# Diversity Boosted Learning for Domain Generalization with A Large Number of Domains

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Machine learning algorithms minimizing the average training loss typically suffer from poor generalization performance. It inspires various works for domain generalization (DG), among which a series of methods work by $O(n^2)$ pairwise domain operations with $n$ domains, where each one is often costly. Moreover, while a common objective in the DG literature is to learn invariant representations against spurious correlations induced by domains, we point out its insufficiency and highlight the importance of alleviating spurious correlations caused by *objects*. Based on the observation that diversity helps mitigate spurious correlations, we propose a **D**iversity boosted tw**O**-level sa**M**pl**I**ng framework (`DOMI`) to efficiently sample the most informative ones among a large number of domains and data points. We show that `DOMI` helps train robust models against spurious correlations from both domain-side and object-side, substantially enhancing the performance of five backbone DG algorithms on Rotated MNIST and Rotated Fashion MNIST.

## 1 Introduction

The effectiveness of machine learning algorithms that minimize the average training loss relies on the IID hypothesis. However, distributional shifts between test and training data are usually inevitable. Under such circumstances, models trained by minimizing the average training loss are prone to sink into spurious correlations. These misleading heuristics only work well on some data distributions but can not be generalized to others that may appear in the test set. In domain generalization (DG) tasks, the data distributions are denoted as different domains. The goal is to learn a model that can generalize well to unseen ones after training on several domains. While lots of methods have been derived to efficiently achieve this goal and show good performance, there are two main drawbacks.

**Scalability.** With an unprecedented amount of applicable data nowadays, many datasets contain a tremendous amount of domains, massive data in each domain, or both. For instance, DrugOOD (Ji et al., 2022) is an out-of-distribution dataset curator and benchmark for AI-aided drug discovery. Datasets of DrugOOD contain hundreds to tens of thousands of domains. In addition to raw data with multitudinous domains, domain augmentation, leveraged to improve the robustness of models in DG tasks, can also lead to a significant increase in the number of domains. For example, HRM (Liu et al., 2021) generates heterogeneous domains to help exclude variant features, favoring invariant learning. Under such circumstances, training on the whole dataset in each epoch is computationally prohibitive, especially for methods training by pairwise domain operations, of which the computational complexity is $O(n^2)$ with $n$ training domains.

**Objective.** Numerous works in the DG field focus entirely on excluding or alleviating domain-side impacts. A general assumption in the DG field is that data in different domains share some "stable" features to form causal correlations. And a large branch of studies holds that the relationship between

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

these "stable" features and the outputs is **domain-independent** given certain conditions (Long et al., 2015; Hoffman et al., 2018; Zhao et al., 2018, 2019; Mahajan et al., 2021). We state that this objective is insufficient, and a simple counterexample is given as follows. We highlight the importance of mitigating spurious correlations caused by the objects for training a robust model.

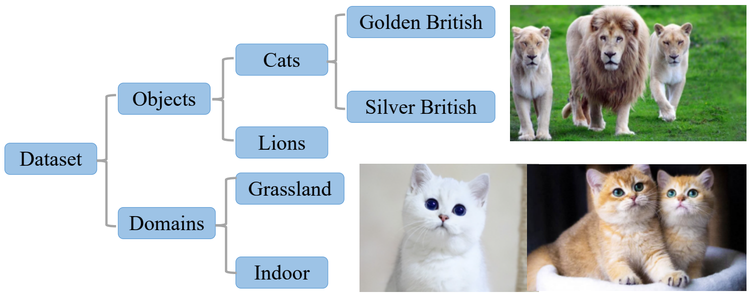



Figure 1: The training set of the counterexample. Cats are mainly silver British shorthair (body color of which is silvery white), rarely golden British shorthair (tan), and lions are all tan. As for the background, most lions are on the grassland while most cats are indoors.

Suppose our learning task is training a model to distinguish between cats and lions. The composition of the training set is shown in Figure 1, and the domain here refers to the images' backgrounds. In this example, the correlation between features corresponding to the body color of the objects and class labels is undoubtedly independent of domains. Moreover, it helps get high accuracy in the training set by simply taking the tan objects as lions and the white ones as cats. Unfortunately, if this correlation is mistaken for the causal correlation, the model is prone to poor performance once cat breed distribution shifts in the test set.

To tackle these two issues, we propose a diversity boosted two-level sampling framework named `DOMI` with the following major contributions: 1) To our best knowledge, this is the first paper to take impacts from the object side into account for achieving the goal of DG. 2) We propose `DOMI`, a diversity-boosted two-level sampling framework to select the most informative domains and data points for mitigating both domain-side and object-side impacts. 3) We demonstrate that `DOMI` substantially enhances the test accuracy of the backbone DG algorithms on two benchmarks.

## 2 Methods

We introduce our method `DOMI` by firstly presenting two key observations.

