DIVERSITY BOOSTED LEARNING FOR DOMAIN GENERALIZATION WITH A LARGE NUMBER OF DOMAINS

Anonymous authors
Paper under double-blind review

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 the insufficiency of it 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 assumption that the testing and training data are identically drawn from the same distribution, known as the IID hypothesis. However, distributional shifts between testing and training data are usually inevitable due to data selection biases or unobserved confounders that widely exist in real-life data. Moreover, the data distribution of the training set is likely to be imbalanced. Specific clusters may contain the majority of data samples, while the others are only a tiny fraction of the training set. Under such circumstances, models trained by minimizing average training loss are prone to sink into spurious correlations. These misleading heuristics only work for most training examples but can not generalize to data from other distributions that may appear in the test set and thus suffer from poor generalization performance. 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. For example, an image classifier should be able to discriminate the objects whatever the image's background is. 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, or massive data in each domain, or both. For instance, WILDS (Koh et al., 2021) is a curated collection of benchmark datasets representing distribution shifts faced in the wild. Among these datasets, some contain thousands of domains and OGB-MolPCBA (Hu et al., 2020b) contains even more than one hundred thousand. Besides WILDS, 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., 2021a) 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 such as MatchDG (Mahajan et al., 2021) and FISH (Shi et al., 2021b), training by pairwise 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 the causal correlations. And a large branch of studies hold that the relationship between 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 induced from the object side for training a robust model.

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.

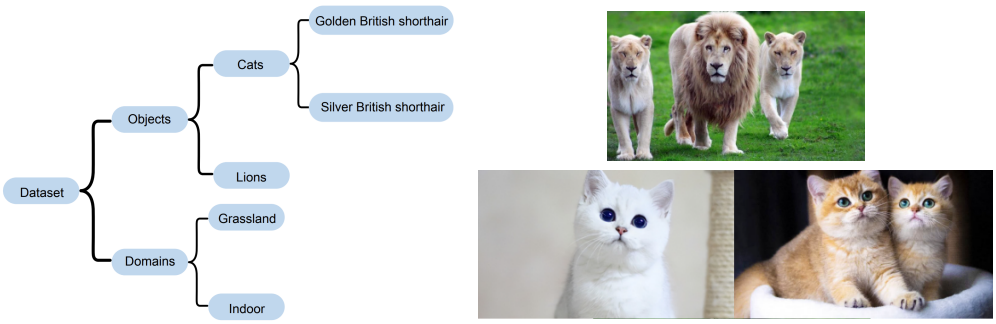


Figure 1: The training set of the counterexample. Cats are mainly silver British shorthair (body color of which is silvery white), rarely golden British shorthair (tan), and lions are all tan. As for the background, most lions are on the grassland while most cats are indoors.

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, a sampling framework that selects the most informative domains or data points to help mitigate impacts from both domain-side and object-side is essential. Under the setting of a large number of domains and domains with massive data points, we propose a diversity boosted two-level sampling framework named `DOMI`. Since we will later state that diversity helps alleviate spurious correlations, a sampling scheme to select diverse domains or data points is a critical part of `DOMI`. In this paper, we incorporate Determinantal Point Process (DPP) (Kulesza et al., 2012) sampling into `DOMI` as one choice of diversity sampling methods. As one option for the diversity sampling method in `DOMI`, DPP sampling can be substituted with other sampling methods. In the first level of `DOMI`, with domain labels, we utilize the inverse version of DANN (Ganin et al., 2016) denoted as `invDANN` to train a featurizer to capture the information of the domains, based on which a subset of domains is selected by the diversity sampling method. In the second level, it is infeasible to leverage `invDANN` again to learn about the attributes incurring object-side spurious correlations without available labels. So we instead use an ERM model to infer the spurious attributes as in (Zhang et al., 2022). With the information extracted by the ERM model, the second level of `DOMI` samples a subset among the data points of the domains selected by the first level. Extensive experiments show that `DOMI` helps efficiently alleviate both domain-side and object-side spurious correlations, substantially enhancing the performance of the backbone DG algorithms on Rotated MNIST and Rotated Fashion MNIST. Our contributions can be summarized as follows:

