

# ODP-Bench: Benchmarking Out-of-Distribution Performance Prediction

Han Yu<sup>1,†</sup>, Kehan Li<sup>1,†</sup>, Dongbai Li<sup>1</sup>, Yue He<sup>2</sup>, Xingxuan Zhang<sup>1</sup>, Peng Cui<sup>1,\*</sup>

<sup>1</sup>Department of Computer Science, Tsinghua University

<sup>2</sup>School of Information, Renmin University of China

yuh21@mails.tsinghua.edu.cn, lkh20@mails.tsinghua.edu.cn, ldb22@mails.tsinghua.edu.cn

hy865865@gmail.com, xingxuanzhang@hotmail.com, cuip@tsinghua.edu.cn

## Abstract

Recently, there has been gradually more attention paid to Out-of-Distribution (OOD) performance prediction, whose goal is to predict the performance of trained models on unlabeled OOD test datasets, so that we could better leverage and deploy off-the-shelf trained models in risk-sensitive scenarios. Although progress has been made in this area, evaluation protocols in previous literature are inconsistent, and most works cover only a limited number of real-world OOD datasets and types of distribution shifts. To provide convenient and fair comparisons for various algorithms, we propose Out-of-Distribution Performance Prediction Benchmark (ODP-Bench), a comprehensive benchmark that includes most commonly used OOD datasets and existing practical performance prediction algorithms. We provide our trained models as a testbench for future researchers, thus guaranteeing the consistency of comparison and avoiding the burden of repeating the model training process. Furthermore, we also conduct in-depth experimental analyses to better understand their capability boundary.

## 1. Introduction

Although deep learning has achieved significant progress in many applications [1, 10, 55, 83], their performance heavily relies on the assumption that test data follows the same distribution as training data does, known as the I.I.D. assumption. Yet in wild environments, such an assumption can be easily violated and models are likely to encounter severe performance degradation in the face of distribution shifts [74], which severely hinders applications of deep learning models in risk-sensitive areas like autonomous driving [13] and medical imaging [92]. In recent years, there have been many algorithms proposed to improve the Out-of-Distribution (OOD) generalization ability of mod-

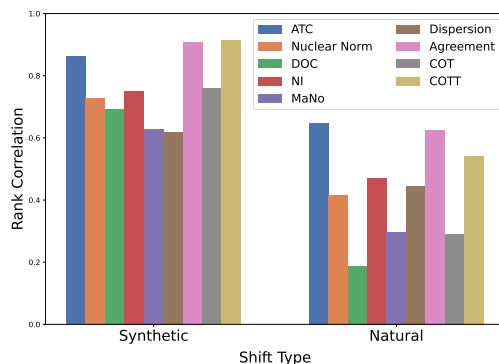


Figure 1. Rank correlation between predicted scores and ground truth performances. We can see that most of current algorithms do well in synthetic corruptions, but generally fail in natural shifts.

els, including invariant learning [2, 20, 51], domain generalization [11, 40, 66], distributionally robust optimization [28, 76, 85], stable learning [21, 94, 102], etc. Nevertheless, none of these algorithms could substantially improve the OOD performance [31, 42] and learn trustworthy models that can be deployed under risk-sensitive scenarios.

Instead of developing algorithms and training new models to improve the OOD generalization ability, with a tremendous number of off-the-shelf trained models nowadays, it is efficient and meaningful to directly leverage them by taking a look at the other side of the coin [95]: Given trained models, we aim to predict their performance on unlabeled OOD datasets, i.e. *OOD performance prediction*. With successful performance prediction, we could safely apply models to their well-performing scenarios and use them with caution in their poorly-performing scenarios, e.g. simply avoiding their usage or cooperating with humans in such scenarios. Meanwhile, we could conduct model selection from a pool of models when facing unseen environments. Therefore, we could broaden the application of off-the-shelf deep learning models in risk-sensitive areas.

Recently, more attention has been paid to the area of

<sup>†</sup>Equal contribution, <sup>\*</sup>Corresponding author

OOD performance prediction from various perspectives. They try to leverage model confidence [29, 30], distribution discrepancy [24, 58], model agreement [3, 69], etc., to carry out the task of performance prediction. Some literature tries to predict the value of model performance [25, 29, 60] while more proposes to calculate a score as the surrogate of performance [62, 87, 88]. Although their improvement and effectiveness have been preliminarily verified through experiments on real-world datasets [34, 36, 67, 68], the evaluation protocols in previous works are inconsistent, including details of training models whose performance needs to be predicted, OOD test datasets, and evaluation metrics. Meanwhile, both the number of real-world test datasets and the types of distribution shifts are relatively limited. For example, datasets widely used in domain generalization [46, 52, 84] and subpopulation shift [43, 56, 70], which are two vital sub-areas in OOD generalization, are rarely covered in previous works of OOD performance prediction. In such cases, it becomes difficult to make fair and comprehensive comparisons between different performance prediction algorithms. Moreover, the application range and capability boundary of current performance prediction algorithms have not been well-determined and clearly explored.