**Observation 1.** *Diverse domains of data help exclude spurious correlations.*

Consider a dataset $D_n = \{D^1, D^2, ..., D^n\}$ which is a mixture of data $D^d = \{(x_i^d, y_i^d)\}_{i=1}^{n_d}$ where $d$ is one domain of the ground set $D$ ($|D| = n$), $x_i^d$ and $y_i^d$ are the $i^{th}$ data and label from domain $d$ respectively, and $n_d$ is the number of data points in $D_d$. Suppose we now have dataset $D_k$ consisting of $k$ domains. On $D_k$, the distribution of data is $P^k(X, Y)$. A "good" set denoted by $C_k$ is a set containing "good" correlations that get high accuracy on $D_k$. The set of causal correlations is $C$. $C \subseteq C_k$ since causal correlations can definitely get good performance but "good" correlations for the $k$ domains may not be held in other domains, i.e., spurious correlations. The goal is to exclude as many spurious correlations as possible.

Given another domain $d_{k+1}$ to form dataset $D_{k+1}$ together with the former $k$ domains. The corresponding data distribution and the "good" set are $P^{k+1}(X, Y)$ and $C_{k+1}$, respectively. If $P^{k+1}(X, Y)$ is close to $P^k(X, Y)$, then most of the correlations in $C_k$ will still be "good" for $D_{k+1}$ and thus preserved in $C_{k+1}$. Nevertheless, if $d_{k+1}$ is a heterogeneous domain that can significantly change the distribution of data, then the "good" set after being constrained would be obviously smaller than the original one, i.e., $|C_{k+1}| << |C_k|$, showing that diverse domains help exclude spurious correlations and training on which helps obtain robust models.

We formally derive Proposition 1 to support Observation 1 that diversity helps mitigate spurious correlations, based on which `DOMI` is a diversity boosted sampling framework and the sampling scheme to obtain a heterogeneous subset is a critical part of `DOMI`. Determinantal Point Process (DPP) (Kulesza et al., 2012) sampling is a powerful diversity sampling method. Based on the similarity matrix (DPP kernel) among the samples, a draw from a DPP yields diversified subsets. Thus we incorporate DPP sampling into `DOMI`. As one option for the diversity sampling method in `DOMI`, DPP sampling can also be substituted with other sampling methods, which will be left as an interesting future direction.

### 2.1 invDANN

Domain-Adversarial Neural Networks (DANN) proposed by Ganin et al. (2016) is composed by Featurizer, Classifier, and Discriminator. Featurizer extracts features of data samples, Classifier learns to classify class labels of data, and Discriminator learns to discriminate domains. DANN set a gradient reversal layer between Featurizer and Discriminator to ensure Featurizer captures object-side features. Using the architecture of DANN, we let Classifier learn to classify domain labels while Discriminator learns to discriminate class labels. As an inverse version of DANN, `invDANN` trains a model whose Featurizer extracts only domain-side features, which serves as an important component of the proposed method `DOMI`.

**Observation 2.** *Excluding domain-induced spurious correlations is insufficient for achieving OOD generalization under the setting of DG.*

Figure 2 shows a structural causal model (SCM) that describes the data-generating process for the domain generalization task with object-side spurious correlations. The SCM divides data into two parts: domain-side and object-side. $\overline{x}$ of domain-side is the reason for domain-induced spurious correlations. For the object-side, the feature is further divided into $\dot{x}$ (causal features) and $\widehat{x}$ where $\widehat{x}$ is the reason for object-induced spurious correlations, just like the body color of objects in the lion-cat example. The three parts together make up the observed data. Thus even if we exclude all the domain-induced spurious correlations, i.e., entirely remove the effect from $\overline{x}$, we may still obtain object-induced spurious correlations resulting from $\widehat{x}$.

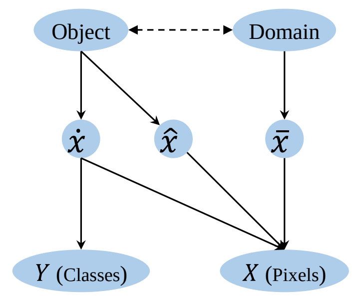



Figure 2: The Structural Causal Model for the data-generating process with a node $\hat{\mathbf{x}}$ leading to object-induced spurious correlations.

As Observation 2 shows that excluding only domain-induced spurious correlations is insufficient, we select diverse data batches among the selected domains to help mitigate object-induced spurious correlations in the level-two sampling. In the level-two-sampling, since we do not have available labels just like domain labels in the level-one-sampling, it is infeasible to utilize `invDANN` again to train a featurizer. So we instead use an ERM model since ERM is prone to taking shortcuts and learning spurious correlations (Zhang et al., 2022). Zhang et al. (2022) also leverages an ERM model to infer the spurious attributes in the unsupervised DG setting. Moreover, since domains attained by the level-one sampling contain diverse data with respect to the domain side, ERM can avert learning domain-induced spurious correlations. Combining these two, the ERM model is prone to relying on object-induced spurious correlations and thus can extract their informative representations. Then a similarity matrix between data batches is constructed with respect to this information. Based on which DPP sampling selects the data batches helping exclude object-induced spurious correlations.

### 2.2 `DOMI`: Diversity Boosted Two-level Sampling

Figure 3 shows the sampling procedure of `DOMI`, a diversity boosted two-level sampling framework. We present the details in Algorithm 1.

