Mixture Data for Training Cannot Ensure Out-of-distribution Generalization

Songming Zhang, Yuxiao Luo, Qizhou Wang, Haoang Chi,
Xiaofeng Chen, Bo Han, Junbin Gao, and Jinyan Li

***Abstract*—Deep neural networks often face generalization problems to handle out-of-distribution (OOD) data, and there remains a notable theoretical gap between the contributing factors and their respective impacts. Literature evidence from in-distribution data has suggested that generalization error can shrink if the size of mixture data for training increases. However, when it comes to OOD samples, this conventional understanding does not hold anymore—Increasing the size of training data does not always lead to a reduction in the test generalization error. In fact, diverse trends of the errors have been found across various shifting scenarios including those decreasing trends under a power-law pattern, initial declines followed by increases, or continuous stable patterns. Previous work has approached OOD data qualitatively, treating them merely as samples unseen during training, which are hard to explain the complicated non-monotonic trends. In this work, we quantitatively redefine OOD data as those situated outside the convex hull of mixed training data and establish novel generalization error bounds to comprehend the counterintuitive observations better. The new error bound provides a tighter upper bound for data residing within the convex hull compared to previous studies and is relatively relaxed for OOD data due to consideration of additional distributional differences. Our proof of the new risk bound agrees that the efficacy of well-trained models can be guaranteed for unseen data within the convex hull; More interestingly, but for OOD data beyond this coverage, the generalization cannot be ensured, which aligns with our observations. Furthermore, we attempted various OOD techniques (including data augmentation, pre-training, algorithm power, *etc*.) to underscore that our results not only explain insightful observations in recent OOD generalization work, such as the significance of diverse data and the sensitivity to unseen shifts of existing algorithms, but it also inspires a novel and effective data selection strategy.**

***Index Terms*—Out-of-distribution, Deep neural network, Generalization risk.**

I. Introduction

REAL-WORLD data are often sourced from diverse domains, where each source is characterized by a different distribution shift, and unknown shifts are hidden in their test distributions (see Fig. 1) [1, 2]. While deep neural networks (DNNs) demonstrate proficiency with in-domain data, they have difficulties in gaining generalization on unknown data shifts. Most out-of-distribution (OOD) generalization methods typically assume their model's capability to extrapolate across every unseen shift and have been working on algorithm improvements such as regularization [3], robust optimization [4], and adjustments in model architecture [5]. Despite these efforts, theoretical analysis of the key factors influencing unseen data and their impacts on model performance is still lacking. Previous empirical evidence such as neural scaling law [6, 7] suggests that all generalization errors follow the same decreasing trend as a power of training set size. Thus, it means that the addition of training data can effectively minimize generalization errors even for unknown distributions.

However, it appears that the scaling law conclusion may not hold true in practice, potentially leading to counterintuitive outcomes. In fact, the OOD generalization error can have a diverse range of scenarios, including decreases under a power-law pattern, initial decreases followed by increases, or remaining stable. We thus propose that *not all generalization errors in unseen target environments will decrease when training data size increases.* As the generalization error decreases, the model's accuracy improves, indicating better generalization capability on OOD data. In other words, we conjecture that simply increasing the volume of training data may not necessarily enhance generalization to OOD data. To our knowledge, no theoretical or empirical methods in the literature have addressed this phenomenon.

To formally understand such problems, this paper first presents empirical evidence for non-decreasing error trends under various experimental settings on the MNIST, CIFAR-10, PACS, and DomainNet datasets. The results are then used to illustrate that with the expansion of the training size, the generalization error decreases when the test shift is relatively minor, mirroring the performance observed under in-distribution (ID) data. Yet, when the degree of shift becomes substantial, the generalization error may not decrease monotonically. Previous works often indiscriminately categorized data unseen during training as OOD data without acknowledging the underlying causes of the non-monotonic patterns that may contain. This motivates us to revisit fundamental OOD generalization setups to elucidate such non-decreasing trends.

We propose a novel theoretical framework within the context of OOD generalization. Given a set of training environments, we first argue that OOD data can be redefined as a type of data lying outside the convex hull of source domains. Based on this new definition, we prove new bounds for OOD generalization errors. Our revised generalization bound estab-

Corresponding author: Jinyan Li.

The work was done at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. S. Zhang and X. Chen are with Department of Mathematics, Chongqing Jiaotong University, China (sm.zhang1@siat.ac.cn). Q. Wang and B. Han are with Hong Kong Baptist University. H. Chi is with National University of Defense Technology. J. Gao is with Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Australia (junbin.gao@sydney.edu.au). Y. Luo and J. Li are with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (jinyan.li@siat.ac.cn)

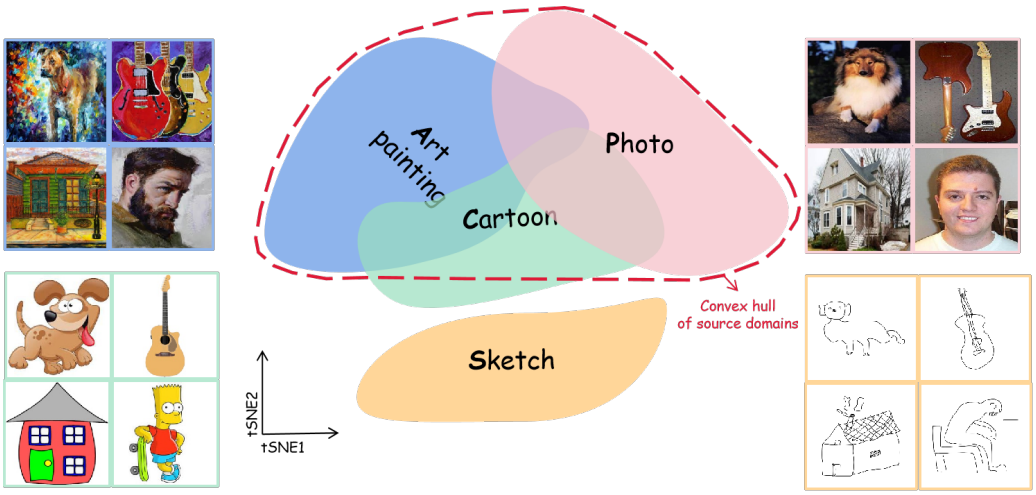


Fig. 1. A schematic diagram of a multi-domain sample in practice which consists of source and target domains. Suppose we can only have access to `Painting` and `Photo`, the model exhibits different generalization abilities at different OOD domains `Cartoon` and `Sketch` according to the distance to the mixture of training data. We draw a counterintuitive conclusion that the efficacy of well-trained models cannot be guaranteed for OOD data beyond the convex hull of training mixture, which is consistent with our experimental observations in Section II.

lishes a tighter upper bound for data within the convex hull, while being relatively relaxed for OOD data when considering additional distributional divergences. Further analysis of this new bound yields some theoretical insights into the influence of shift degree and data size on generalization error, and even on model capacity. By effectively distinguishing between ID and OOD cases in the unseen data, we demonstrate that model performance can be guaranteed for unseen yet ID data with extrapolation across training environments. However, our results collectively highlight that **being trivially trained on data mixtures cannot guarantee the OOD generalization ability of the models**, *i.e.*, the model cannot infinitely improve its OOD generalization ability by increasing training data size. Thus, achieving comparable performance on OOD data remains a formidable challenge for the model.

With our new theoretical framework in place, the ensuing question pertains to enhancing its predictive efficacy in scenarios devoid of prior target knowledge and providing coherent explanations. First, we employ widely used techniques to assess model adaptability, including data augmentation, pre-training, and algorithms. Drawing from our new theoretical findings, the efficacy of these invaluable tools can be elucidated by their ability to expand the coverage of the training mixture and its associated convex hull. Pre-training enables the model to acquire broader and more generalized representations from pre-trained datasets, while data augmentation facilitates an increase in the diversity of representations by expanding the size of training data. In contrast, hyperparameter optimization yields poor results since it solely modifies the training hyperparameters without providing more insight regarding the OOD samples. Moreover, it is important to note that existing algorithms are also sensitive to unseen test shifts.

We also note that this work distinguishes itself not just by offering theoretical understandings for error scenarios on DNNs and common techniques used for OOD generalization, but also by presenting novel insights for algorithm design. Specifically, inspired by the analysis of data diversity in our new definition, we proceed to evaluate a novel data selection algorithm that relies only on training data. By selectively choosing samples with substantial differences alone, the coverage of the training mixture can be effectively expanded, consequently broadening representations learned by the models. Remarkably, this algorithm surpasses the baseline, particularly in the case of large training size, whether chosen randomly or using reinforcement learning techniques. As our focus is on data preprocessing without the requirement of environmental labels, this allows for a smooth combination with other OOD generalization methods to improve further the model's capability to handle unseen shifts.

