Evaluating the Impact of Geometric and Statistical Skews on Out-Of-Distribution Generalization

Aengus Lynch, Jean Kaddour, Gbètondji J-S Dovonon, Ricardo Silva

University College London
London, UK
{aengus.lynch.17, jean.kaddour.20, gbetondji.dovonon.22}@ucl.ac.uk

Abstract

Out-of-distribution (OOD) or domain generalization is the problem of generalizing to unseen distributions. Recent work suggests that the marginal difficulty of generalizing to OOD over in-distribution data (OOD-ID generalization gap) is due to spurious correlations, which arise due to statistical and geometric skews, and can be addressed by careful data augmentation and class balancing. We observe that after constructing a dataset where we remove all conceivable sources of spurious correlation between interpretable factors, classifiers still fail to close the OOD-ID generalization gap.

1 Introduction

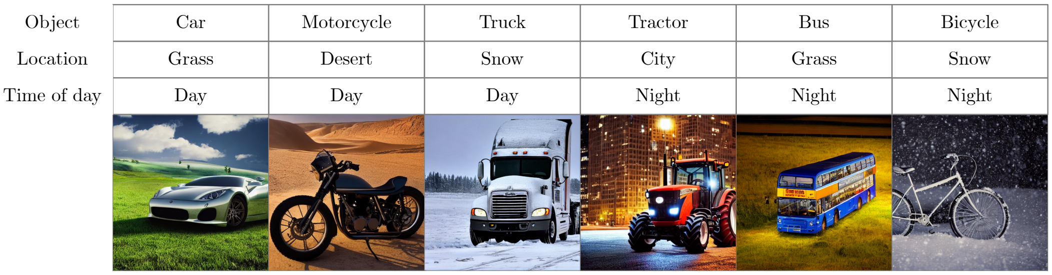
Out-of-distribution (OOD) or *domain generalization* is the problem of improving the ability of machine learning prediction models to generalize to unseen test domains. A key challenge in OOD generalization is to remove spurious correlations in the training data; patterns that discriminate between classes within the training domain, and not necessarily in the test domain [5]. To this end, a myriad of approaches that learn environment-invariant models have been proposed, see e.g., [6, Chapter 3].

However, recent work has questioned the efficacy of current OOD methods and reported sobering experimental results under rigorous examination [3, 10]. Nagarajan et al. [8] analyzed OOD failure modes, and found that spurious correlations induce two kinds of skews in the data: *geometric* and *statistical* skew. Geometric skew occurs when there is an imbalance between groups of types of data points (such as data points from different environments) which induces a spurious correlation, and leads to misclassification when the balance of groups changes. This understanding has motivated simply removing data points from the training data to balance between groups of data points [2]. Statistical skew occurs whenever there is a feature that is spuriously predictive of the label in the training data, yet there already exist invariant features in the training data which are fully sufficient to perfectly predict the label.

Based on this understanding, we would expect that a model trained on training environments without geometric and statistical skews would yield in-distribution-like generalization capabilities when tested in OOD domains. In other words, if we removed spurious correlations by accounting for geometric and statistical skews, our model's performance should not differ between in-distribution (ID) and OOD test data.

Surprisingly, we observe empirically that this does not hold, and we observe gaps between ID and OOD generalization, even after balancing object categories and environmental changes. We conduct experiments with synthetic yet photorealistic images, generated by text-to-image diffusion models [9], with human-interpretable factors in the data-generating process. We observe that *removing spurious correlations from the data may not be enough for OOD generalization*, we conjecture that the model

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).



Object	Car	Motorcycle	Truck	Tractor	Bus	Bicycle
Location	Grass	Desert	Snow	City	Grass	Snow
Time of day	Day	Day	Day	Night	Night	Night

Figure 1: **Examples from TILO** : We generate data according to three interpretable generating factors, ***t**ime of day*, ***l**ocation* and ***o**bject* (class label).

fails to extrapolate to unseen feature values, even if these features are themselves not predictive of the label [12].