To fully understand and investigate existing algorithms and better promote future research in this area, in this paper we establish Out-of-Distribution Performance Prediction Benchmark (ODP-Bench). We provide 1,444 trained models of various architectures, weight initializations, and training algorithms, which are open-sourced and directly available for future researchers so that they do not need to repeat the model training process. Our benchmark includes 29 OOD datasets that cover diverse types of distribution shifts, and 10 performance prediction algorithms. Besides, we provide our codebase<sup>1</sup> where a new proposed algorithm can be easily added with small efforts of code implementation. In this way, we enable convenient and fair comparisons for both current and future algorithms. Furthermore, we conduct empirical analyses of performance prediction algorithms. One pivotal conclusion is that, as revealed in Figure 1, current algorithms show their effectiveness on OOD datasets with synthetic corruptions, which are exactly the most commonly adopted datasets in previous works of performance prediction, but cannot address diverse and complex natural distribution shifts well. This suggests the necessity and significance of our comprehensive benchmark for promoting future research on performance prediction. Our contributions are listed below:

- We propose ODP-Bench, a large comprehensive benchmark of OOD performance prediction covering 29 OOD datasets and 10 algorithms, where diverse types of distribution shifts are included.
- Our benchmark provides 1,444 off-the-shelf models and

a codebase for future researchers to easily add and test newly proposed algorithms on. This enables fair and convenient comparisons between different algorithms.

- We conduct further experimental analyses to better understand the capability boundary of current algorithms.

## 2. Related Work

**Out-of-Distribution Generalization** There have been multiple branches of research devoted to OOD generalization of machine learning models [74]. Invariant learning aims to capture the invariant relationship between the outcome and input covariates given training data from different environments [2, 48, 51], but usually limited in the range of tabular data. Distributionally robust optimization (DRO) pursues a different objective, i.e. finding the worst performing distribution in a ball centered around the training distribution and optimizing under it [28, 76], yet it suffers from over-pessimism. Focusing on visual data, domain generalization proposes diverse solutions, including invariant representation learning [66, 72], data augmentation [106], meta learning [4, 53, 54], flatness-aware optimization [11, 104, 105], etc. Another branch is stable learning which aims to decorrelate covariates via sample reweighting to remove spurious correlations [21, 102] but suffers from variance inflation [94, 97].

**Out-of-Distribution Performance Prediction** Despite progress has been made in OOD generalization, none of current algorithms have empirically shown a significant improvement over simple empirical risk minimization (ERM) [31, 42]. As suggested [95], we can put more emphasis on the evaluation of OOD generalization, among which OOD performance prediction is an important aspect. Some take advantage of different model output properties like model confidence [29], prediction dispersity [26], matrix norm [88], feature separability [87], neighborhood invariance [60], rotation invariance [25], etc. Some measures the discrepancy between covariates of training distribution and those of OOD test distribution [24, 58]. Another series leverages the phenomenon of agreement-on-the-line [3, 69] given a number of trained models.

## 3. Notations and Problem

**Notations** We use  $X \in \mathcal{X}$  to denote input variables and  $Y \in \mathcal{Y}$  to denote the outcome, where  $\mathcal{X}$  and  $\mathcal{Y}$  denote their support.  $P^{tr}(X, Y)$  denotes training distribution and  $P^{te}(X, Y)$  denotes test distribution. In the OOD circumstance,  $P^{te}(X, Y) \neq P^{tr}(X, Y)$ . A trained model is denoted as  $f_{\theta_0} : \mathcal{X} \rightarrow \mathcal{Y}$  with fixed parameters  $\theta_0$ . A validation dataset is denoted as  $\{x_i^{va}, y_i^{va}\}_{i=1}^{n_{va}}$  and an OOD test dataset is denoted as  $\{x_i^{te}, y_i^{te}\}_{i=1}^{n_{te}}$ . Usually validation data follows the training distribution, i.e.  $(x_i^{va}, y_i^{va}) \sim$

<sup>1</sup><https://github.com/h-yu16/Performance.Prediction/>

Table 1. Information of all datasets included in ODP-Bench. For domain generalization (DG), there is multiple training and test settings for each dataset, so the sample size indicates the entire dataset. “DG” is short for domain generalization. “#” indicates the number or size.