Overall, we have made three significant contributions in this study:

1) Contrary to the widely held "more data, better performance" paradigm, we draw a counterintuitive picture: simply increasing training data cannot ensure model performance especially when distribution shifts occur in test data. Our empirical conclusion stems from the complicated non-decreasing trends of OOD generalization errors.
2) We proposed a novel definition for OOD data and proved new error bounds for OOD generalization. Our further analysis of this new bound revealed the main factors with their influence leading to non-decreasing tendency and provided strong support for our empirical conclusions.
3) We explored and validated popular techniques such as data augmentation, pre-training, and algorithm tricks, demonstrating that our new theoretical results not only explain their effectiveness but also guide a novel data selection method for superior performance.

II. Generalization Error of Patterns Observed from OOD Scenarios

To thoroughly assess the model's generalization ability, especially its performance in the presence of distribution shifts,

Table I. Summary of network architectures used in the experiments.
We select different architectures for different tasks to better observe OOD generalization error patterns.

EXPERIMENT	NETWORK(S)	#CLASSES	IMAGE SIZE	BATCH SIZE
ROTATED CIFAR-10	WRN-10-2	2	$(3, 32, 32)$	128
BLURRED CIFAR-10	WRN-10-2	2	$(3, 32, 32)$	128
SPLIT-CIFAR10	SMALLCONV, WRN-10-2	2	$(3, 32, 32)$	128
ROTATED MNIST	SMALLCONV	10	$(1, 28, 28)$	128
PACS	WRN-16-4	4	$(3, 64, 64)$	16
DOMAINNET	WRN-16-4	40	$(3, 64, 64)$	16
CINIC-10	WRN-10-2	10	$(3, 32, 32)$	128

we examine generalization patterns from different scenarios of OOD data. OOD generalization refers to the model's capacity to perform well on those test data containing a distribution distinct from those occurring in the training data [8]. Traditionally, it has been intuitively assumed that as the training size increases, the model's generalization performance improves and the generalization error decreases accordingly, which was also verified experimentally by neural scaling law [6]. However, this assumption needs to be verified by specific experiments to ensure whether it is still applicable to OOD data.

A. Experimental Settings for OOD Evaluation

To systematically evaluate the model's ability to generalize on OOD samples, we designed a series of experiments. Our objective through these experiments is to reveal the relationship between model's OOD capability and distribution shift and to explore whether and why the model can maintain its performance under unknown shifts. The summary of the datasets, network architectures, and training details used in the experiments are presented as follows.

1) Datasets containing OOD distributions: The OOD subtasks follow the setting outlined by De Silva et al. [9] and are constructed from the images available at CIFAR-10, CINIC-10, and several datasets at DomainBed [10] such as Rotated MINST [11], PACS [12] and DomainNet [13]. Two types of settings are attempted to examine the impact of training data size on OOD data.

OOD data arising due to correlation shift. We investigate how correlation shift affects a classification task from a transformed version of the same distribution. We consider the following scenarios for this purpose.

1) **Rotated CIFAR-10:** 0° and 60°-rotated images as training distribution, and θ_1°-rotated images $(30^\circ - 150^\circ)$ as unseen target. This scenario tests how rotation changes the appearance of natural images.
2) **Blurred CIFAR-10:** 0 and 3-blurred images as training distribution and σ_1-blurred images with a range of blurring levels from 2 to 20 as the unseen target. This scenario tests how blur changes the clarity of natural images. Here, "blur" refers to adding a corresponding degree of Gaussian blur to the original images.
3) **Rotated MNIST:** 0° and 30°-rotated digits as training distribution, and θ_1°-rotated digits $(15^\circ - 60^\circ)$ as unseen target. This scenario tests how rotation changes the appearance of handwritten digits.

OOD data caused by diversity shift. We also study how diversity drifts affect a classification task using data samples from source distributions and OOD samples from a different distribution. Diversity shift is a change in diversity or variability of the data distribution between training data and OOD data [14]. We consider the following scenarios for this purpose.

1) **CINIC-10:** The construction of the dataset motivates us to consider two sub-tasks from CINIC-10: (1) Distribution from CIFAR images to ImageNet, and (2) Distribution from ImageNet images to CIFAR. This scenario tests how well a model can recognize images from another data distribution.
2) **Split-CIFAR10:** We use two of the five binary classification tasks from Split-CIFAR10 as training data and another as the unseen task. This scenario tests how well a model can distinguish between different categories of natural images.
3) **PACS:** We use two of the four-way-classification from four domains (Photo, Art Painting, Cartoon and Sketch) and images from one of the other unused domains as the unseen task. This scenario tests how well a model can generalize to different styles and depictions of images.
4) **DomainNet:** Same settings as PACS, we use a 40-way classification from 6 domains in DomainNet (`clipart`, `infograph`, `painting`, `quickdraw`, `real`, and `sketch`). This scenario tests how well a model can adapt to different domains of images with varying levels of complexity and diversity

2) Details for Training: For each random seed, we randomly select samples of different sizes n_0 and n_1 from the source distribution, making their total number as $N = n_0 + n_1$. Next, we select OOD samples of a fixed size M from the unseen target distribution. For Rotated MNIST, Rotated CIFAR-10, and Blurred CIFAR-10, the unseen distribution refers to never appearing at the rotated or blurred level. For PACS and DomainNet, images are down-sampled to $(3, 64, 64)$ during the training.

3) Neural Architectures: In our experiments, we utilize three different network architectures: (a) a small convolutional network with 0.12M parameters (*SmallConv*), (b) a wide residual network of depth 10 and a widening factor 2 (*WRN-10-2*), and (c) a larger wide residual network of depth 16 and widening factor 4 (*WRN-16-4*) [15]. SmallConv consists of 3 convolution layers with a kernel size of 3 and 80 filters, interweaved with max-pooling, ReLU, batch-norm layers, and

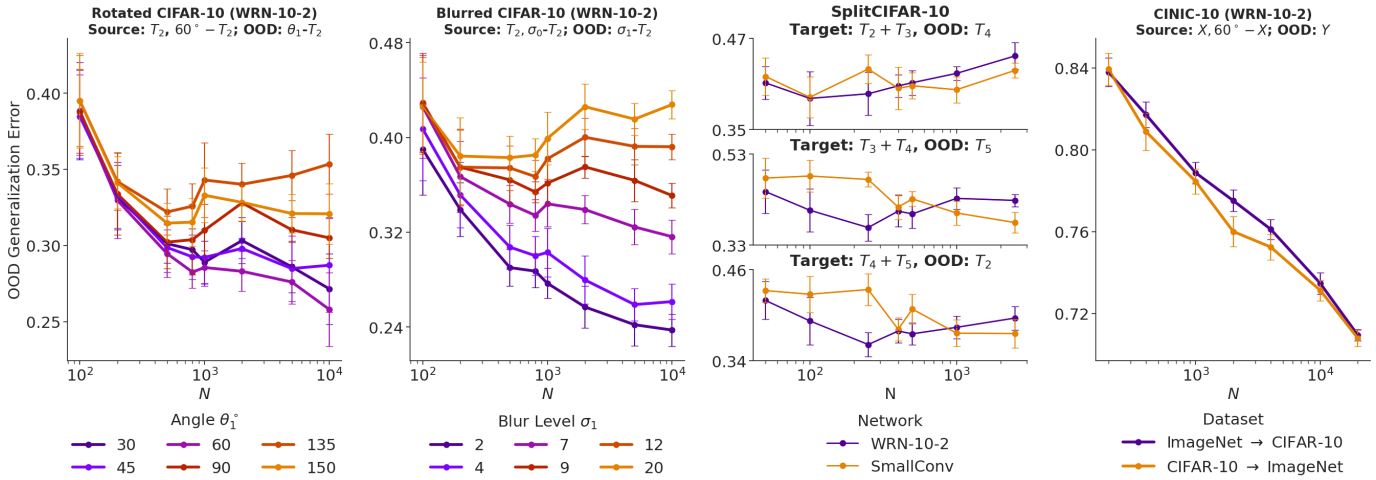


Fig. 2. The lower the OOD generalization error, the better the model is at handling unseen targets. Error bars indicate 95% confidence intervals (10 runs). **(a)** Different angles θ_1 as unseen samples obtained by rotating images in OOD sub-task T2 (`Bird` *vs.* `Cat`) in CIFAR-10, with 0° and 60° as training samples, $M = 400$. For small θ_1, increasing training data size improves the OOD generalization ability of the model. However, beyond a certain value of Δ_1, the error with large rotation has a non-monotonic trend, which means overfitting on unseen rotation. **(b)** $2-20$ level of Gaussian blur are unseen samples, and the training blur levels are at 0 and 3, $M = 400$. The model is resilient to unobserved blur, yet for extreme levels of blur, non-monotonic scenarios are evident, indicating that the model is misaligned with data due to noise. **(c)** Generalization error of two separate networks, WRN-10-2 and SmallConv, concerning a given unseen task. Our plots involve 3 different task pairs from Split-CIFAR10 and exhibit the generalization error as a function of the number of training samples. All 3 pairs demonstrated a non-decreasing trend in OOD generalization for both network models. **(d)** Generalization error of two separate datasets in CINIC-10, consisting of CIFAR-10 and ImageNet subsuet. We set one as the training environment and the other as OOD. While the purple curve shows higher error due to distribution shift, we did not observe any non-monotonic trend when testing on the unseen samples. Even when transferring between different datasets, the degree of distribution shift is still the main factor.

a fully-connected classifier layer.