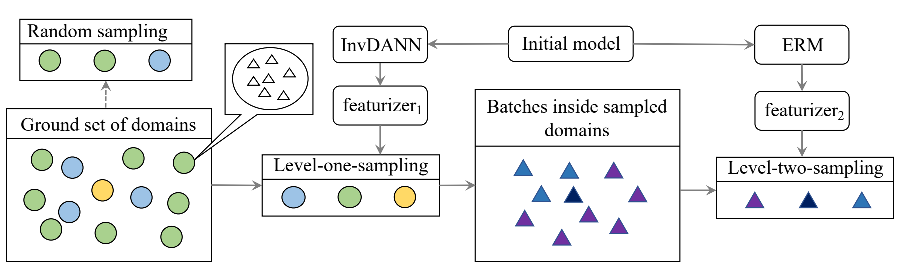



Figure 3: Illustration of the sampling procedure of `DOMI`. The solid arrow indicates the actual sampling flow, while the dotted arrow is used to demonstrate the difference between random sampling and `DOMI`.

**Algorithm 1:** Sampling Procedure of `DOMI`

**Input:** The whole training dataset:
$T = [\{(x_i^d, y_i^d)\}_{i=1}^{n_d} \text{ for } d \in \mathbf{D}]$
the proportion of domains ($\beta$) and batches ($\delta$) to be sampled

**Level-one-sampling**
Train an `invDANN` featurizer $f_{\overline{\theta}}$ on a randomly sampled subset of $T$ ;
**for** *d in* ***D*** **do**
  feat$_d \leftarrow \vec{0}$ ;
  **for** *i from 1 to* $n_d$ **do**
    feat$_d \leftarrow$ feat$_d + f_{\overline{\theta}}(x_i^d)$ ;
  feat$_d \leftarrow$ feat$_d \cdot \frac{1}{n_d}$ ;
Initialize similarity matrix $L_d = 0_{|\mathbf{D}|\times|\mathbf{D}|}$ ;
**for** $d_i$ *in* ***D*** **do**
  **for** $d_j$ *in* ***D*** **do**
    $L_d[i][j] = \|\text{feat}_{d_i} - \text{feat}_{d_j}\|_2$ ;
Obtain $\Omega = \text{DPP}(L_d, \beta \cdot |\mathbf{D}|) = [\{(x_i^d, y_i^d)\}_{i=1}^{n_d}$ for $d \in D], (D \subset \mathbf{D}, |D| = \beta \cdot |\mathbf{D}|)$ ;
**Level-two-sampling**
Divide $\Omega$ into $R = [\{(x_i^b, y_i^b)\}_{i=1}^{n}$ for $b \in \mathbf{B}]$ ;
Train an ERM featurizer $f_{\hat{\theta}}$ on $R$;
**for** *b in* ***B*** **do**
  Compute feat$_b$ in the same way as computing feat$_d$ in **Level-one-sampling**;
Computing similarity matrix $L_b$ ;
Return $S = \text{DPP}(L_b, \delta \cdot |\mathbf{B}|)$ ;

## 3 Experiments

We have investigated the performance of `DOMI` with five backbone DG algorithms on two simulated benchmarks (Rotated MNIST and Rotated Fashion MNIST), which show that `DOMI` can help substantially achieve higher test accuracy. Due to space constraints, and experimental settings, more results and analyses are deferred to Appendix C.1.

**Baselines.** For each one of the backbone algorithms, we set the baseline as training on domains selected by the random sampling scheme and denote it as $level_0$, compared to the level-one-sampling of `DOMI` and the full version of `DOMI` represented as $level_1$ and $level_2$, respectively. The proportion of minibatches selected in level-two-sampling ($\delta$) is a hyperparameter valued from 0 to 1. When $\delta$ equals 1, $level_2$ shrinks to $level_1$.

**Results and analysis** Table 1 shows the empirical results and we make the following observations: *Strong performance across datasets and algorithms.* Considering results on 2 datasets and 5 backbone DG algorithms, $level_1$ gives constant and apparent improvement compared to $level_0$. While $level_2$ may lead to slower growth in accuracy at the initial part of training as shown in Figure 6 because of using a smaller number of minibatches, it keeps outperforming $level_1$ and $level_0$ at later epochs. *The gap between test accuracy and maximal accuracy.* During training we observe that the test accuracy first rises to the peak value and then begins to decline along with the increase of validation accuracy. This reduction indicates a certain degree of overfitting to spurious correlations. Thus we further record the peak value of the test accuracy in each experiment and denote it as maximal accuracy. The distribution of test accuracy and maximal accuracy on MatchDG under different sampling schemes is shown in Figure 5. While the test accuracy of $level_0$ scatters, that of $level_2$ centers, and $level_2$ shrinks the gap between test accuracy and maximal accuracy.