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 Related Work

Domain Generalization. DG aims to learn a model that can generalize well to all domains including unseen ones after training on more than one domains (Blanchard et al., 2011; Wang et al., 2022; Zhou

et al., 2021; Shen et al., 2021). Among recent works on domain generalization, Ben-Tal et al. (2013); Duchi et al. (2016) utilize distributionally robust optimization (DRO) to minimize the worst-case loss over potential test distributions instead of the average loss of the training data. Sagawa et al. (2019) propose group DRO to train models against spurious correlations by minimizing the worst-case loss over groups to avoid suffering high losses on some data groups. Zhai et al. (2021) further use distributional and Outlier Robust Optimization (DORO) to address the problem that DRO is sensitive to outliers and thus suffers from poor performance and severe instability when faced with real, large-scale tasks. On the other hand, as Peters et al. (2016) and Rojas-Carulla et al. (2018) state that the predictor should be simultaneously optimal across all domains, (Arjovsky et al., 2019; Javed et al., 2020; Shi et al., 2021a; Ahuja et al., 2020a) leverage Invariant Risk Minimization (IRM) to learn features inducing invariant optimal predictors over training domains. However, Guo et al. (2021); Rosenfeld et al. (2020); Kamath et al. (2021); Ahuja et al. (2020b) point out that works with IRM lack formal guarantees, and IRM does not provably work with non-linear data. Koh et al. (2021) and Gulrajani & Lopez-Paz (2020) present an analysis to demonstrate that IRM fails to generalize well even when faced with some simple data models and fundamentally does not improve over standard ERM. Risk Extrapolation (V-REx) (Krueger et al., 2021) instead hold the view that training risks from different domains should be similar and achieves the goal of DG by matching the risks. Some works explore data augmentations to mix samples from different domains (Wang et al., 2020; Wu et al., 2020) or generate more training domains (Liu et al., 2021a,b) to favor generalization. Another branch of studies assume that data from different domains share some "stable" features whose relationships with the outputs are causal correlations and domain-independent given certain conditions (Long et al., 2015; Hoffman et al., 2018; Zhao et al., 2018, 2019). Among this branch of work, Li et al. (2018c); Ghifary et al. (2016); Hu et al. (2020a) hold the view that causal correlations are independent of domain conditioned on class label, and Muandet et al. (2013) propose DICA to learn representations marginally independent of domain.

MatchDG. Mahajan et al. (2021) state that learning representations independent of the domain after conditioning on the class label is insufficient for training a robust model. They propose MatchDG to learn correlations independent of domain conditioned on objects, where objects can be seen as clusters within classes based on similarity. To ensure the learned features are invariant across domains, a term of the distance between each pair of domains is added to the objective to be minimized.

FISH, MMD, CORAL. Another line of works promote agreements between gradients with respect to network weights (Koyama & Yamaguchi, 2020; Parascandolo et al., 2020; Rame et al., 2022; Mansilla et al., 2021; Shahtalebi et al., 2021). Among these works, FISH (Shi et al., 2021b) augments the ERM loss with an auxiliary term of gradient inner product between domains. By minimizing the loss and matching the gradients simultaneously, FISH encourages the optimization paths to be the same for all domains, favoring invariant predictions. MMD (Li et al., 2018b) and CORAL (Sun & Saenko, 2016) are another two matching methods besides MatchDG and FISH. MMD matches the distributions among different domains using Maximum Mean Discrepancy measure. In this way, the learned representation is supposed to be invariant for the training domains. MMD further aligns the matched distribution to an arbitrary prior distribution via adversarial feature learning, aiming to prevent the representation from overfitting to the training domains. Then the learned representation is expected to generalize well on the test domains. CORAL instead matches the second-order statistics of the distributions across domains. Specifically, CORAL concurrently minimizes the ERM loss and the difference in learned feature covariances across domains. As a simple yet effective method, CORAL shows state-of-the-art performance on various tasks for OOD generalization (Gulrajani & Lopez-Paz, 2020).

DANN. Besides gradients, some approaches enforce agreements between features and align the features with adversarial methods (Li et al., 2018a; Gong et al., 2016). As one of these approaches, DANN (Ganin et al., 2016) incorporates the structure named domain discriminator to implement adversarial training based on the theory that an ideal classifier for cross-domain shifts should be able to distinguish different classes while cannot learn to identify the domain. `DOMI` takes use of an inverse version of DANN denoted as `invDANN` to learn domain-side features and help select the most informative domains.