2 Problem Setup

We formalize the problem of out-of-distribution generalization from a causal perspective [6, Chapter 3]. Imagine a dataset $\mathcal{D} := \{(\mathbf{X}^e, Y^e)\}_{e \in \mathcal{E}_{\text{train}}}$ with images $\mathbf{X}^{\boldsymbol{e}}$ and labels $Y^{\boldsymbol{e}}$ collected under training environments $\mathcal{E}_{\text{train}}$. We assume that each environment-specific distribution $p(\boldsymbol{x}^{\boldsymbol{e}}, y^{\boldsymbol{e}})$ corresponds to an interventional distribution from a structural causal model (SCM), i.e., $p(\boldsymbol{x}^{\boldsymbol{e}}, y^{\boldsymbol{e}}) := p_{\widetilde{\mathcal{M}}}(\boldsymbol{x}, y \mid \mathrm{do}(\boldsymbol{e}))$, where $\mathbf{E}$ describes the environment-specific components, $\mathcal{M} := (\mathbf{S}, p(\boldsymbol{\epsilon}))$ is an SCM with structural assignments $\mathbf{S} := \{f_i\}_{i=1}^N$ and $\widetilde{\mathcal{M}} := \mathcal{M}_{\mathrm{do}(\boldsymbol{e})} = \left(\widetilde{\mathbf{S}}, p(\boldsymbol{\epsilon})\right)$ is its modified variant with modified structural equations $\widetilde{\mathbf{S}}$ after interventions took place [6, Chapter 2].

Now, imagine we want to learn a classifier $f_{\boldsymbol{\theta}}(\mathbf{X}) \approx Y$ with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ using the training environments such that it generalizes well to novel interventional distributions (test environments) $\mathcal{E}_{\text{test}}$ corresponding to unseen $\mathbf{E}$. In general, we refer to this as the *out-of-domain (OOD) generalization* problem. Methods addressing this problem typically aim to identify environment-invariant features that are causal parents of Y and predictive under all interventional distributions [6, Chapter 3].

3 Data-Generating Process over Time, Location, and Object

The goal of our investigation is to measure a classifier's performance on OOD test data after accounting for common skews in the training data. Further, we impose three desiderata for the distribution shifts:

1. **Realistic:** We wish to study distribution shifts that are directly relevant to real-world impact.
2. **Unambiguous:** For debugging purposes, we wish to understand them clearly. For example, some existing OOD test datasets assume distribution shifts across different hospitals or countries [7]. Yet, it is difficult to qualitatively examine these shifts' characteristics and their severity.
3. **Controlled:** We wish to be able to control the environment's influence on the training dataset.

To this end, we design the following data-generating process (DGP) over images generated with text-to-image models [9]. We present the DGP in Figure 2 and image examples in Figure 1. We consider three generative factors (*object* = Y, *location* $= L$ and *time of day* $= T$) causing the data $\mathbf{X}$. We will refer to this DGP as **TI**me **L**ocation **O**bject (TILO).

Further, we study two types of confounding bias [6, Chapter 2]:

1. **Full confounding:** In Figure 2(a), the environment node E is a confounder of all three factors, thus $\mathbf{X} \sim p(\boldsymbol{x} \mid y, l, t, e)p(y, l, t \mid e)$. We construct a statistical skew between *location* and *object* by inserting object imbalances: removing *tractor* examples from *city* locations, and removing *bicycles*, *cars* and *buses* from *snow* and *grass* locations. We induce a group imbalance by imbalancing the training data between *time of day* groups not present in the test data (see Figure 3 for an illustration of the object imbalancing).