Source	Test Dataset	#Classes	#Train Set	#Val Set	#Test Set	Shift Type	#Trained Models
CIFAR-10	CIFAR-10-C	10	50000	10000	19*50000	Corruption	57
	CIFAR-10.1	0			2021+2000	Data collection	
	CIFAR-10.2	10			2000	Data collection	
	CINIC-10	10			70000	Data collection	
	STL-10	10			7200	Data collection	
CIFAR-100	CIFAR-100-C	100	50000	10000	19*50000	Corruption	108
ImageNet	TinyImageNet-C	200	1281167	50000	15*5*10000	Corruption	109
	ImageNet-C	1000			19*5*50000	Corruption	
	ImageNet-S	1000			50889	Style	
	ImageNet-R	200			30000	Style	
	ObjectNet	313			18574	Camera location	
	ImageNet-V2	1000			3*10000	Data collection	
	ImageNet-A	200			7500	Adversarial	
	ImageNet-Vid	30			22179	Time	
WILDS	iWildCam	182	129809	7314	42791	Camera location	90
	FMoW	62	76863	11483	22108	Time, region	90
	Camelyon17	2	302436	33560	85054	Data collection	90
	RxRx1	1139	40612	40612	34432	Batch effect	90
	Amazon	5	245502	46950	100050	Data collection	39
	CivilComments	2	269038	45180	133782	Demographics	30
DG	PACS	7		9991		Style	120
	OfficeHome	65		15588		Style	120
	DomainNet	345		586920		Style	90
	NICO++	60		88926		Background	90
	VLCS	5		10729		Data collection	120
	TerraInc	10		24330		Camera location	120
Subpop	Waterbirds	2	4795	1199	642	Background	30
	CelebA	2	162770	19867	180	Demographics	30
	CheXpert	2	167093	22280	661	Demographics	30

$P^{tr}(X, Y)$ , and test data  $(x_i^{te}, y_i^{te}) \sim P^{te}(X, Y)$ .

**Problem** Given a trained model  $f_{\theta_0}$ , a labeled validation dataset  $\{x_i^{va}, y_i^{va}\}_{i=1}^{n_{va}}$ , and an unlabeled OOD test dataset  $\{x_i^{te}\}_{i=1}^{n_{te}}$ , the goal is to predict performance of the model on the test dataset so that it is close to ground truth performance, or to calculate a score positively correlated with ground truth performance. For a few algorithms of direct performance estimation, we can calculate the gap between estimated performances and ground truth performances for evaluation. For algorithms predicting a surrogate score instead of estimating the value of performance, since usually there are multiple trained models or multiple OOD test datasets, we can measure the correlation between surrogate scores and ground truth performances for evaluation.

## 4. Benchmark Design





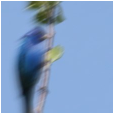

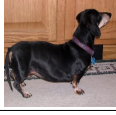

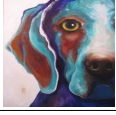

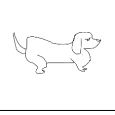


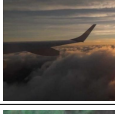




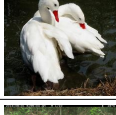

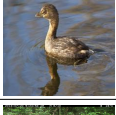


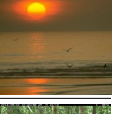


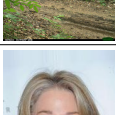





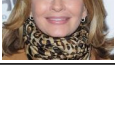



In this section, we introduce the organization of our benchmark. More details are in Appendix A.1.

### 4.1. Datasets

In previous works of OOD performance prediction, they only evaluate on a limited number of datasets, and they

mostly focus on the case where ImageNet and CIFAR-10 are treated as the training dataset. They seldom evaluate performance prediction algorithms on OOD datasets that are widely adopted in domain generalization and subpopulation shift. Besides, the types of distribution shifts previous works have covered are not diverse enough. They have mostly covered shifts of synthetic corruptions, style, and the process of data collection, but they seldom investigate shifts caused by camera locations, image backgrounds, and demographic attributes. To establish a comprehensive benchmark, we include 29 OOD datasets in our benchmark, covering most commonly used datasets in the area of OOD generalization and OOD performance prediction and more diverse types of distribution shifts. We list detailed information of the included OOD datasets in Table 1. We also provide examples from representative OOD datasets for common types of shifts in Table 2. For variants of CIFAR [50], we include CIFAR-10-C and CIFAR-100-C [34] as synthetic distribution shifts induced by corruptions, and we include CIFAR-10.1 [67], CIFAR-10.2 [57], CINIC-10 [22], and STL-10 [19] as representatives of real-world shifts. For variants of ImageNet [23], we include ImageNet-C and TinyImageNet-C [34] as synthetic distribution shifts also caused by corrup-

Table 2. Examples of representative datasets for various types of distribution shift. For the shift type of demographics, it represents demographic attributes like sex. For example, in training and validation data of CelebA, they mostly consist of blond hair female and black hair male, but in test data it mostly consists of blond hair male.

Type	Dataset	Val			Test		
Corruption	ImageNet-C						
Style	PACS						
Background	NICO++						
Data collection	VLCS						
Camera location	iWildCam						
Demographics	CelebA						

tions, and we include ImageNet-V2 [68], ImageNet-S [86], ImageNet-A [36], ImageNet-R [35], ObjectNet [6], and ImageNet-Vid-Robust [73] covering various types of real-world shifts. For WILDS [47], we choose iWildCam [8], FMoW [18], Camelyon17 [5], RxRx1 [78], Amazon [61], and CivilComments [9] as representative datasets. For domain generalization, we include all commonly used ones: PACS [52], VLCS [46], OfficeHome [84], TerraInc [7], DomainNet [63], and NICO++ [103]. For subpopulation shift, we include three most commonly used ones: Waterbirds [70], CelebA [56], and CheXpert [43]. Note that for variants of ImageNet and CIFAR, and datasets from WILDS, each dataset corresponds to only one fixed setting of training and testing, while for domain generalization datasets, each employs leave-one-domain/group-out practice, i.e. testing on one domain or one group of domains and training on the rest domains, thus leading to multiple settings. For subpopulation shift, we treat the worst subgroup as the OOD test dataset, which is fixed and determined in advance according to experiments in previous literature.