Table I provides a summary of the network architectures used. All networks are trained using stochastic gradient descent with Nesterov's momentum and cosine-annealed learning rate scheduler. The training hyperparameters include a learning rate of 0.01 and a weight-decay of 10^{-5}.

B. Generalization error scenarios for deep learning benchmark datasets

Not all generalization errors decrease due to correlation shift. The shift in correlation refers to the change in the statistical correlations between the source and unseen target distribution [14]. Five binary classification sub-tasks use CIFAR-10 to explore the generalization scenario of unseen data. Our research focuses on a CIFAR-10 sub-task T_2 (`Bird` *vs.* `Cat`), where we use rotated images with 0° and 60° as training environment. Also, we use the rotated images with fixed angles from $30^\circ - 150^\circ$ as OOD samples. We also investigate the OOD effect of applying Gaussian blur with different levels to sub-task T_2 from the same distribution. Our results in Fig. 2(a)-(b) both show a monotonically decreasing trend within the generalization error for the low-level shift, *i.e.*, small rotation, and low blur. However, for a high-level shift, it is a non-monotonical function of training sample size on the target domain. Despite enlarged training data, the generalization error remains relatively high. This can be explained in terms of overfitting training data: the model learns specific patterns and noise of the training data (such as rotation or blur) but fails to capture the underlying representations of the data. This compromises the model's robustness to variations in distribution and its generalization to new data.

Not always a decreasing trend can occur when OOD samples are drawn from a different distribution. OOD data may arise due to categories evolving with new appearances over time or drifting in underlying concepts (for instance, an airplane in 2024 with a new shape or even images of drones) [16]. Split CIFAR-10 with 5 binary sub-tasks is used to study the generalization scenario of unseen data, such as `frog` *vs.* `horse` and `ship` *vs.* `truck`. We consider sub-task combinations (T_i, T_j) as the training domain and evaluate the trend of error on T_k. As the sample size grows, see Fig. 2c, the generalization error shrinks slightly or even does not show a falling trend, either in WRN-10-2 or SmallConv. Interestingly, WRN-10-2 initially outperforms SmallConv but it is overtaken by the latter as the number of samples increases.

Non-decreasing trend also occurs for OOD benchmark datasets. The different trends of generalization error motivate us to further investigate three popular datasets in the OOD works [14, 17]. First, we examine Rotated MNIST benchmark when the OOD samples are represented by θ-rotated digit images in MNIST, while 0° and 30°as training angles. We observe a decreasing trend in Fig. 3 (left), but the error lower bound keeps rising and the slope keeps getting smaller as the testing angle increases. This shows that even in real-world datasets, model generalization is also vulnerable to unknown shifts. Thus, we explore the PACS [12] and DomainNet [13] dataset from DomainBed benchmark [10]. The dataset contains subsets of different domains, two of which we selected as training domains and the other as an unseen target domain. When training samples consist of sketched and painted images, the generalization error on the clipart domain falls exponentially (*i.e.*, S/A $\rightarrow$ C in Fig. 3 (Middle)).

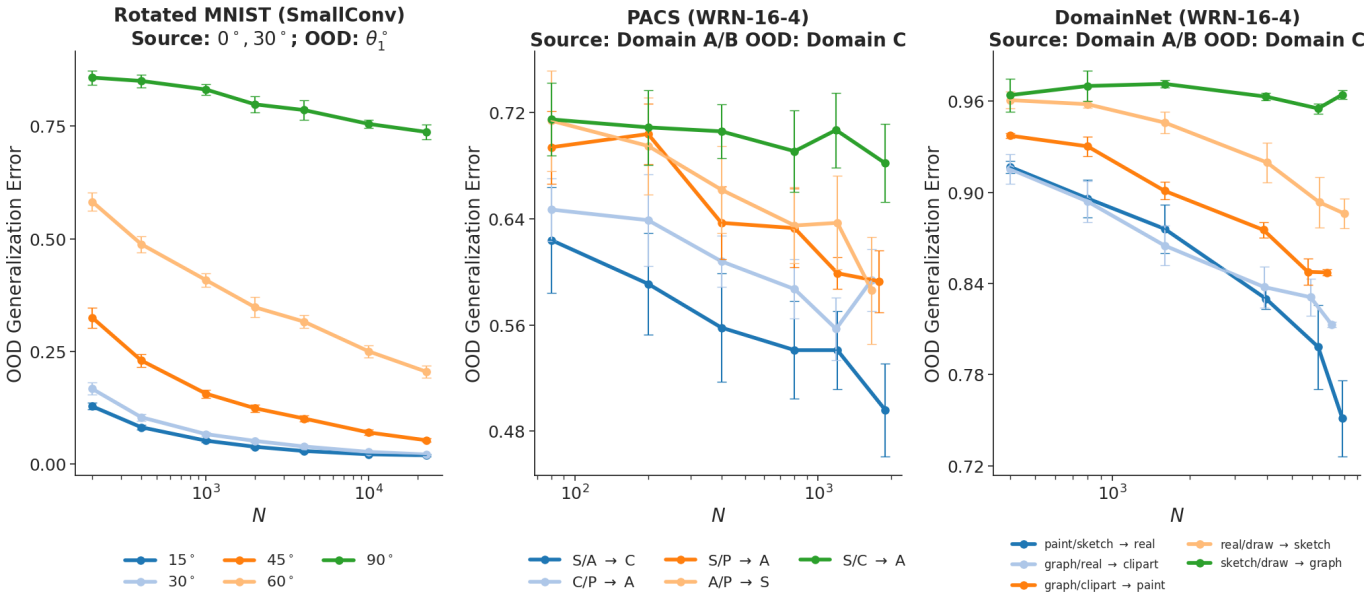


Fig. 3. Different error trends in OOD generalization error on three DomainBed benchmarks. **Left:** Rotated MNIST (10 classes, $M = 2,000$, SmallConv), **Middle:** PACS (4 classes, 4 domains {A, C, P, S}, $M = 25$, WRN-16-4), **Right:** DomainNet (40 classes, 6 domains {paint, sketch, real, graph, clipart, draw}, $M = 25$, WRN-16-4). Error bars indicate 95% confidence intervals (10 runs for Rotated MNIST and PACS, 3 runs for DomainNet). As the number of training samples increases, the various distances between distributions and how they are combined lead to different decreasing trends in OOD generalization error.

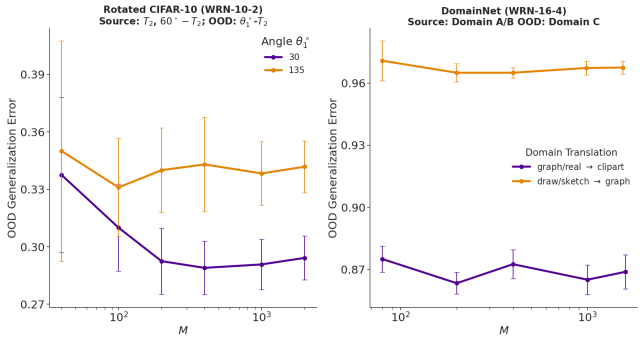


Fig. 4. From two benchmark datasets, we plot their OOD generalization error (y-axis) as a function of the OOD sample sizes per class (M) (x-axis), namely **Left:** a classification task from Rotated CIFAR-10, where the OOD rotation is $\theta_1 = 30°$ and $135°$. **Right:** a classification task from DomainNet with OOD environment of `graph` and `clipart` respectively. We calculate the OOD generalization error over 10 runs and 3 seeds for the two datasets respectively. We found a decrease at lower M across all the pairs, and the average error is stable with a decreasing variance for larger values of M. Error bars indicate 95% confidence intervals.

Moreover, an interesting observation is that the generalization error tested on the graphical images drops only slightly and remains consistently high when learned from parts of drawings and sketches. Similar trends are also observed in DomainNet, which is a comparable benchmark to PACS; See Fig. 3 (Right).

Generalization error decreases in power law despite shifts in dataset distributions. We take CINIC-10 as an example of distribution shifts between different datasets. CINIC-10 is a dataset consisting of images selected from CIFAR-10 and down-sampled from ImageNet. We train a network on one subset of CINIC-10, use another subset as OOD samples, and test on it. Fig. 2d shows that all situations exhibit a consistently decreasing trend, signifying that the OOD generalization error declines with an increase in the number of samples from the training dataset. Consequently, all models can perform well on another dataset without prior knowledge. The reason for this intriguing phenomenon is that although the datasets are separate (CIFAR-10 *vs*. ImageNet), the data distributions may not be significantly different.