DPP. Determinantal Point Process (DPP) (Kulesza et al., 2012) is a point process that mimics repulsive interactions. A draw from a DPP yields diversified subsets based on a similarity matrix (DPP kernel) of samples to be selected. While it shows powerful performance in selecting heterogeneous data, DPP sampling relies on an eigendecomposition of the DPP kernel, whose cubic complexity is a huge impediment. To address this problem, Li et al. (2016) suggest first construct an approximate

probability distribution to the true DPP and then efficiently samples from this approximate distribution. As one choice of diversity sampling, DPP sampling is incorporated into DOMI to help select the most informative domains and data points. It can be replaced with other diversity sampling schemes.

Discussions. Although MatchDG, FISH, MMD and CORAL perform well in domain generalization tasks, the matching procedure between domains means their computational complexity is $O(n^2)$ with n training domains. When n is large, it will inevitably slow down the training process. Therefore, we must select the most informative domains from all the training domains. Inspired by Liu et al. (2021a) that heterogeneous training domains help to learn invariant features since more variant features can be excluded, we conduct an analysis of diversity and spurious correlations to further state it.

3 Diversity Helps Mitigate Spurious Correlations

Spurious correlations essentially result from imbalanced data. While specific clusters may contain the majority of data points, the others are only a tiny fraction of the training set. Suppose a correlation is easy to be found and is held by most of the data. In that case, algorithms minimizing the average loss like ERM may simply take this correlation as the causal correlation. When we sample diverse data, we in fact re-balance them and help mitigate spurious correlations. We verify this observation with a toy example.

For the task and dataset mentioned above (Figure 1), we further suppose our featurizer extracts 4 features with a binary value as shown in Table 1.

Table 1: Details of the features and the label. X_1 to X_3 correspond to features of the object and X_4 corresponds to features of the domain.

	X_1 : Mane	X_2 : Proportion of face	X_3 : Body color	X_4 : Background	y
0	no mane	short face	white	indoors	cat
1	have a mane	long face	tan	grassland	lion

Then $X_1 + X_2 \geq 1 \Rightarrow y = 1$ is the causal correlation since the proportion of lions' faces is longer than that of cats, and X_2 may be wrongly computed to 0 for male lions because of the existence of mane. $X_3 = 1 \Rightarrow y = 1$ is the object-induced spurious correlation (Abbrev. OSC) and $X_4 = 1 \Rightarrow y = 1$ is the domain-induced spurious correlation (Abbrev. DSC). Details of our simulated dataset is shown in Appendix A.

Suppose we get 6 of these 12 data samples for training where 3 of 6 come from cats and the other 3 are from lions. There are 4 sampling methods denoted as S_1 to S_4 to be picked: random sampling, diverse sampling with respect to the object features (X_1, X_2 and X_3), diverse sampling with respect to (X_4), and diverse sampling with respect to all 4 features. For S_2 to S_4, we use Manhattan Distance on the corresponding features to measure the similarity between data. After conducting the similarity matrix, we finally use DPP sampling to select data points. Table 2 shows the average training accuracy of OSC and DSC. When the spurious correlations get lower training accuracy, they are less likely to be mistaken for causal correlation, favoring exploration of the causal correlations.

Table 2: We use each sampling method to select 30 batches of data for training, on which the average accuracy of two kinds of spurious correlations is computed. When spurious correlations get lower accuracy during training, they are more likely to be excluded. Here we use DSC and OSC to denote domain-induced spurious correlation and object-induced spurious correlation, respectively.