## 4.2. Training

We provide a total of 1,444 off-the-shelf trained models for subsequent researchers to directly test their performance prediction algorithms on. Note that this not only enables

fair comparisons between algorithms on the same testbench, but also greatly reduces their burden of training models repeatedly. The number of trained models for each dataset can also be referred to in Table 1. For variants of ImageNet, we directly leverage 109 open-sourced models of different architectures included in Torchvision. For variants of CIFAR-10 and CIFAR-100, we train from scratch with each architecture three random seeds following default settings in two public repositories<sup>23</sup>. The former yields  $19 \times 3$  models while the latter  $36 \times 3$  models. For WILDS, we follow its default setting [47] to train one model initialized from ImageNet supervised pretrained weights for 30 different model architectures. For domain generalization datasets, we follow Gulrajani and Lopez-Paz [31] and Yu et al. [96] to adopt the leave-one-domain-out setting for PACS, VLCS, OfficeHome, and TerraInc, and the leave-one-group-out setting for DomainNet and NICO++, and we use supervised, MoCo-v2 [14], and CLIP [64] pretrained weights as initialization for ResNet-50 [32], and supervised, MoCo-v3 [15], and CLIP pretrained weights as initialization for ViT-B/16 [27]. Each setting leads to 5 models with different random seeds, yielding 30 trained models. For subpopulation shift datasets, we employ the same training

<sup>2</sup><https://github.com/kuangliu/pytorch-cifar>

<sup>3</sup><https://github.com/weiaicunzai/pytorch-cifar100>



protocol as that of domain generalization, thus each also yielding 30 trained models.

### 4.3. Algorithms

In our benchmark, we evaluate 10 performance prediction algorithms, covering almost all practical ones. We do not include algorithms that require additional training of models [12, 25, 98]. We list adopted algorithms below:

- Average Thresholded Confidence (ATC) [29]: Learn a threshold on model confidence and employ fraction of samples with higher confidence as the predicted accuracy.
- Difference of Confidences (DoC) [30]: Calculate the difference between average model confidence on the OOD test dataset and the validation dataset.
- Nuclear Norm [26]: Adopt nuclear norm of prediction matrix to measure both confidence and dispersity.
- Neighborhood Invariance (NI) [60]: Calculate the label invariance across augmented examples in the neighborhood of a given test sample.
- Matrix Norm (MaNo) [88]: Calculate the  $L_p$  norm of the prediction matrix.
- Dispersion Score (Dispersion) [87]: Measure inter-class feature separability with average distances between each feature cluster centroid induced by pseudo labeling and the centroid of all features.
- Meta-Distribution Energy (MDE) [62]: Calculate average energy of unlabeled test data via probability density of the Gibbs distribution.
- Agreement [3]: Leverage agreement-on-the-line to predict model performance based on model agreement.
- Confidence Optimal Transport (COT) [58]: Calculate Wasserstein distance between predicted label distribution on test data and true label distribution on validation data.
- Confidence Optimal Transport with Thresholding (COTT) [58]: A variant of COT that applies thresholding to transport costs to improve estimation.

### 4.4. Evaluation Metrics

In previous literature, only a few methods are capable of directly estimating the value of accuracy. Actually, there is usually a series of trained models or OOD datasets available. Thus it is generally practical enough to provide a series of surrogate scores so that the predicted best-performing model or dataset can be selected. Therefore, for each OOD dataset, we calculate **Spearman’s rank correlation**  $\rho$  between predicted scores and ground truth performances of a set of trained models to evaluate the effectiveness of performance prediction algorithms:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (R(\hat{S}_i) - R(Acc_i))^2}{n(n^2 - 1)} \quad (1)$$

Where  $\hat{S}_i$  represents the  $i$ -th predicted score,  $Acc_i$  represents the  $i$ -th ground truth accuracy, and  $R(\cdot)$  implies the

rank of  $\hat{S}_i$  or  $Acc_i$  in their sequence. This metric has been widely adopted before [3, 26, 30, 62, 87–89]. We do not choose coefficient of determination  $R^2$  as the main evaluation metric since it is quite sensitive to outliers and it cannot address the nonlinear correlation. We will conduct analyses of  $R^2$  and  $\rho$  in Section 5.2. Besides, some previous works calculate  $R^2$  and  $\rho$  of models only when the models share the same architecture, and most of the results are larger than 0.95, even quite close to 1 [26, 62, 87–89]. Although this demonstrates the effectiveness of the proposed algorithms to some extent, it might not be practical enough since there are usually different architectures in a given pool of trained models. Thus in our benchmark ODP-Bench, we calculate these metrics for models across different architectures, which is both more challenging and more meaningful.