No effect of OOD sample size on generalization error. Unlike the above experiments where the number of training samples is fixed, we here investigate the impact of OOD samples on generalization error with two representative training environments in OOD datasets (30° and 150° as OOD in Rotated CIFAR-10, `graph` and `clipart` as OOD in DomainNet). As depicted in Fig. 4, the generalization error declines with fewer target samples, however, as the number of OOD samples increases, the generalization error stays flat and is related to the degree of distribution shift. Thus, OOD generalization error is not associated with the number of samples in the unseen target domain, but rather with the training domains and the degree of shift in the target domain.

Discussion. Even restricted to OOD benchmark datasets, different scenarios of generalization errors can be observed. We have observed that the models can perform well on the interpolated distribution of training mixtures, that is, the error decreases as the sample size increases. DNNs are highly non-convex, which makes it difficult to find the global optima, but they can generalize well on interpolated distributions. The intuitive reason is that DNN performs well on the original distribution and as a continuous function, DNN with locally optimal values at the boundary of the original distributions may generalize equally well on the interpolated distributions. Meanwhile, if the two distributions are similar enough, a well-

trained DNN is also likely to perform well on the interpolated distributions because it has learned the common features between their original distributions. When the unseen shift occurs, there is not always a downward trend. Models that perform well on the interpolated distribution fail when the test distribution is shifted significantly. In other words, the error may not decrease as the volume of training data grows. In the next section, we revisit the OOD generalization problem and seek theoretical explanations for the non-descending phenomenon.

III. Revisit OOD generalization problem

We first review those definitions and theories related to the OOD generalization problem.

A. Formulation of the OOD generalization problem

Consider a set of $N_{\mathcal{E}}$ environments (domains) $\mathcal{E} = \{e_i\}_{i=1}^{N_{\mathcal{E}}}$. Let $\mathcal{E}_s$ and $\mathcal{E}_t$ be source and target environments collected from $\mathcal{E}$, respectively. That is, $\mathcal{E}_s \subset \mathcal{E}$ and $\mathcal{E}_t \subset \mathcal{E}$. For each source environment $e \in \mathcal{E}_s$, there exists a training dataset $(X^e, Y^e) = \{(x_i^e, y_i^e)\}_{i=1}^{N_e}$ collected from each training environment. We use x^e and y^e to denote the generic sample and the label variables with respect to the environment e, respectively. Furthermore, denote the overall training dataset as $D_s = \{(X^e, Y^e) : e \in \mathcal{E}_s\}$.

We only have access to $\mathcal{E}_s$ during training, and the target environments are unseen during training and relatively different from the training one. Moreover, we assume that there exists a ground-truth label process h satisfying $h(x^e) = y^e$. Then, in an OOD generalization, we would like to find a proper hypothesis function f that minimizes the worst empirical risk among all the training environments D_s,

$$\arg\min_{f} \sup_{e \in \mathcal{E}_s} R^e[\ell(h(x^e), f(x^e)], \tag{1}$$

where R^e denotes the "empirical" risk calculated over the loss ℓ measuring the difference between the ground truth and the function f for any sample (x^e, y^e) in the training environments. In this paper, we just consider a binary classification where the label y^e is 0 or 1.

We then specify a set of assumptions about the data-generating environmental process and consider the OOD generalization error of interest. The described multi-environment model is general enough to cover both the i.i.d. case ($\mathcal{E}$ contains a single environment) and the OOD setup ($\geq$ 2 environments are allowed) but also supports several other cases. The difference among the environments is measured by $\mathcal{H}$-divergence as well as for domain adaption case [18].

$$d_{\mathcal{H}}[e', e''] = 2 \sup_{f \in \mathcal{H}} |Pr_{x \sim e'}[\mathbf{I}(f)] - Pr_{x \sim e''}[\mathbf{I}(f)], \tag{2}$$

where $\mathcal{H} = \{f : \mathcal{X} \mapsto \{0, 1\}\}$ is a hypothesis class on $\mathcal{X}$ and $\mathbf{I}(f) = \{x : f(x) = 1\}$, i.e., all the inputs $x \in \mathcal{X}$ that are classified as class 1 by the hypothesis f.

In the context of OOD generalization, the test distribution is inaccessible, necessitating certain assumptions about the test environment to enable generalization.

We address this issue in the results below and introduce generalization guarantees for a specific test environment with data mixture of training distributions. Let the training environments $\mathcal{E}_s$ contain N_S training environments, denoted as $\mathcal{E}_s = \{e_s^i\}_{i=1}^{N_S}$. The convex hull $Con(\mathcal{E}_s)$ of $\mathcal{E}_s$ is defined as the set of pooled environments given by

$$Con(\mathcal{E}_s) = \{\hat{e} \mid \hat{e} = \sum_{i=1}^{N_S} \alpha_i e_s^i, \alpha_i \in \Delta_{|N_S|-1}\}, \tag{3}$$

where $\Delta_{|N_S|-1}$ is the $(|N_S|-1)$-th dimensional simplex. The following lemma shows that for any pair of environments such that $e', e'' \in Con(\mathcal{E}_s)$, the $\mathcal{H}$-divergence between e' and e'' is upper-bounded by the largest $\mathcal{H}$-divergence measured between the elements of S.

Lemma III.1 (Paraphrase from [19]). *Suppose $d_{\mathcal{H}}\left[e^i, e^j\right] \leq \epsilon, \forall i, j \in [N_S]$, then the following inequality holds for the $\mathcal{H}$-divergence between any pair of environments $e', e'' \in Con(\mathcal{E}_s)$:*

$$d_{\mathcal{H}}\left[e', e''\right] \leq \epsilon. \tag{4}$$

It is suggestive that the $\mathcal{H}$-divergence between any two environments in $Con(\mathcal{E}_s)$ (*i.e.*, the maximum pairwise $\mathcal{H}$-divergence) can be used to measure the difference from the target environment and can affect OOD generalization ability of the model.

One generally refers to all unseen data as OOD data in literature work, provided such data is not available during training [8, 20]. The classic OOD data is defined as follows

Definition III.1 (Out-of-distribution data (General)). *Let $\mathcal{E}_s$ and $\mathcal{E}_t$ be the sets of source and target environments, respectively, and let $\mathcal{Y} = [0, 1]$. Suppose we have a set of n source environments, $\mathcal{E}_s = \{e_s^1, e_s^2, \ldots, e_s^{N_S}\}$. Let $e_t \in \mathcal{E}$ be an unseen target environment, such that for any data in e_t, the following conditions hold*

$$e_t \notin \{e_s^1, e_s^2, \ldots, e_s^{N_S}\}, \tag{5}$$

*then the data in e_t is said as **out-of-distribution data**.*

B. Redefinition for OOD data

It has been typically assumed in the literature that any test environment that does not appear in the training environments is considered OOD data. However, our findings suggest that OOD data exhibit heterogeneous generalization scenarios despite never having been encountered during training. We now use Lemma III.1 to redefine whether the unseen data is OOD. Formally,

Definition III.2 (Out-of-distribution data (Refined)). *Let $\mathcal{E}_s$ and $\mathcal{E}_t$ be the sets of source and target environments, respectively, and $\mathcal{Y}$ be the output space. Suppose that we have a set of n source environments, $\mathcal{E}_s = \{e_s^1, e_s^2, \ldots, e_s^{N_S}\}$ and that there exists a real number $\epsilon > 0$ such that $d_{\widetilde{\mathcal{H}}}[e^i, e^j] \leq \epsilon$, $\forall e^i, e^j \in Con(\mathcal{E}_s)$. Let $e_t \in \mathcal{E}$ be an unseen target environment such that, for any data in e_t, the following conditions hold:*

$$d_{\widetilde{\mathcal{H}}}(e_t, \hat{e}) > \epsilon, \quad \forall \hat{e} \in Con(\mathcal{E}_s), \tag{6}$$

then the data in e_t is said as ***out-of-distribution data****, where $\hat{e} := \sum_{i=1}^{N_S} \alpha_i e_s^i$, where $\hat{e} \in Con(\mathcal{E}_s)$, $\sum_{i=1}^{N_S} \alpha_i = 1$ is the convex combination of training environments, and $\alpha_i \geq 0$ for all i.*

Remark III.1. *We discuss ϵ in this definition, noting that the mixture of training environments, $Con(\mathcal{E}_s)$, can affect generalization to the target distribution. Threshold estimates are typically derived by assessing the model's performance on a mixture of training data, which reflects the model's ability on a specific distribution. However, unknown shifts can influence these estimates and are not directly measurable in practice. It is important to indicate that the data mixture in training may not cover all samples in the unseen target environment. Consequently, a model trained solely on this mixture may not be insufficient for generalizing to the target environment.*

Remark III.2 (An intuitive explanation of OOD data definition)**.** *The conventional intuitive way to understand OOD data is to treat it as inaccessible and not included in the training distribution. However, the various scenarios of generalization errors prompt us to introduce "convex hull" empirically, which helps define OOD data and the expected generalization. A classifier works via decision boundaries, which learns from the set of all mixtures obtained from given training distributions,* i.e.*, convex hull. The distance between the target distribution and the decision boundary is a determinant of the classifier's performance. In the process of OOD generalization, unlike domain adaptation settings, no data from the test distribution can be observed. Furthermore, not all test samples are located outside the convex hull constructed by the training set, as defined above. For example, in Fig. 1, the model learning from* `Painting` *and* `Photo` *predicts better in* `Cartoon` *than* `Sketch` *due to* `Cartoon` *is closer to the training mixture.*

Next, we present the error bounds in the following theorem.