Table 1: Average test accuracy. We repeat the experiment for 5 times on FISH and 20 times on the other algorithms with random seeds.

| Dataset | Sampling scheme | DANN | MatchDG | FISH | MMD | CORAL |
|---|---|---|---|---|---|---|
| Rotated MNIST | $level_0$ | 74.5 | 81.5 | 65.2 | 84.2 | 85.6 |
| | $level_1$ | 76.5 $\uparrow$ 2.0 | 83.6 $\uparrow$ 2.1 | 66.5 $\uparrow$ 1.3 | 87.2 $\uparrow$ 3.0 | 89.2 $\uparrow$ 3.6 |
| | $level_2$ | **78.6** $\uparrow$ 4.1 | **84.2** $\uparrow$ 2.7 | **66.6** $\uparrow$ 1.4 | **87.7** $\uparrow$ 3.5 | **89.6** $\uparrow$ 4.0 |
| Rotated Fashion MNIST | $level_0$ | 40.3 | 38.2 | 33.2 | 39.0 | 38.7 |
| | $level_1$ | 42.8 $\uparrow$ 2.5 | 39.7 $\uparrow$ 1.5 | 34.5 $\uparrow$ 1.3 | 41.8 $\uparrow$ 2.8 | 40.8 $\uparrow$ 2.1 |
| | $level_2$ | **43.5** $\uparrow$ 3.2 | **40.7** $\uparrow$ 2.5 | **35.8** $\uparrow$ 2.6 | **42.8** $\uparrow$ 3.8 | **42.1** $\uparrow$ 3.4 |

## Ethics statement

This study does not involve any of the following: human subjects, practices to dataset releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues.

## Reproducibility statement

To ensure the reproducibility of our empirical results, we present the detailed experimental settings in Appendix C.1 in addition to the main text. Besides, we will further provide the source codes for reproducing results in our paper.

## References


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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. Pmlr, 2018.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022.

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*, pp. 5637–5664. PMLR, 2021.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pp. 6804–6814. PMLR, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pp. 7728–7738. PMLR, 2020.

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

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

# Appendix of DOMI

## Contents



## A Theoretical analysis

**Preliminaries.** Consider the universal set of domains $\mathcal{D}$, where each domain $d \in \mathcal{D}$ corresponds to a distribution $P_d$ over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{X}$ being the space of inputs and $\mathcal{Y}$ that of outputs. Our goal is to find a predictor $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ while we can only access the domains in $\mathcal{D}_{tr}$ where $\mathcal{D}_{tr} \subset \mathcal{D}$. We measure the quality of a prediction with a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to R_{\geq 0}$; and the quality of a predictor by its population loss on domain $d \in \mathcal{D}$, given by $\mathcal{L}_d(f) := E_{(x,y)\sim P_d}[\ell(f(x), y)]$.

**Definition 1** (Correlation)**.** *We define a correlation as a predictor $f = \omega \circ \phi$ where $\phi : \mathcal{X} \to \mathcal{Z}$ is a data representation and $\omega^* : \mathcal{Z} \to \mathcal{Y}$ is a classifier. The causal correlation $f^*$ satisfies $\phi^*$ elicits a invariant predictor (Arjovsky et al., 2019) on $\mathcal{D}$: $\omega^*$ simultaneously optimal for all domains, i.e., $\forall d \in \mathcal{D}, \omega^* \in argmin_{\omega:\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_d(\omega \circ \phi^*)$.*

Notably, Definition 1 requires that $\phi$ and $\omega$ are unrestricted in the space of all (measurable) functions. However, we learn $\phi$ and $\omega$ being restricted to only access domains in $\mathcal{D}_{tr}$, a small subset of $\mathcal{D}$. For this to be feasible, it is natural to add a restriction that $\phi \in \Phi$ and $\omega \in \Omega$ for suitable classes $\Phi$ of functions mapping $\mathcal{X} \to \mathcal{Z}$ and W of functions mapping $\mathcal{Z} \to \widehat{\mathcal{Y}}$.

**Assumption 1.** *$argmin_{\omega:\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_d(\omega \circ \phi) = \{\omega | \mathcal{L}_d(\omega \circ \phi) \leq \delta\}$, where $\delta > 0$ is a constant.*

**Definition 2.** *Consider a domain set $\mathcal{D}_s$, on which the set of invariant predictors, $\mathcal{I}(\mathcal{D}_s)$, is the set of all predictors $f$ satisfies following:*

- *$f = \omega \circ \phi$ with $(\omega, \phi) \in \Omega \times \Phi$;*
- *for all $d \in \mathcal{D}_s$, $\omega \in argmin_{\bar{\omega}:\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_d(\bar{\omega} \circ \phi)$.*

**Lemma A.1.** *Based on Definition 1 and Definition 2, we can trivially derive: for any nonempty set $\bar{\mathcal{D}} \subseteq \mathcal{D}, f^* \in \mathcal{I}(\bar{\mathcal{D}})$.*

**Definition 3** (Diversity)**.** *We use Integral Probability Metric (Müller, 1997) to measure the diversity between domains. For domain d and $\bar{d}$, the diversity is defined as:*

$$Div(P_d, P_{\bar{d}}) = Div(P_d, P_{\bar{d}}, \mathcal{G}) = \sup_{g\in\mathcal{G}} |E_{P_d}[g(x,y)] - E_{P_{\bar{d}}}[g(x,y)]|$$