## 5. Experiments

In this section, we present detailed results of our established benchmark ODP-Bench and conduct other experimental analyses of evaluation metrics, model architectures, pretrained weights, etc.

### 5.1. Benchmark results

The complete results are shown in Table 3, where the datasets are sorted in a descending order of average rank correlations achieved by the 10 algorithms. The last column is the number of effective algorithms for a specific OOD dataset, where we defined “effective” as  $\rho > 0.7$ . We can see that many algorithms achieve a considerable rank correlation on OOD datasets whose shifts are created by synthetic corruptions, including CIFAR-10-C, CIFAR-100-C, and ImageNet-C. These are exactly the most commonly adopted datasets in previous literature. This indicates that it would be better to shift focus to more complex real-world distribution shifts when evaluating performance prediction algorithms in the future. For OOD datasets whose shifts are induced by image style, effectiveness of current algorithms gradually decrease with the increasing complexity and diversity of styles. For PACS and OfficeHome that have four different styles, about half of algorithms achieve satisfying results. For DomainNet with six styles, only two algorithms exhibit effectiveness. For OOD datasets whose shifts come from the process of data collection, including CINIC-10, STL-10, CIFAR-10.1/10.2, and VLCS, fewer than half of algorithms work well. This might be due to the complexity of data collection process. Nevertheless, for NICO++ composed of six different domains whose shifts are generated by image background, the shifts should have been complex enough, while almost all algorithms are capable of achieving a high rank correlation. This could be because the complex shifts of background contribute mostly to the covariate shift, i.e. shift caused by  $P(X)$  instead of concept shift, i.e. shift caused by  $P(Y|X)$  in NICO++, since Zhang et al.

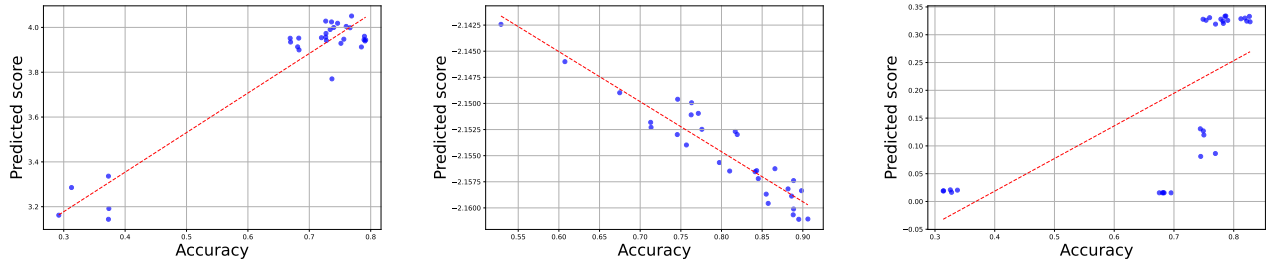
Table 3. Complete results of ODP-Bench including 29 OOD datasets and 10 OOD performance prediction algorithms. The adopted metric is Spearman’s rank correlation between predicted scores and ground truth performances. The last column represents the number of effective algorithms on the corresponding dataset, where “effective” is defined as  $\rho > 0.7$ .