Theorem III.1 (Upper-bounding the risk on unseen data)**.** *Let $\mathcal{E}_s$ be the set of training environments and let $\mathcal{Y} = [0, 1]$ For any unseen environment $e_t \in \mathcal{E}_t$ and any hypothesis $f \in \mathcal{H}$, the risk $R_t(f)$ can be bounded in the following ways:*

(i) If $e_t \in Con(\mathcal{E}_s)$, data in e_t is considered as ID, then

$$R_t(f) \leq \sum_{i=1}^{N_S} \alpha_i R_s^i(f) + 2\epsilon + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}, \tag{7}$$

(ii) If $e_t \notin Con(\mathcal{E}_s)$, data in e_t is considered as OOD, then

$$R_t(f) \leq \sum_{i=1}^{N_S} \alpha_i R_s^i(f) + \delta + \epsilon + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}. \tag{8}$$

where ϵ is the highest pairwise $\widetilde{\mathcal{H}}$-divergence measured between pairs of environments within $Con(\mathcal{E}_s)$ under $\widetilde{\mathcal{H}} = \{sign(|f(x) - f'(x)| - t) \mid f, f' \in \mathcal{H}, 0 \leq t \leq 1\}$, $\delta := \min_{\alpha_i} d_{\widetilde{\mathcal{H}}}[e_t, \sum_{i=1}^{N_S} \alpha_i e_s^i]$ with minimizer α_i be the distance of e_t from convex hull $Con(\mathcal{E}_s)$, $\hat{e} := \sum_{i=1}^{N_S} \alpha_i e_s^i$ is the "projection" of e_t onto convex hull $Con(\mathcal{E}_s)$ with $\alpha_i \geq 0$ for all i, $h_{s'}(x) = \sum_{i=1}^{N_S} \alpha_i h_{e_s^i}(x)$ is the labeling function for any $x \in Supp(\hat{e})$ derived from $Con(\mathcal{E}_s)$ with weights α_i, and h_t is the ground-truth labeling function for e_t.

Proof. Let the source environment and target environments be $\mathcal{E}_s$ and $\mathcal{E}_t$, respectively. The risk $R_t(h)$ can be bounded by Zhao et al. [21] for single-source and single-target domain adaptation cases as follows:

$$R_t(f) \leq R_s(f) + d_{\widetilde{\mathcal{H}}}[e_s, e_t] + \min\{\mathbb{E}_{e_s}\|h_s - h_t\|, \mathbb{E}_{e_t}\|h_t - h_s\|\}. \tag{9}$$

where $\widetilde{\mathcal{H}} = \{sign(|f(x) - f'(x)| - t) \mid f, f' \in \mathcal{H}, 0 \leq t \leq 1\}$. To design a generalized constraint for the risk of any unseen domain based on quantities associated with the distribution seen during training, we need to start by rewriting Equation (9) and considering e_t and its "projection" onto the convex hull of $\hat{e} \in Con(\mathcal{E}_s) = \{\hat{e_s} | \hat{e_s} = \sum_{i=1}^{N_S} \alpha_i e_s^i, \sum_{i=1}^{N_S} \alpha_i = 1, \alpha_i \geq 0\}$.

For that, we introduce the labeling function $h_{s'}(x) = \sum_{i=1}^{N_S} \alpha_i h_{e_s^i}(x)$ which is an ensemble of the respective labeling functions and each weighted by the responding mixture coefficients from $Con(\mathcal{E}_s)$. $R_t(f)$ can thus be bounded as

$$R_t(f) \leq R_{\hat{e}}(f) + d_{\widetilde{\mathcal{H}}}[\hat{e}, e_t] + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}. \tag{10}$$

Similarly to Lemma III.1 for the case where $\mathcal{H} = \widetilde{\mathcal{H}}$, $e' = e_t$ and $e'' = \hat{e}$, it follows that

$$R_t(f) \leq \sum_{i=1}^{N_S} \alpha_i R_s^i(f) + \sum_{i=1}^{N_S} \alpha_i d_{\widetilde{\mathcal{H}}}[e_s^i, e_t] + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}. \tag{11}$$

Using the triangle inequality for the $\widetilde{\mathcal{H}}$-divergence along with Lemma III.1, we can bound the $\widetilde{\mathcal{H}}$-divergence between e_t and any source environment e_s^i, $d_{\widetilde{\mathcal{H}}}[e_s^i, e_t]$, according to our new Definition III.2.

If $e_t \in Con(\mathcal{E}_s)$, the data in an environment e_t is defined as in-distribution data, which means even e_t has never been seen at training, it is still located in the convex hull of training mixture. According to Lemma III.1 for case I where $e' = e_t$ and $e'' = \hat{e}$, and case II where $e' = e_s^i$ and $e'' = \hat{e}$, the following inequality holds for the $\mathcal{H}$-divergence between any pair of environments $e', e'' \in Con(\mathcal{E}_s)$.

$$\begin{aligned} d_{\widetilde{\mathcal{H}}}[e_s^i, e_t] &\leq d_{\widetilde{\mathcal{H}}}[e_s^i, \hat{e}] + d_{\widetilde{\mathcal{H}}}[\hat{e}, e_t] \\ &\leq \epsilon + \epsilon = 2\epsilon. \end{aligned} \tag{12}$$

And if $e_t \notin Con(\mathcal{E}_s)$, the data in an environment e_t is defined as out-of-distribution data, and we have

$$\begin{aligned} d_{\widetilde{\mathcal{H}}}[e_s^i, e_t] &\leq d_{\widetilde{\mathcal{H}}}[e_s^i, \hat{e}] + d_{\widetilde{\mathcal{H}}}[\hat{e}, e_t] \\ &\leq \epsilon + \delta, \end{aligned} \tag{13}$$

where $\delta := \min_{\alpha_i} d_{\widetilde{\mathcal{H}}}[e_t, \sum_{i=1}^{N_S} \alpha_i e_s^i]$ with minimizer α_i be the distance of e_t from convex hull $Con(\mathcal{E}_s)$. Thus, we get an upper-bounded $d_{\widetilde{\mathcal{H}}}[e_s^i, e_t]$ based on our new definition.

Finally we rewrite the bound on $R_t(f)$: if $e_t \in Con(\mathcal{E}_s)$, the data in an environment e_t is defined as in-distribution data

$$R_t(f) \leq \sum_{i=1}^{N_S} \alpha_i R_s^i(f) + 2\epsilon + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}. \tag{14}$$

And if $e_t \notin Con(\mathcal{E}_s)$, the data in an environment e_t is defined as out-of-distribution data

$$R_t(f) \leq \sum_{i=1}^{N_S} \alpha_i R_s^i(f) + \delta + \epsilon + \min\{\mathbb{E}_{\hat{e}}\|h_{s'} - h_t\|, \mathbb{E}_{e_t}\|h_t - h_{s'}\|\}. \tag{15}$$

□

Remark III.3 (Intuitive interpretation of Theorem III.1)**.** *Compared with the upper bound proposed by Albuquerque et al. [19], our theorem derives different upper bounds for unseen data in different cases based on the new definition. For unseen data within the convex hull of training mixture, the upper bound is tighter because e_t is assumed to be close enough to the training mixture. The upper bound is more relaxed for OOD data in e_t because additional differences between the target environment and the source environment need to be taken into account (denoted by δ). Such boundaries explain why data mixture in training cannot guarantee models' OOD capability. The intuitive interpretation of this theorem can also be verified in Fig. 1. Different risk bounds on unseen data mean that in some cases, the performance of the model may be affected by different bounds even if a large amount of training data is provided, especially if the multi-domain data collected in the real world has different data quality, task difficulty, and distribution gap. This all means that even with more training samples, the OOD generalization ability of the model may be limited. In other words, the model will not improve indefinitely even if more training data is added.*

Remark III.4 (The importance of diverse data)**.** *We further highlight that the introduced results also provide insights into the value of obtaining diverse datasets for generalization to OOD samples in practice. More diverse the environments where the dataset occurs during training, it is more likely that the unseen distribution falls within the convex hull of the training environments. That is, even though the dataset has never been seen before, it may still belong to ID data. Specifically, the diversity of training samples can help the model learn more robust and general feature representations that are critical for identifying and processing new, unseen data distributions. Thus, as well as the dataset size, the diversity of training samples is also crucial for better generalization.*

Remark III.5 (Widely used OOD techniques)**.** *We next discuss widely used OOD techniques based on the introduced framework in order to demonstrate the verification of our theory and interesting observations in related algorithms. Such techniques are crucial for enhancing the robustness and OOD generalization of DNNs, including data augmentation, pre-training, and hyperparameter optimization. We then define a novel selection algorithm that relies only on training data. While the risk of utilizing the training mixture can be minimized, it is able to measure the value of data and adapt the learning scope dynamically. It is worth noting that our empirical result is that the proposed algorithm can achieve success even when no information is known about the target environment.*

IV. Can we break OOD limitation to improve model's capability?

Can we make OOD samples as generalizable as possible, so that we can improve the generalization ability of the models and we can provide reasonable explanations when we do not have any prior knowledge about the target distribution?