Train env.	ID. accuracy		Test env.	OOD. accuracy	
	ERM	gDRO		ERM	gDRO
C_N, G_D, D_D	**99%**	92%	S_D, S_N	**93.37%**	72.65%
C_N, G_D, S_N	**98%**	83%	D_D, D_N	**76.77%**	56.99%
C_N, S_D, G_N	**99%**	84%	D_D, D_N	**82.87%**	61.52%
G_N, D_N, S_D	**99%**	87%	C_D, C_N	**73.25%**	70.98%
Avg.	**99%**	87%	Avg.	**81.57%**	65.53%

(a) **Object and time imbalanced**: spurious correlations exist between location and objects, and there is a geometric skew between time of day groups

Train env.	ID. accuracy		Test env.	OOD. accuracy	
	ERM	gDRO		ERM	gDRO
C_D, G_D, D_D	**99%**	95%	S_D	**96.31%**	90.62%
C_D, G_D, S_D	**100%**	98%	D_D	**94.47%**	80.67%
C_N, S_N, G_N	**99%**	97%	D_N	**96.42%**	89.92%
G_D, D_D, S_D	**99%**	98%	C_D	**93.01%**	88.13%
Avg.	**99%**	97%	Avg.	**95.05%**	87.34%

(b) **Object and time balanced**: any statistical and geometric skew among the interpretable factors has been removed

Table 1: **Empirical results for generalization performance of ERM and gDRO on ResNet-18**: We denote our data distributions in the format $\text{Location}_{\text{time of day}}$: city ($C$), grass ($G$), desert ($D$), city ($C$); day ($D$), night ($N$). In all three tables, we train on three locations and test on a held-out location. The validation distribution is the same as the training distribution.

2. **No confounding**: In Figure 2(), all factors have been intervened upon, removing any environmental influence from the generation process, and $\mathbf{X} \sim p(\boldsymbol{x} \mid \mathrm{do}(y), \mathrm{do}(l), \mathrm{do}(t))$. Hence, there exist no geomebtric and statistical skews among the interpretable factors in the DGP.

These two variants (a-c) reflect common assumptions about potential sources of real-world distribution shifts throughout the OOD generalization literature: in environment b) of Figure 2, there exists a causal generating factor whose distribution is invariant to the environment, and fully predictive of the label. Throughout all experiments, we kept the number of objects the same across environments, such that the number of samples in any object, location, and time combination is the same across the whole dataset. This was done to control the sources of confounding in our dataset as much as possible.

Our data-generating process relies on the text-to-image diffusion model *stable diffusion* [9]. Thereby, we can produce a dataset satisfying all three desiderata by intervening on the prompts. The prompts followed the following template:

"a/an [object], in a [location], at a [time of day], highly detailed, with cinematic lighting, 4k resolution, beautiful composition, hyperrealistic"

4 Results

We conduct tests in which we aim to evaluate the impact of statistical and geometric skews on the performance of a vision classifier. To achieve this, we use a ResNet-18 [4] as our base model architecture and construct four generalization tasks for each skew regime given in Figure 2. Further, we compare the performance of ERM against gDRO: while ERM shuffles all the data in its training domain, gDRO assigns an environment label e to each environment and optimizes the worst-case loss. Results are shown in Table 1, and experiments were performed using NVIDIA 3090 GPUs.

We observe that removing the sources of statistical and geometric skews positively impacts generalization performance, as we would expect. In the case where both geometric and statistical skews have been removed (Table 1 b)), we still observe a generalization gap, where the OOD average accuracy (95.05%) is below the ID average accuracy (99%). We suspect that such a gap remains because of the failure of the model to extrapolate to unseen values of *location* features, similar to how a regression model can succeed at interpolation within the support of the data points, and fail at extrapolating away from the support.

5 Future Work

More object categories. We would like to conduct experiments with a greater number of object categories and understand how the diversity of object categories affects generalization performance.

Evaluating the problem of unseen feature values Further, we would like to evaluate the impact of including a sample of the test domain in the training data in improving generalization performance. In such a setting, the problem morphs from an OOD generalization problem to a multi-environment generalization problem, since the test distribution will have been observed in the training environments. Yet, this will allow us to reason about the impact of including full support of feature values in the training data on the test domain performance.