Dataset	ATC	Nu. Norm	DOC	NI	MaNo	Dispersion	MDE	Agreement	COT	COTT	Avg.	#Effective
CIFAR-10-C	0.951	0.934	0.884	0.818	0.836	0.903	-0.832	0.991	0.980	0.990	0.746	9
CIFAR-100-C	0.948	0.829	0.874	0.868	0.738	0.718	-0.703	0.971	0.892	0.985	0.712	9
NICO++	0.970	0.797	0.742	0.929	0.846	0.802	-0.777	0.932	0.844	0.965	0.705	9
OfficeHome	0.911	0.594	0.465	0.912	0.754	0.541	-0.409	0.936	0.693	0.933	0.633	5
ObjectNet	0.794	0.674	0.640	0.875	0.672	0.689	-0.670	0.973	0.706	0.845	0.620	5
ImageNet-C	0.933	0.596	0.556	0.790	0.453	0.581	-0.506	0.933	0.583	0.839	0.576	4
ImageNet-V2	0.993	0.260	0.958	0.845	0.149	0.323	-0.216	0.996	0.357	0.986	0.565	5
ImageNet-S	0.961	0.500	0.480	0.846	0.343	0.438	-0.384	0.838	0.582	0.987	0.559	4
PACS	0.708	0.769	0.657	0.825	0.489	0.772	-0.736	0.909	0.498	0.636	0.553	5
CIFAR-10.1	0.717	0.526	0.423	0.766	0.524	0.533	-0.524	0.850	0.556	0.718	0.509	4
CIFAR-10.2	0.780	0.511	0.364	0.800	0.511	0.511	-0.508	0.807	0.541	0.750	0.507	4
ImageNet-R	0.742	0.325	0.338	0.859	0.414	0.232	-0.278	0.926	0.404	0.885	0.485	4
TinyImageNet-C	0.617	0.553	0.454	0.530	0.487	0.274	-0.343	0.740	0.583	0.839	0.473	2
DomainNet	0.841	0.563	0.344	0.235	0.624	0.563	-0.565	0.433	0.601	0.874	0.451	2
STL-10	0.718	0.432	0.003	0.825	0.412	0.419	-0.402	0.771	0.567	0.745	0.449	4
FMoW	0.964	0.716	0.172	0.299	-0.305	0.678	-0.674	0.859	0.681	0.955	0.435	4
CINIC-10	0.706	0.318	-0.075	0.867	0.320	0.307	-0.301	0.794	0.423	0.728	0.409	4
VLCS	0.472	-0.036	0.142	0.699	0.520	0.145	0.060	0.777	0.112	0.457	0.335	1
iWildCam	0.669	0.103	0.341	0.540	0.023	0.273	0.012	0.508	0.190	0.618	0.328	0
ImageNet-Vid	-0.095	0.712	-0.690	0.556	0.606	0.648	-0.597	0.654	0.743	0.737	0.328	3
RxRx1	0.983	0.882	-0.738	-0.538	-0.356	0.863	-0.888	0.979	0.885	0.979	0.305	6
ImageNet-A	0.503	0.135	0.208	0.818	0.209	0.143	-0.158	0.197	0.122	0.530	0.270	1
Amazon	0.749	0.077	-0.225	-	0.091	0.073	-0.131	0.810	0.123	0.640	0.245	3
Camelyon17	0.296	0.632	-0.506	-0.513	0.056	0.747	0.315	-0.637	0.796	0.668	0.185	2
TerraInc	0.370	0.440	0.282	0.355	-0.067	0.438	-0.482	0.422	-0.122	-0.062	0.157	0
CivilComments	0.806	-0.779	-0.096	-	0.454	-0.850	0.496	0.818	-0.350	0.537	0.115	3
Waterbirds	0.219	0.449	0.825	-0.194	0.149	0.475	0.037	0.473	-0.960	-0.883	0.059	1
CelebA	0.392	0.498	0.369	0.027	0.463	0.602	-0.350	0.162	-0.922	-0.825	0.042	0
CheXpert	0.052	0.327	-0.752	-0.786	-0.472	0.782	0.535	-0.530	-0.786	-0.853	-0.248	1

[103] find that NICO++ has smaller concept shift and larger covariate shift compared with other domain generalization datasets. In addition, it is worth noting that all three subpopulation datasets seem to be extremely “hard” with no or only one algorithm achieving a rank correlation higher than 0.7. Many algorithms even achieve a negative rank correlation on these datasets. This might be due to over-confidence in certain subpopulations. For example, in CelebA, ERM-trained models tend to classify faces of blond hair male as not blond hair with high confidence since it learns the strong spurious correlation between blond hair and female. This indicates that more attention could be paid to performance

prediction associated with subpopulation shifts.

From the perspective of the evaluated algorithms, we can see that ATC, NI, Agreement, and COTT achieve relatively desirable overall results in performance prediction, even some of them are quite simple and straightforward compared with the rest of algorithms. Nevertheless, all of them still fail on many datasets, indicating that they are not universal enough to address diverse and complicated scenarios. It is also worth noting that the four algorithms adopt completely different practices: ATC employs model confidence, NI leverages invariance under data augmentations, Agreement takes advantage of the property of generalization dis-



(a) MaNo applied on Real domain of OfficeHome, where  $R^2 = 0.912$ ,  $\rho = 0.475$ . (b) MDE applied on Art domain of PACS, where  $R^2 = 0.881$ ,  $\rho = -0.911$ . (c) Dispersion applied on Product domain of OfficeHome, where  $R^2 = 0.475$ ,  $\rho = 0.793$ .

Figure 2. Scatter plots of predicted scores against ground truth accuracies.

Table 4. Rank correlation measured on PACS with models trained by different OOD generalization algorithms. Overall, performance prediction algorithms achieve slightly higher rank correlation for models trained with OOD generalization algorithms than with ERM.

OOD Algorithm	ATC	Nu. Norm	DOC	NI	MaNo	Dispersion	MDE	Agreement	COT	COTT	Avg.
ERM	0.708	0.769	0.657	0.825	0.489	0.772	-0.736	0.909	0.498	0.636	0.553
RSC	0.744	0.762	0.653	0.728	0.788	0.778	-0.716	0.893	0.534	0.785	0.595
Mixup	0.784	0.758	0.678	0.798	0.610	0.766	-0.687	0.854	0.804	0.831	0.620
SWAD	0.698	0.730	0.642	0.778	0.589	0.715	-0.631	0.856	0.605	0.704	0.569
CORAL	0.740	0.719	0.672	0.740	0.460	0.729	-0.619	0.885	0.599	0.770	0.569

agreement equality [44], and COTT focuses on pseudo label shift. This implies that it is still not clear which direction is the correct and most promising one in the area of OOD performance prediction, and the development of performance prediction algorithms is far from convergence.