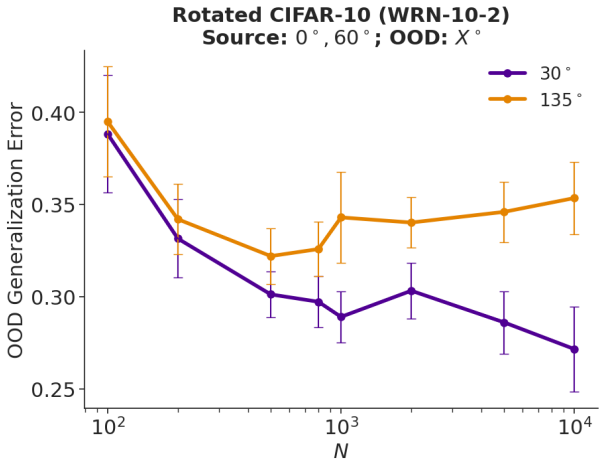


Fig. 5. For 0° and 60° as source samples, and 135° and 30° as OOD samples in Rotated CIFAR-10 sub-task T_2 respectively, we investigate the effect of **hyper-parameter tuning**. We record the best set of hyper-parameters with a validation set and test it on an unseen target. It can still be observed that the same error trend in our previous results since manipulating the training set is irrelevant for the test set, and the distribution distance is the main influencing factor.

A. Observation for widely used OOD methods

When given unknown samples that are never seen during training, the number of available options to alleviate the degradation caused by OOD samples is limited. Thus we need to use some popular techniques to process unseen data to make them as "close" as possible to the data mixture in training, such as hyperparameter optimization, data augmentation, pre-training, and other DomainBed algorithms.

Effect of hyperparameter optimization. The first technique we aim to question is whether the best-performing model trained on the training set can effectively reduce the generalization error on an unseen distribution. Similarly as by Kumar et al. [22], we employ an easy two-step strategy of linear probing and then full fine-tuning (50-50 epochs). It can be observed in Fig. 5 that relying on hyperparameter optimization alone is not sufficient to improve model performance in the case of handling OOD samples. This implies that the optimal performance of a model on training or validation data does not guarantee its success on unseen samples. The reason for this is that it adjusts primarily for the nature of training data, but does not offer enough insight regarding the OOD samples.

Effect of data augmentation. To evaluate whether augmentation works, we use two different rotations as unseen tasks for WRN-10-2 on Rotated CIFAR-10, *i.e.*, 135° and 30° as OOD tasks. The results in Fig. 6 (medium color) show that the effectiveness of data augmentation increases as the amount of data increases. Initially, for a small dataset, augmentation may have a negative effect. However, as the size of training data grows, augmentation helps break the bias (digit *vs*. rotation) in the training data. Augmentation overcomes the overfitting phenomenon for semantic noise, especially on 135° as an OOD task. Its effectiveness stems from the ability to introduce diversity into the training data patterns. Data augmentation helps create augmented samples that better represent real-world variations and challenges. This enables the model to

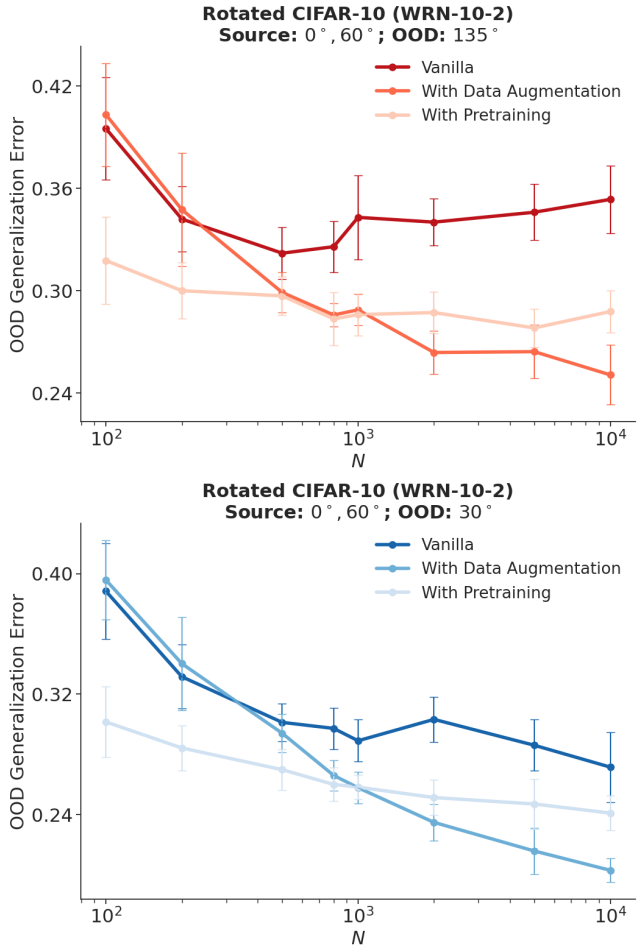


Fig. 6. For $0°$ and $60°$ as training samples, and $135°$(Upper) and $30°$(Bottom) as OOD samples in Rotated CIFAR-10 sub-task T_2 respectively, we train a WRN-10-2 model with target samples $M = 400$, under the following settings: (1) *Vanilla*, that is, without any popular techniques (darkest color), (2) *Data augmentation* with random copping, flips and padding (medium color), and (3) *Pre-training* followed by fine-tuning (lightest color). WRN-10-2 is pre-trained on ImageNet images from CINIC-10 (100 epochs and 0.01 learning rate followed De Silva et al. [9]). For smaller N, augmentation worsens the OOD generalization error but shows improvement with increasing samples, not disturbed by rotation. After pre-training, WRN-10-2 initially has a dramatic drop in error but still rises for unseen samples, especially on $135°$. Error bars indicate 95% confidence intervals (10 runs).

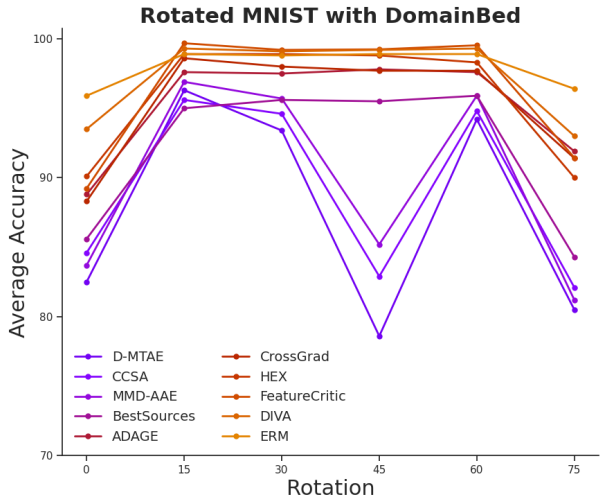


Fig. 7. Five of the six domains ($\{0°, 15°, 30°, 45°, 60°, 75°\}$) in Rotated MNIST are used as training sets and the rest as an unseen task. We validate the performance data of 10 algorithms in DomainBed on unseen rotations. The higher the OOD generalization accuracy, the better the algorithm in DomainBed is at handling unseen targets. It can be observed that all algorithms perform poorly for $0°$ and $75°$, demonstrating sensitivity to unseen shifts.

learn more robust and generalizable features, improving its performance during tests on OOD samples.

Effect of pre-training. We repeat the same target tasks for pre-training on Rotated CIFAR-10. As shown in Fig. 6 (lightest color), even with a limited sample size, pre-training demonstrated a more robust improvement. And as the sample size increases, the improvement of pre-training becomes more minor, but still better than the baseline. We can conclude that pre-training is a useful tool for improving the ability to unseen distributions. Pre-training on large and diverse datasets can enhance the model's ability to learn a broad and generalized set of representations, which can subsequently improve its performance on OOD samples. However, its effectiveness depends on how well they can transfer the learned representations to target tasks. The quality of the representations must have been tested carefully based on target tasks.