*Where $\mathcal{G}$ is a class of bounded functions. When we let $g(x,y) = \ell(f(x), y)$ and $\mathcal{G} = \mathcal{F} = \Omega \times \Phi$, the diversity is:*

$$\begin{aligned} Div(P_d, P_{\bar{d}}) = Div(P_d, P_{\bar{d}}, \Omega, \Phi) &= \sup_{\omega\in\Omega,\phi\in\Phi} |\mathbb{E}_{P_d}[\ell(\omega \circ \phi(x), y)] - \mathbb{E}_{P_{\bar{d}}}[\ell(\omega \circ \phi(x), y)]| \\ &= Div(P_d, P_{\bar{d}}, \mathcal{F}) = \sup_{f\in\mathcal{F}} |\mathcal{L}_d(f) - \mathcal{L}_{\bar{d}}(f)| \end{aligned}$$

Consider we have a domain set $\mathcal{D}_k = \{d_1, d_2 ..., d_k\}$ and the corresponding $\mathcal{I}(\mathcal{D}_k) = \{f_1, f_2 ..., f_m\}$. And now we get one more domain $d_{k+1}$ to form $\mathcal{D}_{k+1}$ . According to Lemma A.1, the causal correlation $f^* \in \mathcal{I}(\mathcal{D}_{k+1})$, so a informative domain $d_{k+1}$ which helps exclude spurious correlations leads to $|\mathcal{I}(\mathcal{D}_{k+1})|] < m$.

**Proposition 1** (Diverse domains help exclude spurious correlations)**.** *If* $d_{k+1}$ *satisfies that:* $\max_{d \in \mathcal{D}_k, f_i \in \mathcal{I}(\mathcal{D}_k)} Div(d_{k+1}, d) + \mathcal{L}_d(f_i) \leq \delta$, *then* $\mathcal{I}(\mathcal{D}_k) = \mathcal{I}(\mathcal{D}_{k+1})$.

*Proof.* Without loss of generality, we first conduct analysis on $f_t$ of $\mathcal{I}(\mathcal{D}_k)$. For $f_t$:

$$\max_{d \in \mathcal{D}_k} |\mathcal{L}_{d_{k+1}}(f_t) - \mathcal{L}_d(f_t)| + \mathcal{L}_d(f_t) \leq \max_{d \in \mathcal{D}_k} Div(d_{k+1}, d) + \mathcal{L}_d(f_t)$$

$$\max_{d \in \mathcal{D}_k} Div(d_{k+1}, d) + \mathcal{L}_d(f_t) \leq \max_{d \in \mathcal{D}_k, f_i \in \mathcal{I}(\mathcal{D}_k)} Div(d_{k+1}, d) + \mathcal{L}_d(f_i) \leq \delta$$

When

- $\mathcal{L}_{d_{k+1}}(f_t) - \mathcal{L}_d(f_t) < 0:$

$$\mathcal{L}_{d_{k+1}}(f_t) < \mathcal{L}_d(f_t) \leq \delta$$

- $\mathcal{L}_{d_{k+1}}(f_t) - \mathcal{L}_d(f_t) \geq 0:$

$$\mathcal{L}_{d_{k+1}}(f_t) = \max_{d \in \mathcal{D}_k} \mathcal{L}_{d_{k+1}}(f_t) - \mathcal{L}_d(f_t) + \mathcal{L}_d(f_t) = \max_{d \in \mathcal{D}_k} |\mathcal{L}_{d_{k+1}}(f_t) - \mathcal{L}_d(f_t)| + \mathcal{L}_d(f_t) \leq \delta$$

$\mathcal{L}_{d_{k+1}}(f_t) \leq \delta$, we get $f_t \in \mathcal{I}(\mathcal{D}_{k+1})$ for any $t \in \{1, 2..., m\}$, thus $\mathcal{I}(\mathcal{D}_k) = \mathcal{I}(\mathcal{D}_{k+1})$ $\square$

# B The Simulated Dataset

Table 2: The simulated dataset of the toy example. From these 12 data points, we sample 6 for training.

| | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | $D_6$ | $D_7$ | $D_8$ | $D_9$ | $D_{10}$ | $D_{11}$ | $D_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| $X_2$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| $X_3$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $X_4$ | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| $Y$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

# C Experimental Details

## C.1 Settings and results

**Datasets.** To satisfy the setting of a large number of domains, we extend the original simulated benchmarks on MNIST and Fashion MNIST by Piratla et al. (2020) from rotating images 15° through 75° in intervals of 15° to intervals of 1° in the training set, i.e., 61 domains in total. And we get test accuracy on the test set which rotates images either 0° or 90°. Moreover, while the original datasets rotate the same images for different degrees, we extend them to fit the real cases in DG tasks. We generate indices using different random seeds to select images from MNIST and Fashion MNIST for each domain before rotating. Appendix D gives examples to show how spurious correlations can occur in the two datasets.