## 5.2. Analysis of evaluation metrics

We mainly focus on the comparison between Spearman’s rank correlation  $\rho$  and the coefficient of determination  $R^2$ . We pick three showcases for analyses shown in Figure 2: MaNo applied on Real domain of OfficeHome, MDE applied on Art domain of PACS, and Dispersion applied on Product domain of OfficeHome. In Figure 2a, it shows that  $R^2$  is relatively high but  $\rho$  is quite low, and we can clearly see that there are several outliers with accuracies lower than 0.4. When there are outliers,  $R^2$  could become rather large without really exhibiting a strong linear correlation. In such a case,  $\rho$  is more appropriate. In Figure 2b, the scores and the ground truth accuracies are reversely correlated but  $R^2$  is very high since there is a strong linear correlation. In such a case,  $\rho$  is also more appropriate. Figure 2c shows the opposite, i.e.  $R^2$  is relatively low but  $\rho$  is quite high. We can see that although the scores and the ground truth accuracies are not linearly correlated, they still show a relatively monotonic pattern, which also matches our requirement of evaluation. Overall, considering these failure cases of  $R^2$ , we treat Spearman’s rank correlation  $\rho$  as a more proper and effective evaluation metric for OOD performance prediction, and we adopt it in our main experiments. Discussion related

to other metrics can be referred to in Appendix A.2.1.

## 5.3. Analysis of OOD generalization methods

In previous works, performance prediction algorithms were only applied on models trained with simple ERM. To investigate the effect of OOD generalization methods that are used to train models, we choose four representative OOD generalization methods: RSC [40], Mixup [99], SWAD [11], and CORAL [77], and train models on PACS with them. Results are in Table 4. We can see that the change of OOD generalization methods has a large influence in some cases, e.g. MaNo applied to models trained with ERM and RSC, COTT applied to models trained with ERM and Mixup. Overall, when applied to models trained with OOD generalization methods, performance prediction algorithms could achieve higher rank correlations compared with models trained with ERM. This could be because OOD generalization methods help reduce the performance gap between training and test distribution, which might be easier for performance prediction algorithms to work.

## 5.4. Analysis of pretraining strategies

We investigate pretraining strategies including supervised pretraining, MoCo (MoCo-v2 for ResNet-50 and MoCo-v3 for ViT-B/16), and CLIP on PACS. Results of ATC, NI, Agreement, and COTT, the four most effective algorithms, are shown in Figure 3a. We can see that pretraining strategies have a large influence on the effectiveness of performance prediction. For ATC, NI, and COTT, the rank

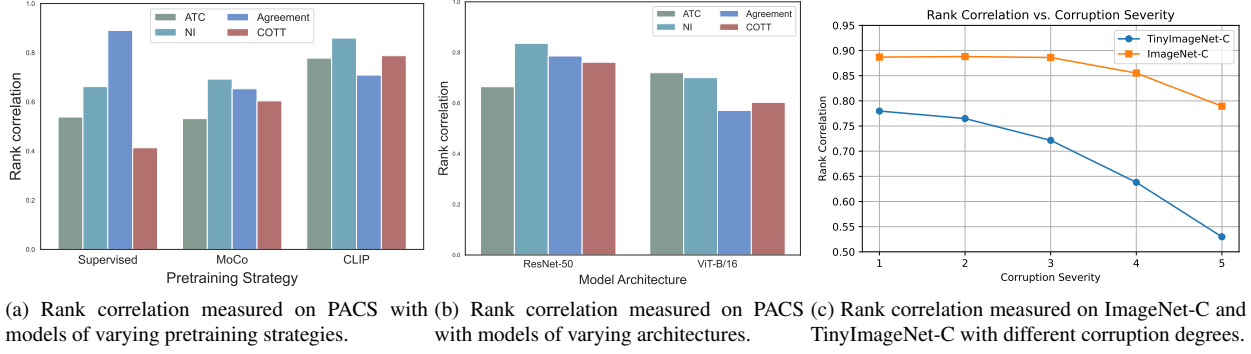


Figure 3. Analyses of pretraining strategies, model architectures, and degree of distribution shift.

correlation increases as the pretraining strategy changes from supervised to MoCo and CLIP. This could be because MoCo and CLIP are contrastive learning strategies that do not directly rely on the  $y$  label, so models initialized with these pretrained weights are less likely to encounter overconfidence and could be better calibrated, making it easier for performance prediction algorithms to apply to. For Agreement, the supervised pretraining strategy performs best, which could be because the property of generalization disagreement equality [44] holds better with supervised pre-trained weight initialization.

