2±0.4	98.3±0.3	96.9±0.3	90.7±0.2	89.0±0.3	87.1±0.6	81.4±0.9	77.8±1.2	71.3±0.2	64.2±0.9	57.7±2.0
IN_3	98.5±0.5	98.7±0.6	97.4±0.8	95.9±0.8	89.9±0.3	86.7±0.9	83.2±0.3	79.8±0.9	73.7±0.8	67.8±0.8	60.9±0.5
IN_4	98.8±0.3	98.5±0.8	97.6±0.2	95.5±1.2	92.9±1.4	87.2±0.5	84.7±0.1	79.2±1.3	75.3±0.7	69.1±0.8	62.6±0.6
CLIP_0	94.4±0.1	92.1±0.3	88.8±0.6	81.9±0.4	72.7±1.1	62.9±1.0	53.7±1.0	45.5±1.1	39.7±1.7	32.3±1.8	27.6±0.7
CLIP_1	94.6±0.5	93.7±0.4	90.0±0.4	85.5±0.6	76.6±1.7	69.7±2.6	60.2±2.2	50.9±4.0	42.4±3.2	35.0±1.8	28.9±2.9
CLIP_2	94.5±0.8	93.8±0.6	91.8±0.6	87.0±0.4	81.4±0.9	73.5±0.7	65.4±2.3	55.0±1.3	44.5±1.7	38.1±1.4	30.3±1.5
CLIP_3	94.8±1.1	94.6±0.7	90.7±0.6	87.4±1.2	82.3±0.9	75.7±1.0	69.1±0.6	60.8±0.6	51.9±1.9	42.8±1.3	35.7±0.8
CLIP_4	94.4±1.3	94.2±1.0	90.4±0.5	86.2±1.5	82.1±1.4	77.7±1.1	71.5±0.2	64.4±1.5	55.2±1.1	46.6±1.3	40.7±1.8

Table 8: ResNet-50 results on LR-IMAGENET15.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	97.9±0.2	93.8±0.5	91.6±0.3	87.9±0.5	84.1±0.5	78.5±1.3	71.5±2.0	64.2±2.8	55.8±3.6	43.4±4.6	32.9±2.9
IN_1	98.0±0.3	97.4±0.4	92.8±0.3	90.3±0.5	87.9±0.3	82.9±0.3	76.4±1.2	69.6±1.5	59.3±2.0	50.3±1.8	39.7±1.6
IN_2	96.8±0.4	97.3±0.2	96.4±0.4	91.4±0.3	89.4±0.1	85.6±0.5	81.0±0.8	75.7±0.5	67.4±1.4	57.0±1.1	45.8±0.8
IN_3	97.4±0.6	96.7±0.7	97.6±0.7	95.3±1.0	89.7±0.3	86.4±0.6	82.5±0.2	78.1±0.4	69.0±0.9	58.1±0.9	47.4±0.3
IN_4	97.3±0.4	97.4±1.1	96.6±0.5	95.3±0.3	94.6±0.9	87.1±0.4	82.6±0.3	78.4±0.9	69.9±0.9	58.2±1.2	46.5±2.1
CLIP_0	93.5±0.3	90.8±0.7	89.2±0.5	86.1±0.8	83.0±0.2	78.3±0.8	71.3±1.5	60.2±1.5	49.2±1.4	38.0±1.9	27.3±2.5
CLIP_1	93.3±0.3	92.4±0.2	90.7±0.4	88.8±0.5	86.0±0.7	81.3±0.4	73.8±0.5	64.9±0.9	54.5±1.2	44.3±2.2	34.3±1.3
CLIP_2	93.7±1.4	92.7±0.8	91.8±1.1	88.7±0.1	85.7±0.9	81.7±0.7	74.1±0.4	63.6±1.6	52.5±1.2	42.1±0.8	31.1±2.1
CLIP_3	93.2±0.3	92.2±0.9	92.3±0.3	91.0±2.0	86.4±0.7	83.0±0.4	75.8±0.3	66.3±1.2	55.4±1.9	44.7±1.0	33.1±1.8
CLIP_4	92.2±0.5	93.4±0.9	91.5±0.2	91.3±1.4	89.4±1.0	83.5±0.2	77.3±0.9	68.4±0.3	58.0±0.6	46.7±0.5	34.2±1.1

Table 9: ViT-B/32 results on LR-IMAGENET15.

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$	$\mathcal{D}_{10}$
IN_0	98.5±0.2	94.2±0.1	93.6±0.2	92.2±0.1	90.3±0.2	88.4±0.3	85.6±0.1	81.2±0.8	77.0±1.1	70.9±1.3	59.8±1.4
IN_1	98.1±0.4	98.6±0.3	93.7±0.1	92.4±0.1	91.1±0.3	89.2±0.2	86.5±0.2	83.4±0.3	79.5±0.2	73.6±0.3	62.7±0.5
IN_2	98.2±0.8	98.3±0.4	97.9±0.3	92.8±0.2	91.0±0.2	89.2±0.2	86.9±0.2	84.3±0.2	80.5±0.0	74.6±0.3	63.6±0.7
IN_3	98.5±1.3	98.3±0.8	98.4±0.0	98.2±0.2	91.3±0.4	90.0±0.5	87.4±0.3	83.9±0.0	80.4±0.4	74.4±0.3	63.6±0.5
IN_4	98.9±0.3	98.3±0.5	98.3±0.7	98.3±0.1	97.6±0.6	90.6±0.1	88.1±0.3	85.1±0.4	81.1±0.5	75.2±0.3	66.2±0.7
CLIP_0	94.5±0.3	93.3±0.1	92.1±0.4	91.0±0.1	89.9±0.4	88.3±0.3	85.7±0.3	80.0±0.7	73.6±1.0	62.7±1.2	54.1±1.9
CLIP_1	94.1±0.7	94.8±0.5	92.4±0.1	91.6±0.1	90.3±0.1	88.6±0.8	86.2±1.8	79.3±1.6	73.0±2.1	63.2±1.4	54.0±0.8
CLIP_2	94.8±1.1	94.2±1.0	93.9±1.0	92.1±0.1	91.0±0.2	89.9±0.3	86.6±0.3	79.8±0.3	74.0±0.3	64.6±0.5	55.2±0.5
CLIP_3	94.0±0.7	94.4±0.9	94.9±0.3	94.1±0.8	90.8±0.1	89.3±0.4	86.0±1.2	80.0±0.9	73.7±1.4	64.4±1.1	55.1±2.6
CLIP_4	94.9±0.4	93.6±1.0	94.3±1.2	92.9±0.2	92.1±0.6	90.0±0.2	86.4±0.3	81.0±0.7	75.6±0.6	64.6±0.3	56.2±0.3