Sampling Method	Accuracy of OSC	Accuracy of DSC
S_1	0.86	0.68
S_2	0.77	0.66
S_3	0.85	0.50
S_4	0.78	0.49

We take the data batches sampled by S_1 as base-batches. Random sampling preserves the imbalance of data since a data point is more likely to be sampled into a subset when it appears more often in the whole dataset. For base-batches sampled by S_1, both OSC and DSC get high accuracy and are thus prone to be wrongly treated as causal correlations. S_2 selects diverse data pertaining to object

features. Data batches sampled by S_2 get lower accuracy for OSC than base-batches, which means S_2 reduces the probability of taking OSC as causal correlation. However, data batches sampled by S_2 get almost the same result for DSC. S_3 selects diverse data on domain-feature X_4. For these batches of data, DSC gets lower accuracy than base-batches and is less likely to be taken as causal correlation, while OSC has a similar result. S_4 selects data with heterogeneity with regard to all 4 features. Compared to base-batches, the data batches selected by S_4 get lower accuracy on both spurious correlations.

4 METHODS

Figure 2 shows the sampling procedure of DOMI, a diversity boosted two-level sampling framework.

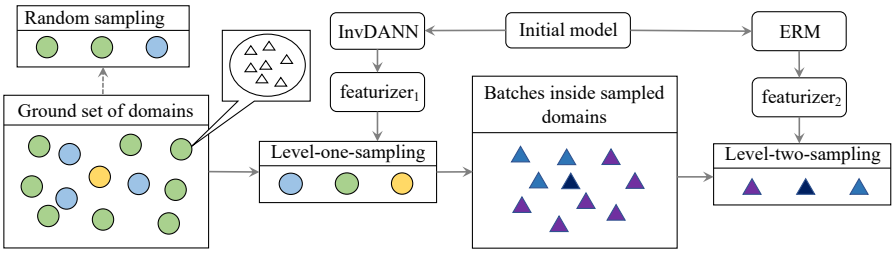


Figure 2: Illustration of the sampling procedure of DOMI. The solid arrow indicates the actual sampling flow, while the dotted arrow is only used to demonstrate the difference between random sampling and DOMI.

4.1 DIVERSITY BOOSTED SAMPLING FRAMEWORK

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 ($|D| = n$), $\mathbf{x}_i^d$ and $\mathbf{y}_i^d$ 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. Our 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.

4.1.1 DIVERSITY SAMPLING METHOD

As observation 1 states that diversity helps mitigate spurious correlations, 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) sampling is a powerful diversity sampling method. Based on the similarity matrix between the samples, a draw from a DPP yields diversified subsets. Thus we incorporate DPP sampling into DOMI as one choice of diversity sampling methods. As one option for the diversity sampling method in DOMI, DPP sampling can also be substituted with other sampling methods.

4.2 LEVEL-ONE-SAMPLING

In the level-one-sampling, we select diverse domains to help mitigate domain-induced spurious correlations. Since we aim to sample diverse domains, we have to learn about the domains. We

propose an inverse version of DANN denoted as `invDANN` to train a model to capture the domain information.

4.2.1 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. Since DANN aims to obtain a model that can not differentiate domains to ensure Featurizer captures domain-independent features, Discriminator is connected to the Featurizer via a gradient reversal layer that multiplies the gradient by a certain negative constant during backpropagation. Gradient reversal ensures that the feature distributions over the two domains are made similar, thus resulting in domain-independent features. Using the architecture of DANN, we let Classifier learn to classify domain labels of data while Discriminator learns to discriminate class labels. As an inverse version of DANN, `invDANN` trains a model that can classify domains while not distinguishing class labels. Thus we can get Featurizer extracting only domain-side features.

4.2.2 SAMPLING PROCEDURE

In the level-one-sampling of `DOMI`, we first use `invDANN` to train a featurizer. As mentioned in Section 4.2.1, the featurizer only extracts domain-side features. Then we use the featurizer to capture the information of domains and construct a similarity matrix between them. Based on the similarity matrix, DPP sampling selects the diverse domains.

4.3 LEVEL-TWO-SAMPLING

Observation 2 *Excluding domain-induced spurious correlations is insufficient for learning a robust model.*

Mahajan et al. (2021) have proposed that correlations independent of domain conditional on class ($\Phi(x) \perp\!\!\!\perp D|Y$) are not necessarily causal correlations if $P(\dot{x}|Y)$ changes across domains. Here $\Phi(x)$ is a featurizer to extract features and $\dot{x}$ represents the causal features. We now further propose that the condition is insufficient even if $\dot{x}$ is consistent across domains. A correlation incorporating features entirely from the objects can still be a spurious correlation. Figure 3 shows a structural causal model (SCM) that describes the data-generating process for the domain generalization task. 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 object-side, feature is further divided into $\dot{x}$ and $\widehat{x}$ where $\widehat{x}$ is the reason for object-induced spurious correlations, just like the body color of objects in the toy 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 $\widehat{x}$.