The sensitivity of distribution distance in DomainBed algorithms. We can see the performance of different algorithms on the rotated MNIST dataset in Fig. 7. Most algorithms (like ERM and DIVA [23]) perform well at rotations of $15° - 60°$, but mediocre at $0°$ and $75°$. This trend verifies that different algorithms are equally sensitive to changes in data distribution. That is, they perform well in the environment within the data mixture but perform poorly in the environment with distant distribution. Besides, the performance of some algorithms fluctuates greatly from different angles. For example, MMD-AAE [24] and BestSources [25] perform well at lower angles, but their performance degrades more after $45°$. This may be due to the different feature extraction processes of the algorithms, resulting in generalization failure at $45°$. The performance of existing algorithms is sensitive to the distribution distance, which can help us understand the generalization ability of different algorithms in OOD settings.

B. Selection of the training samples

Recent work suggests that a careful selection of the most valuable samples can prevent models from focusing too much on the noise in training, thereby having a potential reduction of the overfitting risk [26, 27]. However, data selection for the OOD generalization problem meets a train-test mismatch challenge, since it is impossible to anticipate the distribution of target data that the model will encounter in the future.

1) Diversity learning for OOD generalization: Diversity is a desirable dimension in OOD generalization problems. We can explore dynamically adjusting weight and diversity centered on the training data, so that information from the source can be *"partially adapted"* to the target. For example, style diversity in the dog data, since each style (`picture`, `sketch`, `painting`, *etc.*) is different. Our empirical result suggests that diversity ensures the expansion of the convex hull

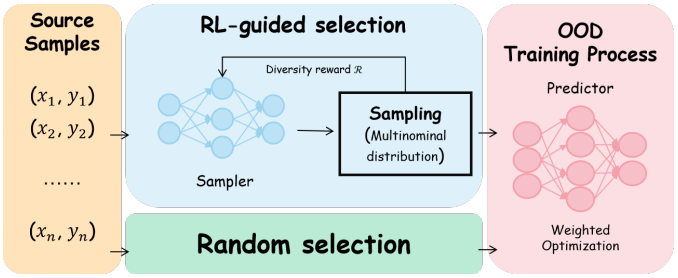


Fig. 8. Block diagram of Selecting the training samples A set of training domain samples serves as training input. (1) *RL-guided selection:* Learns from training samples (with shared parameters between batches), updates with diversity-related rewards, and returns selection vectors (corresponding to a multinomial distribution). (2) *Random selection:* Randomly output select vectors to a batch of training samples. The OOD predictor is trained only on training samples with selection vectors, using gradient descent optimization.

and contains more unseen samples as ID. To increase training diversity, we can use random weighting, since it can help to increase the amount of domain-agnostic information available and promote the robustness of neural networks. Although random selection can capture some of the information in training domains, it depends on the distance between domains and cannot correspond to each training domain sample one-to-one, and therefore cannot be updated using traditional stochastic gradient descent. Fortunately, well-established solutions from the field of reinforcement learning (RL) are readily available to update the selection sampler and weight each sample individually to the selection criteria [27, 28].

2) Framework Details: We demonstrate that our selection performance in two OOD tasks: (1) $0°$ and $60°$ as training environments and $135°$ as OOD samples in Rotated CIFAR-10; (2) `clipart` and `sketch` as training environments and `painting` as OOD samples in PACS. We formalize the components of the sample selection in OOD training set optimization as shown in Fig. 8, including a training dataset D_s with size n, a predictor f for OOD generalization and an encoder f_{en} as a feature extractor. We first shuffle and randomly partition D_s into mini-batches, and then perform the data selection process. (i) As for *random selection*, we provide a random vector for each mini-batch selection. (ii) As for *RL-guided selection*, we adopt REINFORCE [29] algorithm to train a selector $\mathcal{F}$ for the optimization of OOD generalization. Our goal is to learn an optimal policy π to maximize the diversity of selected subsets from each mini-batch.

The process of RL-guided Selection is: *First,* the encoder f_{en} transforms a batch of data B_t into its representation vector $v_t(v_t = f_{en}(B_t))$ at each step t. *Secondly,* the policy π outputs the batch of state s_t, so that each v_t is associated with a probability of diversity representing how likely it is going to be selected. The selected subset $\hat{B}_t$ is then obtained by ranking their probability. *Thirdly,* the selector $\mathcal{F}$ as well as encoder f_{en} are finetuned by the selected subset $\hat{B}_t$. The $\mathcal{F}$ is updated with REINFORCE algorithm and the reward function $\mathcal{R}$. No target knowledge is required, and the reward is only measured via max divergence of source samples. We define the divergence $Diver(B)$ as the total of all distances between

Algorithm 1: RL-guided data selection

Input: Training set D_s, Predictor f (including encoder f_{en}), Epoch M, Reward function $\mathcal{R}$, learning rate α, Discount factor γ
Output: selected set, fine-tuned predictor f, policy π, data selector $\mathcal{F}$

Initialize selection policy π and data selector $\mathcal{F}$;
foreach *episode* m **do**
  Shuffle and randomly partition D_s into mini-batched with same size N: $D_s = \{B_t\}_{t=1}^{T} = \{B_1, B_2, \ldots, B_T\}$;
  Initialize an empty list for episode history Γ;
  foreach $B_t \in D_s$ **do**
    $v_t = f_{en}(B_t)$;
    Obtain state s_t;
    Obtain action a_t by sampling based on $\pi(s_t)$;
    Obtain the selected set $\hat{B}_t$ by ranking a_t;
    Fine-tune predictor f with $\hat{B}_t$;
    Calculate reward $r_t = \mathcal{R}(\hat{B}_t, \mathcal{F})$ with Equation (16);
    Store (s_t, a_t, r_t) to episode history Γ;
  end
  foreach $(s_t, a_t, r_t) \in \Gamma$ **do**
    Update policy weight and selector weights with REINFORCE algorithm with $\mathcal{R}$, α, and γ.
  end
  Clear episode history Γ;
end
return f, π, and $\mathcal{F}$

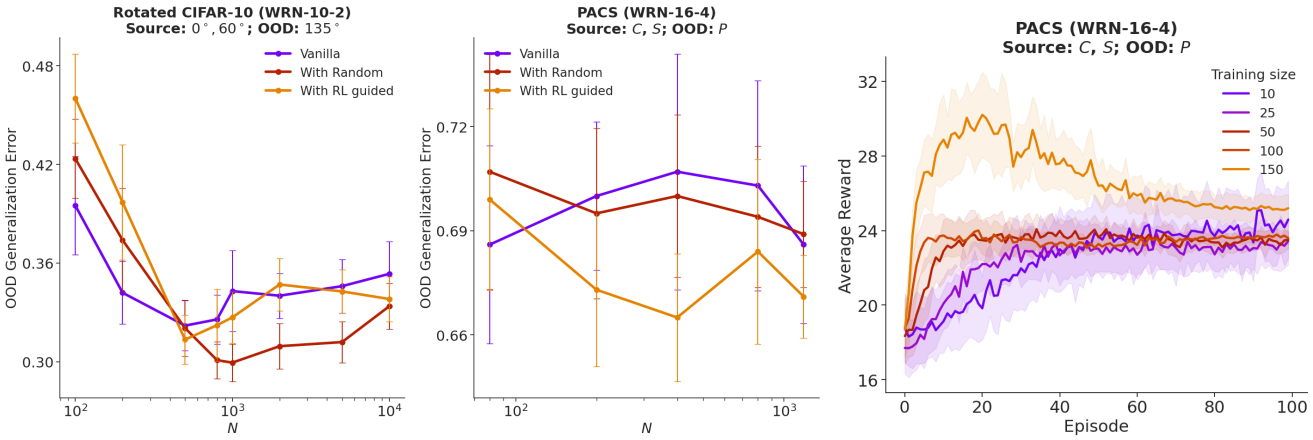
any pairs (v_i, v_j) within the batch B:

$$Diver(B) = \sum d(v_i, v_j), (v_i, v_j) \in B \tag{16}$$

where $d(\cdot, \cdot)$ is the distance function. Maximizing the dispersion of samples can effectively increase the coverage of learned representations of well-trained models, leading to a more diverse content for the target environment. See Algorithm 1 for more details. The expansion of convex hull allows models to learn more diverse types of features and patterns.

3) Results and discussion: We first provide the training reward on PACS in Fig. 9b. We can observe that the proposed algorithm gradually reaches a convergence point at approximately 40 episodes, which exhibits a faster convergence speed. If the training size is large (*e.g.*, 150), the reward first increases and then decreases until it flattens out. Additionally, the variance of average reward changes during the increase in training size. It can be concluded that by designing diversity as the reward function, the neural network can learn more diverse examples Moreover, in the same tasks presented in Section II-B, no noticeable trends were indicating a decrease in performance. However, a significant reduction in error can be observed in Fig. 9a following the selection process.