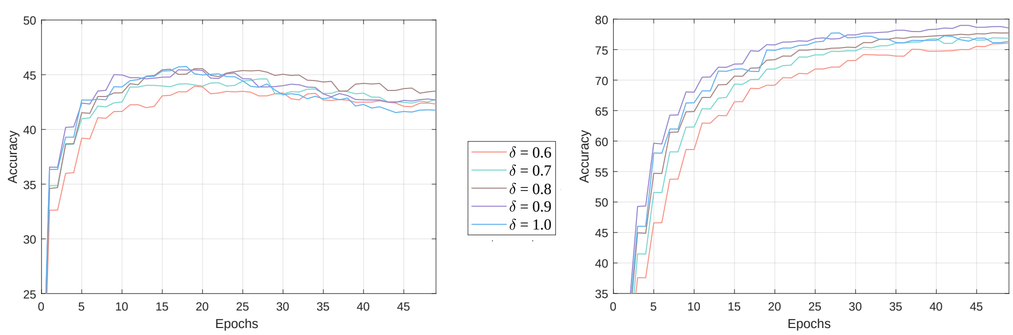
**Backbones.** We take MatchDG (Mahajan et al., 2021), FISH (Shi et al., 2021), CORAL (Sun & Saenko, 2016), MMD (Li et al., 2018) and DANN (Ganin et al., 2016) as backbone algorithms. The former four algorithms work by pairwise domain operations, leading to $O(n^2)$ computational complexity with $n$ domains and thus prohibitive to be scaled to DG tasks with multitudinous domains. It is essential for them to sample the most informative domains. We further incorporate DANN as one of the backbone algorithms since `DOMI` can not only efficiently select domains by its first level of sampling but can help deal with circumstances where each domain contains massive data by the second level of sampling.

**Hyperparameters.** For DANN, the training epochs are set to be 50. MatchDG is a two-phase method, and in our experiment, we set 30 epochs of training for phase 1 and 25 epochs for phase 2. While $level_1$ gets higher accuracy on Rotated MNIST and $level_2$ shows better performance on Fashion MNIST, they all outperform $level_0$, i.e., randomly sampling. The training epochs of FISH are set to be 5. Each epoch contains 300 iterations and we observe test accuracy every 30 iterations. And in Figure 6 we slightly abuse epoch to mean the time we obtain test accuracy. Unlike MatchDG and DANN, fish needs to sample domains in each iteration instead of training on one list of domains. Sampling domains in each iteration will result in great computational overhead compared to randomly sampling. Thus we just sample 30 domain lists containing diverse domains using level-one-sampling of `DOMI` and repeatedly train the model on these domain lists(one list for one iteration) for $level_1$. As for $level_2$, we further utilize level-two sampling to sample data batches of each domain in the domain lists for training. The former 3 DG algorithms utilize SGD optimizer with a learning rate of 0.01, weight decay $5 \times 10^{-4}$, and momentum 0.9. The training epochs of MMD and CORAL are set as 30. These two algorithms leverage Adam optimizer with a learning rate of 0.001 and weight decay of 0. All five algorithms use the Resnet18 model. Within each backbone algorithm, we keep factors including learning rate, batch size, choice of the optimizer, and model architecture the same for $level_0$, $level_1$ and $level_2$ to highlight the effect of different sampling schemes. It's worth noting that we do no comparison between the backbone algorithms since we do not conduct meticulous hyperparameter tuning for them.

**Model selection.** During training, we use a validation set to measure the model's performance. The test accuracy of the model is updated after an epoch if it shows better validating performance. That is, we save the model with the highest validation accuracy after the training procedure, obtain its test accuracy, and report results. For Rotated MNIST and Rotated Fashion MNIST, data from only source domains (rotation degree is from 15 ° to 75 °) are used to form the validation set.

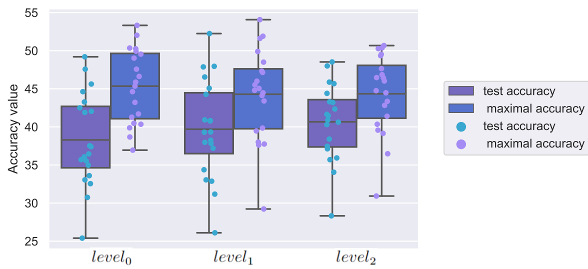
**Empirical results** Figure 5 show test accuracy and maximal accuracy among 20 times of repeated experiments with random seeds leveraging different sampling levels on Rotated Fashion MNIST and Rotated MNIST. Among training epochs, the test accuracy rises to the peak value and then declines along with the increase of validation accuracy. In this figure, maximal accuracy represents the peak value. Each tiny circle represents one time of the experiment, of which the vertical location corresponds to the accuracy value. The horizontal line inside each box indicates the mean value.

**The choice of $\delta$.** A smaller $\delta$ helps efficiently mitigate strong object-induced spurious correlations and speed up training, but when the impact from the object side is weak, a small $\delta$ leads to a waste of training data. In the experiment, we observe that a relatively small $\delta$ is more beneficial for Rotated Fashion MNIST while a large $\delta$ works better on Rotated MNIST. Figure 4 shows the results of different $\delta$.