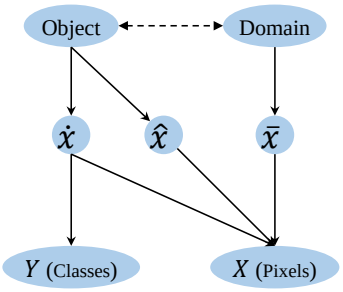


Figure 3: The Structural Causal Model for the data-generating process with a node $\hat{x}$ leading to object-induced spurious correlations.

4.3.1 SAMPLING PROCEDURE

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 leverage an ERM model to infer the spurious attributes in the unsupervised DG field. 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 information. 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.

4.4 DOMI

We present DOMI in Algorithm 1. Combining the two levels, DOMI finally gets a subset of the dataset to tackle the issue of scalability under the setting of tremendous domains and training on which helps obtain robust models against impacts from both domain-side and object-side.

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 (β) and batches (δ) to be sampled

Level-one-sampling
Train an invDANN featurizer $f_{\bar{\theta}}$ on T ;
for *d in* ***D*** **do**
  $\text{feat}_d \leftarrow \vec{0}$;
  for *i from 1 to* n_d **do**
    $\text{feat}_d \leftarrow \text{feat}_d + f_{\bar{\theta}}(x_i^d)$;
  $\text{feat}_d \leftarrow \text{feat}_d \cdot \frac{1}{n_d}$;
Initialize similarity matrix $L_d = O_{|\mathbf{D}| \times |\mathbf{D}|}$;
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 $\Omega = DPP(L_d, \beta \cdot |\mathbf{D}|) = [\{(x_i^d, y_i^d)\}_{i=1}^{n_d} \text{ for } d \in D]$, $(D \subset \mathbf{D}, |D| = \beta \cdot |\mathbf{D}|)$;
Level-two-sampling
Divide Ω into $R = [\{(x_i^b, y_i^b)\}_{i=1}^{n} \text{ for } b \in \mathbf{B}]$;
Train a 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 $S = DPP(L_b, \delta \cdot |\mathbf{B}|)$;

5 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. We also conduct experiments on iwildcam. Due to space constraints, the results and analysis are listed in Appendix B.2. The experimental settings and results are shown as follows.

5.1 CONFIGURATIONS

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 C 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., 2021b), CORAL (Sun & Saenko, 2016), MMD (Li et al., 2018b) 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.

Table 3: Average test accuracy of five algorithms. 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 ↑ 2.0	83.6 ↑ 2.1	66.5 ↑ 1.3	87.2 ↑ 3.0	89.2 ↑ 3.6
	$level_2$	**78.6** ↑ 4.1	**84.2** ↑ 2.7	**66.6** ↑ 1.4	**87.7** ↑ 3.5	**89.6** ↑ 4.0
Rotated Fashion MNIST	$level_0$	40.3	38.2	33.2	39.0	38.7
	$level_1$	42.8 ↑ 2.5	39.7 ↑ 1.5	34.5 ↑ 1.3	41.8 ↑ 2.8	40.8 ↑ 2.1
	$level_2$	**43.5** ↑ 3.2	**40.7** ↑ 2.5	**35.8** ↑ 2.6	**42.8** ↑ 3.8	**42.1** ↑ 3.4