Dataset cleaning. Finally, we would like to improve our method of generating from TILO by removing failure modes of our generative model, including poor quality images and confusion in the generative factors. We illustrate such examples in the appendix (**??**).

6 Related Work

Arjovsky et al. [1] propose the colored MNIST benchmark, which altered the colors of MNIST digits and introduced a spurious correlation between colors and classification labels in the training data that was switched in the test data. Nagarajan et al. [8] introduced modifications on CIFAR10 images to isolate the effect of geometric and statistical skews on the performance of classifiers. The Causal3DIdent dataset [11] is similar to our dataset in its construction, due to their effort to model the data generating process with a causal DAG, yet we suggest that our dataset comes closer to real-world examples of distribution shifts while maintaining a controlled DGP.

Koh et al. [7] propose a selection of benchmarks for approaches to OOD generalization by compiling real-world data examples. However, we note that the data generating process is unknown in these examples, thus they do not provide insights into isolated types of distribution shift.

Recently, there has been a growth in interest in *Counterfactual Data Augmentation* [6, Chapter 4], where some data augmentations can be viewed as generating counterfactuals from a causal data generating process. Indeed, in this work, we can view moving from object imbalanced data to object balanced data as performing just that: sampling counterfactual data points, eg. *what would an image of a tractor look like if it had been generated in a city rather than on a grassy field?*, and appending them to training distributions to remove any spurious correlation between location-time and object distributions.

References

[1] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Arjovsky, M., Chaudhuri, K., and Lopez-Paz, D. Throwing away data improves worst-class error in imbalanced classification. *arXiv preprint arXiv:2205.11672*, 2022.

[3] Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

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

[5] Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

[6] Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022. URL `https://arxiv.org/abs/2206.15475`.

[7] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

[8] Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization, 2020. URL `https://arxiv.org/abs/2010.15775`.

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

[10] Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[11] Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[12] Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Dvijotham, K., and Cemgil, T. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No]
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes]

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

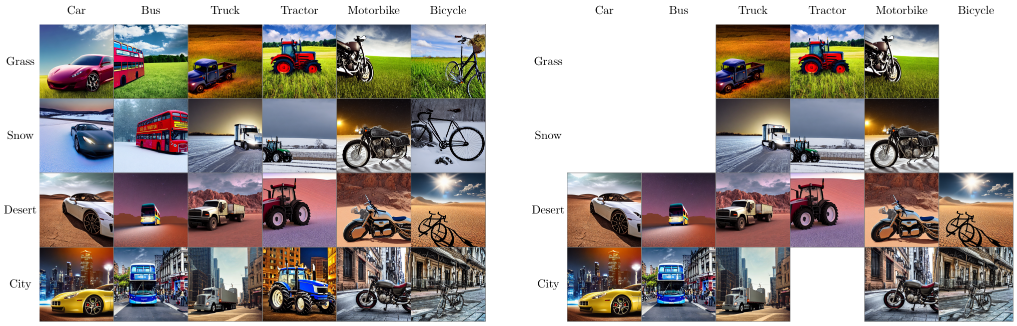
5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] figs/stable_diff_project_DAG_1.pdf.pdf

Figure 2: **DAG of the data generating process**: the generating factors are *object*, denoted by Y, *location* (L) and *time of day* (T), which causes of the data $\mathbf{X}$. The environment variable E varies in how it influences the generating factors depending on how we subsample the dataset, which simulates the effect of interventions.



a) Object balanced dataset

b) Object imbalanced dataset

Figure 3: **The images we present in each location distribution**: We vary the location and time of day in our training distribution as a means of simulating new environments. Within the object-balanced regime, we maintain the same distribution of objects within each time and location combination. In contrast, we specify a selection rule for the object imbalanced regime: cities have no examples of tractors, while grass and snow have no examples of cars, bicycles, or buses. Within this regime, the number of samples between each combination of object, location, and time is identical.