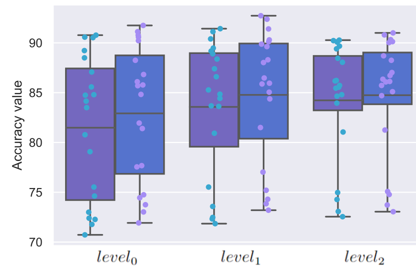



(a) Rotated Fashion MNIST (b) Rotated MNIST

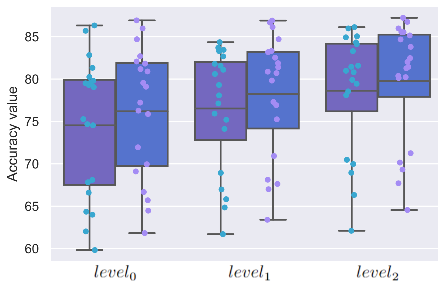
Figure 4: Average test accuracy of 20 experiments with random seeds during 50 epochs with different $\delta$ on Rotated Fashion MNIST and Rotated MNIST of DANN. $\delta = 1.0$ corresponds to `DOMI` with only level one.



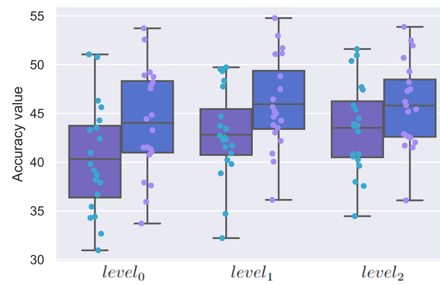

(a) MatchDG on Rotated Fashion MNIST



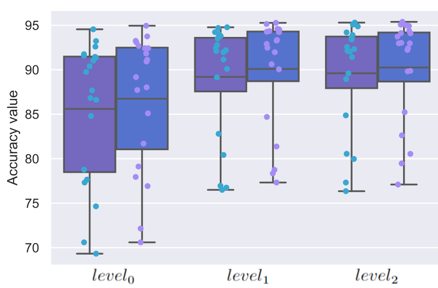

(b) MatchDG on Rotated MNIST



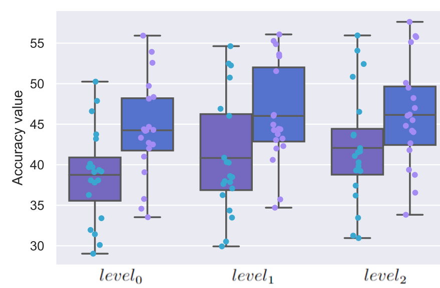

(c) DANN on Rotated MNIST



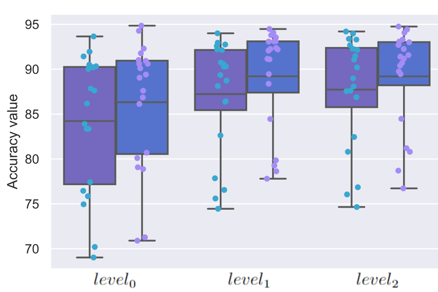

(d) DANN on Rotated Fashion MNIST



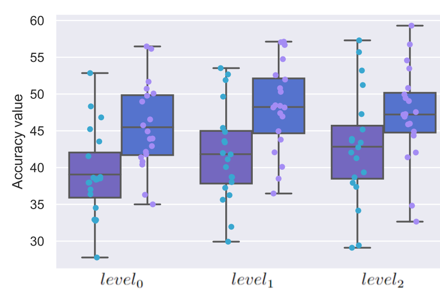

(e) CORAL on Rotated MNIST




(f) CORAL on Rotated Fashion MNIST

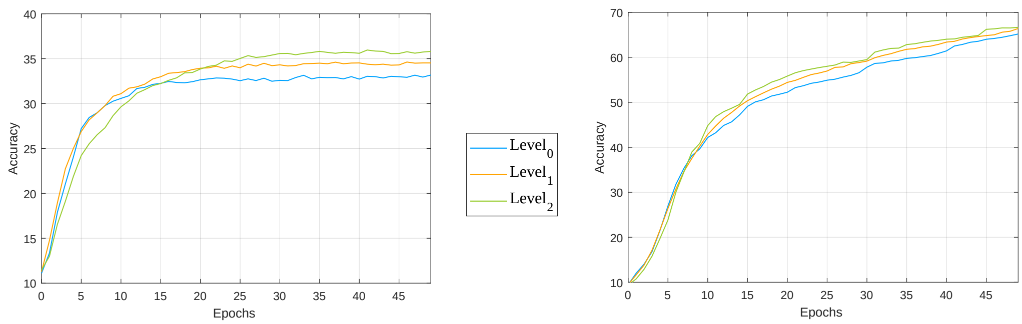



(g) MMD on Rotated MNIST




(h) MMD on Rotated Fashion MNIST

Figure 5: Boxplot of test accuracy and maximal accuracy among 20 repeated experiments with random seeds leveraging different sampling levels on Rotated Fashion MNIST and Rotated MNIST. Among training epochs, the test accuracy rises to the peak value and then declines with the increase of validation accuracy. In this figure, maximal accuracy represents the peak value. Each tiny circle represents one time of the experiment, of which the vertical location corresponds to the accuracy value. The horizontal line inside each box indicates the mean value.