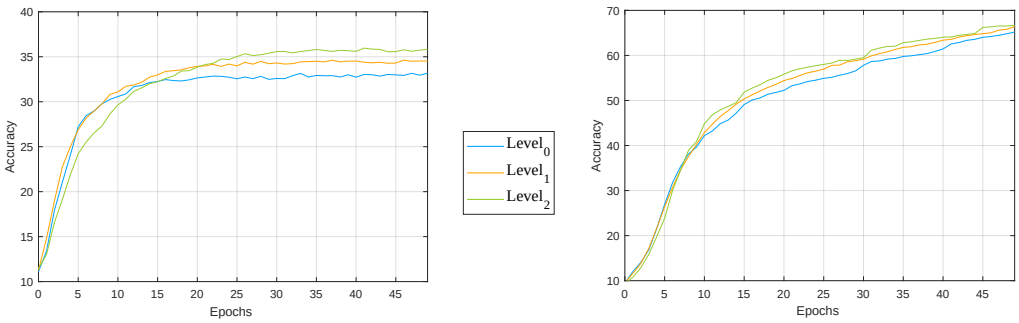
Baselines. For each one of the backbone algorithms, we set the baseline as training on domains selected by the random sampling scheme and denote it as $level_0$, compared to the level-one-sampling of `DOMI` and the full version of `DOMI` represented as $level_1$ and $level_2$, respectively. We sample 5 domains for training on Rotated MNIST and Rotated Fashion MNIST. The proportion of minibatches selected in level-two-sampling (δ) is a hyperparameter valued from 0 to 1. When δ equals 1, $level_2$ shrinks to $level_1$. Within each backbone algorithm, we keep factors including learning rate, batch size, choice of 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.

5.2 Empirical results and analysis

Table 3 shows the empirical results and we make the following observations:

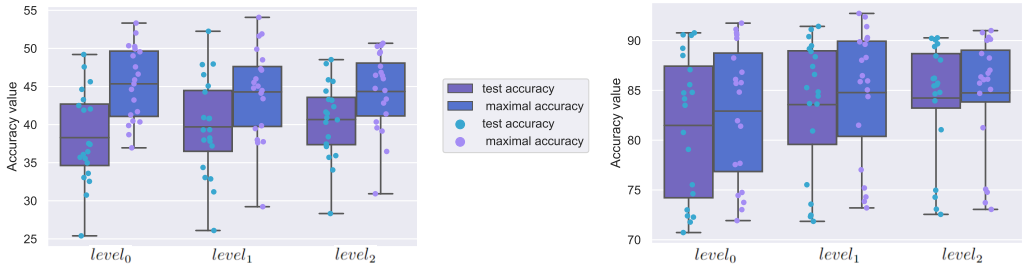
Strong performance across datasets and algorithms. Considering results on 2 datasets and 5 backbone DG algorithms, $level_1$ gives constant and apparent improvement compared to $level_0$. While $level_2$ may lead to slower growth in accuracy at the initial part of training as shown in Figure 4 because of using a smaller number of minibatches, it keeps outperforming $level_1$ and $level_0$ at later epochs.



(a) Rotated Fashion MNIST (b) Rotated MNIST

Figure 4: Average test accuracy of 5 experiments with random seeds during 50 epochs under different sampling schemes of FISH.

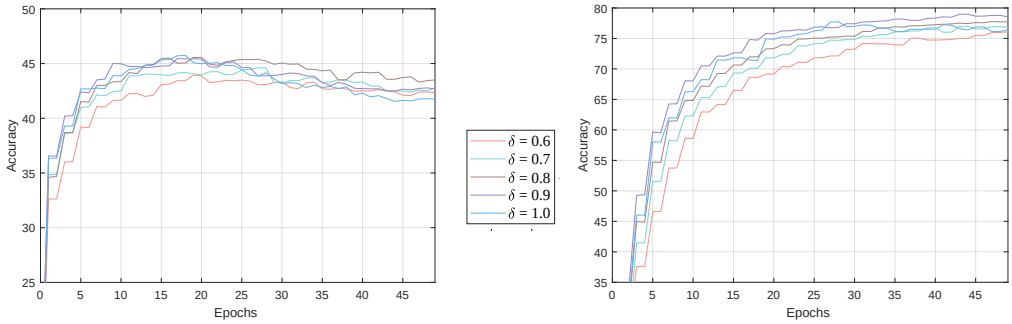
The gap between test accuracy and maximal accuracy. During training we observe that the test accuracy first rises to the peak value and then begins to decline along with the increase of validation accuracy. This reduction indicates a certain degree of overfitting to spurious correlations. Thus we further 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.



(a) Rotated Fashion MNIST (b) Rotated 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 experiment, of which the vertical location corresponds to the accuracy value. The horizontal line inside each box indicates the mean value.