Weighted objectives were once believed to be ineffective for over-parameterized models, including DNN [30] due to zero-valued optimization results (*i.e.*, the model fits the training dataset perfectly, resulting in zero loss). However, our experiments demonstrate the opposite: our method outperforms the vanilla in the case of large training samples, whether random or elaborated weighted. This is perhaps because cross-entropy loss is hard to reach 0 due to multi-distribution distance. Simultaneously, the selection of samples exhibiting



(a) Generalization error on Rotated CIFAR-10 and PACS (b) Training reward on PACS

Fig. 9. The selecting results for OOD training. **(a)** Generalization error on the target distribution for Rotated CIFAR-10 (left) and PACS (right) using weighted subset. Here, we present three settings on OOD tasks: **Vanilla**, **With Random**, and **With RL-guided weighted objective**. All settings only get information from the training environment. For a small number of samples, the error of the latter two is high, but as the sample size grows, they can effectively lower the OOD generalization error. The effect of OOD data arising due to intra-class nuisances (Rotated CIFAR-10) is more sensitive to random weights since there is no semantic-level shift and more correlation-level. Unlike in CIFAR-10 tasks, we observe that in PACS, OOD generalization error falls significantly due to semantic-level shift. In other words, if we use a weighted training subset, then we always obtain some benefit on OOD samples, whether random or guided. **(b)** Training reward on PACS of RL-guided selection method. Reward on all training sizes converge to the optimal reward associated with PACS dataset. Error bars indicate 95% confidence intervals (10 runs).

the largest dissimilarities between distributions has the potential to broaden the coverage of the convex hull in the training domains. This expansion ultimately leads to improved performance outcomes.

By utilizing weighted objectives, we effectively prioritize and emphasize samples that significantly contribute to the overall learning process. This approach allows the model to focus on the most informative and challenging parts of the training mixture, thereby enhancing its capacity to generalize to new, unseen shifts. Furthermore, our selection method operates independently of the necessity for environment labels (*i.e.*, identification of sample domains), making it more flexible and applicable in a broader range of scenarios. It can be seamlessly integrated with other OOD generalization techniques, such as domain generalization or data augmentation, to further enhance the model's ability to generalize to new environments.

V. Related Work

OOD generalization. Theoretical achievements in machine learning has been made all under the assumption of independent and identical distributions (i.i.d. assumption). However, in many real-world scenarios, it is difficult for the i.i.d. assumption to be satisfied, especially in the areas such as medical care [31]. Consequently, the ability to generalize under distribution shift gains more importance. Early OOD studies mainly follow the distribution alignment by learning domain invariant representations via kernel methods [32], or invariant risk minimization [3], or disentangle learning [33]. Research on generalizing to OOD data has been extensively investigated in previous literature, mainly focusing on data augmentation [34] or style transfer [35]. Increasing data quantity and diversity from various domains enhances the model's capability to handle unseen or novel data. Zhang et al. [36] introduced Mixup, which generates new training examples by linearly interpolating between pairs of original samples. Zhou et al. [37] further extended this idea by Mixstyle, a method that leverages domain knowledge to generate augmented samples. Other OOD methods on the level also employ cross-gradient training [38] and Fourier transform [39].

Different from all the methods mentioned above, our work starts with the definition of OOD data and its empirical phenomenon and revisits the generalization problem from a theoretical perspective. It is worth noting that the difference between us and De Silva et al. [9] is that the latter focuses on the effects of adding a few OOD data to the training data, while we focus on the effects of generalizing to unseen target tasks.

Data selection. Data selection is a critical component of the neural network learning process, with various important pieces of work [40]. Careful selection of relevant and representative data is to guarantee that the data used for training accurately captures the patterns and relationships that the network is supposed to learn. Several recent studies have explored different metrics for quantifying individual differences between data points, such as EL2N scores [27] and forgetting scores [41]. There are also influential works on data selection that contribute to OOD and large pre-trained models. Zhu et al. [42] introduced a framework of cross-table pre-training of tabular transformers on datasets from various domains. Shao et al. [43] leveraged manually created examples to guide large language models in generating more effective instances automatically and select effective ones to promote better inference.

While these methods can improve the training data quality, our method leverages existing data with reinforcement learning in source domains for modeling to maximize diversity. Discussions concerning the convex hull and OOD data selection have also been prevalent in literature [14, 44]. For instance, Krueger

et al. [45] aim to optimize the worst-case performance of the convex hull of training mixture. Our work distinguishes itself by focusing on error scenarios resulting from distribution shifts on DNN and highlighting the importance of data diversity. Additionally, we provide novel insights for algorithm design.

VI. Conclusion

This work examined the phenomenon of non-decreasing generalization error when the models are trained on data mixture of source environments and the evaluation is conducted on unseen target samples. Through empirical analysis on benchmark datasets with DNN, we introduced a novel theorem framework within the context of OOD generalization to explain the non-decreasing trends. Furthermore, we demonstrated the effectiveness of the proposed theoretical framework in the interpretation of the existing methods by evaluating existing techniques such as data augmentation and pre-training. We also employ a novel data selection algorithm only that is sufficient to deliver superior performance over the baseline methods.

Acknowledgment

This work is supported by .

References

[1] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.

[2] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is hard: a closer look at subpopulation shift," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[3] Y. Chen, K. Zhou, Y. Bian, B. Xie, B. Wu, Y. Zhang, M. KAILI, H. Yang, P. Zhao, B. Han, and J. Cheng, "Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization," in *International Conference on Learning Representations*, 2023.

[4] W. Azizian, F. Iutzeler, and J. Malick, "Exact generalization guarantees for (regularized) wasserstein distributionally robust models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu, "Sparse mixture-of-experts are domain generalizable learners," in *International Conference on Learning Representations*, 2023.

[6] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[7] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818–2829.

[8] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] A. De Silva, R. Ramesh, C. Priebe, P. Chaudhari, and J. T. Vogelstein, "The value of out-of-distribution data," in *International Conference on Machine Learning*. PMLR, 2023, pp. 7366–7389.

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

[11] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2551–2559.

[12] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[13] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1406–1415.

[14] N. Ye, K. Li, H. Bai, R. Yu, L. Hong, F. Zhou, Z. Li, and J. Zhu, "Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7947–7958.

[15] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference 2016*, 2016.

[16] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.

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

[18] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, pp. 151–175, 2010.

[19] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, "Generalizing to unseen domains via distribution matching," *arXiv preprint arXiv:1911.00804*, 2019.

[20] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[21] H. Zhao, R. T. D. Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 7523–7532.

[22] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features

and underperform out-of-distribution," in *International Conference on Learning Representations*, 2022.

[23] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "Diva: Domain invariant variational autoencoders," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 322–348.

[24] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[25] M. Mancini, S. R. Bulò, B. Caputo, and E. Ricci, "Best sources forward: Domain generalization through source-specific nets," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1353–1357.

[26] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang, "Domain adaptive transfer learning with specialist models," *arXiv preprint arXiv:1811.07056*, 2018.

[27] J. Yoon, S. Arik, and T. Pfister, "Data valuation using reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 842–10 851.

[28] Y. Yu, S. Khadivi, and J. Xu, "Can data diversity enhance learning generalization?" in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 4933–4945.

[29] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[30] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 872–881.

[31] J. Schrouff, N. Harris, S. Koyejo, I. M. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen *et al.*, "Diagnosing failures of fairness transfer across distribution shift in real-world medical settings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 304–19 318, 2022.

[32] S. Hu, K. Zhang, Z. Chen, and L. Chan, "Domain generalization via multidomain discriminant analysis," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 292–302.

[33] H. Zhang, Y.-F. Zhang, W. Liu, A. Weller, B. Schölkopf, and E. P. Xing, "Towards principled disentanglement for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8024–8034.

[34] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 407–25 437.

[35] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8035–8045.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[37] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," in *International Conference on Learning Representations*, 2020.

[38] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *arXiv preprint arXiv:1804.10745*, 2018.

[39] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 383–14 392.

[40] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos, "Beyond neural scaling laws: beating power law scaling via data pruning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 523–19 536, 2022.

[41] M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An empirical study of example forgetting during deep neural network learning," in *International Conference on Learning Representations*, 2019.

[42] B. Zhu, X. Shi, N. Erickson, M. Li, G. Karypis, and M. Shoaran, "XTab: Cross-table pretraining for tabular transformers," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[43] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Synthetic prompting: Generating chain-of-thought demonstrations for large language models," in *International Conference on Machine Learning*, vol. 202. PMLR, 2023, pp. 30 706–30 775.

[44] X. Zhou, Y. Lin, R. Pi, W. Zhang, R. Xu, P. Cui, and T. Zhang, "Model agnostic sample reweighting for out-of-distribution learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 27 203–27 221.

[45] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.