(a) Rotated Fashion MNIST (b) Rotated MNIST

Figure 6: Average test accuracy of 5 experiments with random seeds during 50 epochs under different sampling schemes of FISH.

### C.2 Experiments on iwildcam

WILDS (Koh et al., 2021) is a curated collection of benchmark datasets representing distribution shifts faced in the wild. As one dataset in WILDS, iwildcam contains photos of wild animals and 324 different camera traps are taken as domains. The data of iwildcam is extremely unbalanced, while part of the domains contains less than 20 photos, some domains contain over 2000 ones. In the original experiments of Shi et al. (2021), iwlidcam is divided into batches in each domain. FISH samples 10 batches from different domains for training in each iteration. The sampling probability of one batch in a domain is proportional to the number of batches left in this domain. This sampling scheme is taken as $level_0$ here and we refer to the result of (Shi et al., 2021). In each iteration, $level_1$ samples 10 batches based on DPP using `invDANN`, $level_2$ first samples 10 batches in the level-one-sampling and among them selects 6 batches in the level-two-sampling. Under the same setting in the original experiments, the results on iwildcam of FISH are shown in Table 3 .

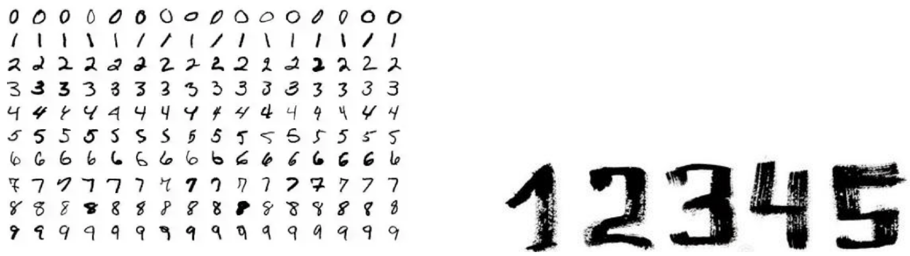
Table 3: Macro F1 score of FISH on iwildcam under three sampling schemes

| | $level_0$ | $level_1$ | $level_2$ |
|---|---|---|---|
| Iwildcam | 22.0 | 22.8 | **23.4** |

Although `DOMI` gets a higher Macro F1 score, it leads to a much larger computational overhead since it needs to do sampling in each iteration. Moreover, for DANN and MatchDG, Macro F1 of diverse domains may be significantly lower than randomly sampled domains because of the unbalanced data, i.e., the diverse domains may contain much fewer data compared to the randomly sampled domains. It would be significant future work to tackle the issues of extremely imbalanced data and computational overhead for algorithms that need to do sampling for multi-times.

## D How can spurious correlations occur in the two datasets?

It's much easier to differentiate the rotation degree than to discriminate the objects. This can be empirically verified since it only needs about 30 epochs for a model to achieve over 98% validation accuracy of classifying 61 different degrees while 50 epochs to achieve no more than 97% and 88% validation accuracy of classifying 10 different objects on rotated MNIST and Fashion MNIST, respectively. Thus if a certain class label is closely associated with a certain rotation degree in the training set, recognizing objects by actually recognizing the rotation degree can be a shortcut and domain-induced spurious correlation, just like classifying cats and lions using the background in the toy example. As for object-induced spurious correlation, on rotated MNIST, the handwriting is the feature of the object, however, it can also be the spurious correlation. For example, in Figure 7, let's focus on the number "1" and "7". After training on Figure 7a, can the model correctly recognize "1" in Figure 7b instead of wrongly taking it as "7"?




(a) Hypothetical training set (b) Hypothetical test set

Figure 7: Two figures to illustrate the impact of object-induced spurious correlations on MNIST.

On Fashion MNIST, assume we take the data in Figure 8 as the training set. The majority of the data points for Shirt are darker than Coat. When differentiating between Shirt and Coat, a model can simply take the bright ones as coats and the dark ones as shirts to obtain high training accuracy. However, what if the color of the Shirt and Coat is similar in the test set?

Thus our proposed framework set two levels of sampling to mitigate the impacts of domain-side and object-side, the sampling is, in fact, a rebalance procedure of data.

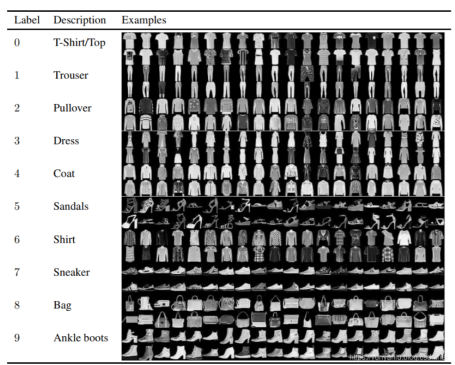



Figure 8: The figure to illustrate the impact of object-induced spurious correlations on Fashion MNIST.