The choice of δ. A smaller δ helps efficiently mitigate strong object-induced spurious correlations and speed up training, but when the impact from object-side is weak, a small δ leads to a waste of training data. In the experiment we observe that a relatively small δ is more beneficial for Rotated Fashion MNIST while a large δ works better on Rotated MNIST. Figure 6 shows the results of different δ.



(a) Rotated Fashion MNIST (b) Rotated MNIST

Figure 6: Average test accuracy of 20 experiments with random seeds during 50 epochs with different δ on Rotated Fashion MNIST and Rotated MNIST of DANN. $\delta = 1.0$ corresponds to `DOMI` with only level one.

6 Conclusion

Under the setting of a large number of domains and domains with massive data points, we propose a diversity boosted two-level sampling algorithm named `DOMI` to help sample the most informative subset of dataset. Empirical results show that `DOMI` substantially enhances the out-of-domain accuracy and gets robust models against spurious correlations from both domain-side and object-side.

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 B.1 in addition to the main text. Besides, we will further provide the source codes for reproducing results in our paper.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020b.

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

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

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.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848. PMLR, 2016.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

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.

Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pp. 292–302. PMLR, 2020a.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Khurram Javed, Martha White, and Yoshua Bengio. Learning causal models online. *arXiv preprint arXiv:2006.07461*, 2020.

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.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2021.

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.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. 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.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Efficient sampling for k-determinantal point processes. In *Artificial Intelligence and Statistics*, pp. 1328–1337. PMLR, 2016.

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, 2018a.

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, 2018b.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018c.

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

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernelized heterogeneous risk minimization. *arXiv preprint arXiv:2110.12425*, 2021b.

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.

Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6630–6638, 2021.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

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.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

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.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pp. 1546–1555. PMLR, 2021a.

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*, 2021b.

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

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3622–3626. IEEE, 2020.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *European Conference on Computer Vision*, pp. 540–555. Springer, 2020.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pp. 12345–12355. PMLR, 2021.

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.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. 2021.

Appendix of DOMI

CONTENTS

A The Simulated Dataset **14**

B Experiments **14**

B.1 Settings and results 14

B.2 Experiments on iwildcam 14

C How can spurious correlations occur in the two datasets? **16**

A THE SIMULATED DATASET

Table 4: 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

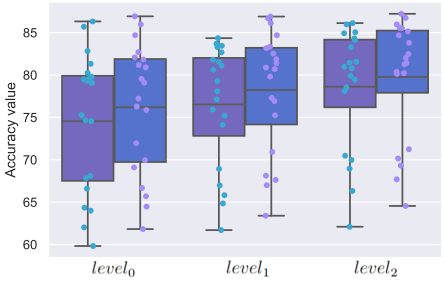
B EXPERIMENTS

B.1 SETTINGS AND RESULTS

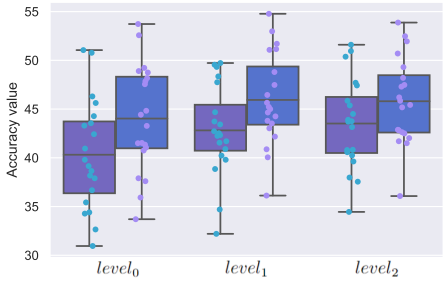
For DANN, the training epochs are set to be 50. MatchDG is a two-phase method, and in our experiment we set 30 epochs of training for phase 1 and 25 epochs for phase 2. While $level_1$ gets higher accuracy on Rotated MNIST and $level_2$ shows better performance on Fashion MNIST, they all outperform $level_0$, 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 4 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 learning rate 0.01, weight decay 5e-4 and momentum 0.9. The training epochs of MMD and CORAL are set as 30. These two algorithms leverage Adam optimizer with learning rate 0.001 and weight decay 0. All five algorithms use Resnet18 model. Figure 7 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 decline along with the increase of validation accuracy. In this figure maximal accuracy represents the peak value. Each tiny circle represents one times of experiment, of which the vertical location corresponds to the accuracy value. The horizontal line inside each box indicates the mean value.

B.2 EXPERIMENTS ON IWILDCAM

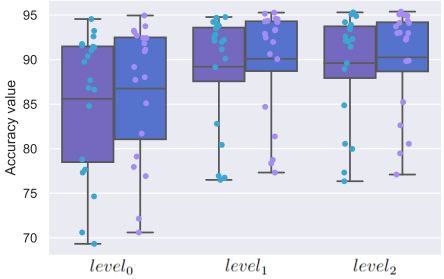
WILDS (Koh et al., 2021) is a curated collection of benchmark datasets representing distribution shifts faced in the wild. As one dataset in WILDS, iwildcam contains photos of wild animals and 324 different camera traps are taken as domains. The data of iwildcam is extremely unbalanced, while



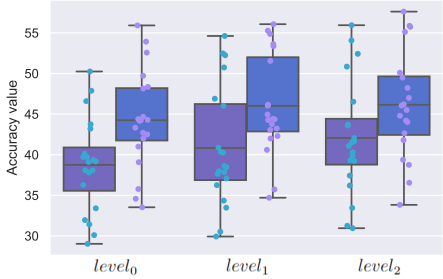
(a) DANN on Rotated MNIST



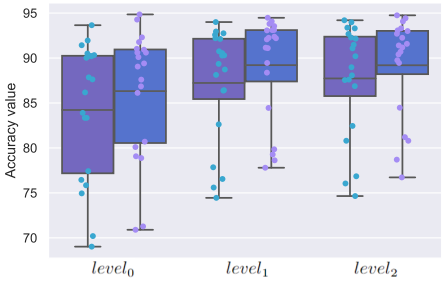
(b) DANN on Rotated Fashion MNIST



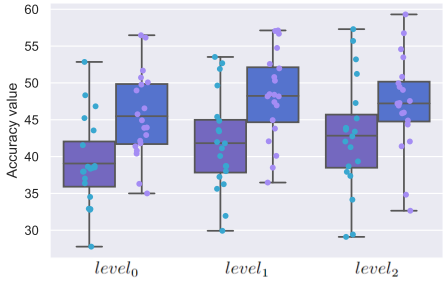
(c) CORAL on Rotated MNIST



(d) CORAL on Rotated Fashion MNIST



(e) MMD on Rotated MNIST



(f) MMD on Rotated Fashion MNIST

Figure 7: Boxplot of DANN, CORAL and MMD.

part of the domains contain less than 20 photos, some domains contain over 2000 ones. In the original experiments of Shi et al. (2021b), iwildcam 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., 2021b). 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 5 .

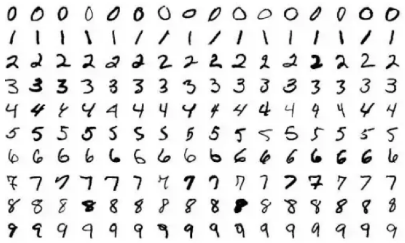
Table 5: Macro F1 score of FISH on iwildcam under three sampling schemes

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

Although `DOMI` gets 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 a significant future work to tackle the issues of extremely imbalanced data and computational overhead for algorithms that need to do sampling for multi-times.

C How can spurious correlations occur in the two datasets?

It's much easier to differentiate the rotation degree than discriminating 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 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 8, let's focus on the number "1" and "7". After training on Figure 8a, can the model correctly recognize "1" in Figure 8b instead of wrongly taking it as "7" ?



(a) Hypothetical training set

(b) Hypothetical test set

Figure 8: Two figures to illustrate the impact of object-induced spurious correlations on MNIST.

On Fashion MNIST, assume we take the data in Figure 9 as the training set. Majority of the data points of Shirt are darker than Coat. When differentiate between Shirt and Coat, a model can simply takes the bright ones as coats and the dark ones as shirts to obtain high training accuracy. However, what if the color of 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.

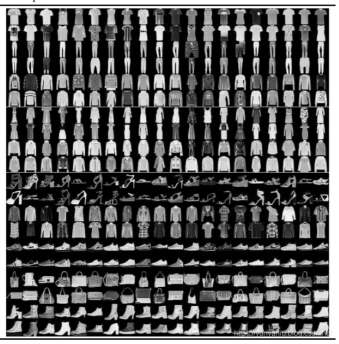
Label	Description	Examples
0	T-Shirt/Top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandals	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boots	

Figure 9: The figure to illustrate the impact of object-induced spurious correlations on Fashion MNIST.