### 5.5. Analysis of model architectures

To investigate the effect of model architectures on the results of performance prediction, we compare ResNet-50 and ViT-B/16 on PACS. From Figure 3b, we can see that for most algorithms, changing to a larger architecture results in even lower rank correlations except for ATC. This indicates that it is generally harder to predict performances of larger models, but confidence-based algorithms are worth being explored for scaling to larger models in future research.

### 5.6. Analysis of degree of distribution shift

To characterize the degree of distribution shift, we investigate NI on TinyImageNet-C and ImageNet-C, both of which have 5 degrees of distribution shift depicted by the severity of corruption. Figure 3c shows that the effectiveness of NI decreases as the corruption severity grows. This could indicate that performance prediction algorithms work better with smaller distribution shifts. However, as mentioned in Section 5.1 and revealed in Table 3, datasets like NICO++ and ObjectNet also exhibits strong distribution shift, but many algorithms show their effectiveness on these datasets, indicating the requirement of further in-depth analyses.

### 5.7. Analysis of subpopulations

In Table 3, results of our benchmark show that current performance prediction algorithms fail on subpopulation shift datasets where subpopulations are induced by attributes and

categories with a strong spurious correlation. Here we conduct experiments by conducting performance prediction on each category, which also formulates subpopulations but without strong spurious correlations. We choose five algorithms that do not require diversity of category labels. From Table 5, we can see that the phenomenon of failure on subpopulation shift datasets does not occur here. This implies that spurious correlations play a key role in the failure of performance predictions under subpopulation shift situations, which could be paid more attention to in the future.

Table 5. Rank correlation measured on each category of PACS.

Category	ATC	DOC	NI	MaNo	Agreement	Avg.
$y = 0$	0.737	0.782	0.822	0.425	0.915	0.736
$y = 1$	0.817	0.811	0.847	0.328	0.895	0.740
$y = 2$	0.737	0.769	0.779	0.445	0.892	0.725
$y = 3$	0.815	0.752	0.760	0.360	0.916	0.721
$y = 4$	0.770	0.671	0.790	0.277	0.677	0.637
$y = 5$	0.878	0.829	0.831	0.511	0.880	0.786
$y = 6$	0.798	0.798	0.851	0.409	0.838	0.739

## 6. Conclusion

In this paper, we propose a large and comprehensive benchmark named ODP-Bench. It includes 29 OOD test datasets and 10 OOD performance prediction algorithms, covering diverse types of distribution shifts. It provides a testbench of 1,444 off-the-shelf trained models, which greatly reduces the burden of model training for future researchers and enables fair comparisons between different algorithms. It also provides a codebase so that a newly proposed algorithm can be easily added without complicated code implementation. The experimental results show that although current performance prediction algorithms exhibit effectiveness on certain types of shifts, e.g. synthetic corruptions, they are not universal enough to address all kinds of complex real-world shifts, which are left for future research.



## Acknowledgement

This work was supported by NSFC (No. 62425206), Tsinghua-Toyota Joint Research Fund, NSFC (No. 62141607), and Beijing Municipal Science and Technology Project (No. Z241100004224009). Peng Cui is the corresponding author. All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2
- [3] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35: 19274–19289, 2022. 2, 5
- [4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 2
- [5] Peter Bandi, Oscar Geessink, Quirine Manson, Marcorij Van Dijk, Maschenka Balkenhol, Meyke Hermesen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 4, 1
- [6] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 4, 1
- [7] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 4, 1
- [8] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 4, 1
- [9] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019. 4, 1
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [11] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 1, 2, 7
- [12] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34: 14980–14992, 2021. 5
- [13] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4
- [15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 4
- [16] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. <https://arxiv.org/abs/1707.01629>. 2
- [17] François Chollet. Xception: Deep learning with depth-wise separable convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. <https://arxiv.org/abs/1610.02357>. 2, 3
- [18] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 4, 1
- [19] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 3, 1
- [20] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. 1
- [21] Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022. 1, 2
- [22] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018. 3, 1
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 1

- [24] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15069–15078, 2021. 2
- [25] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pages 2579–2589. PMLR, 2021. 2, 5
- [26] Weijian Deng, Yumin Suh, Stephen Gould, and Liang Zheng. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In *International Conference on Machine Learning*, pages 7658–7674. PMLR, 2023. 2, 5
- [27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4
- [28] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. 1, 2
- [29] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. 2, 5
- [30] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1134–1144, 2021. 2, 5
- [31] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 1, 2, 4
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 2, 3
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. <https://arxiv.org/abs/1603.05027>. 2, 3
- [34] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 2, 3, 1
- [35] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 4, 1
- [36] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 2, 4, 1
- [37] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. <https://arxiv.org/abs/1709.01507>. 2, 3
- [38] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 3
- [39] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. <https://arxiv.org/abs/1608.06993>. 2, 3
- [40] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 1, 7
- [41] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5MB model size, 2016. <https://arxiv.org/abs/1602.07360>. 2, 3
- [42] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 1, 2
- [43] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 2, 4, 1
- [44] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2021. 7, 8
- [45] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022. 4
- [46] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 2, 4, 1
- [47] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 4, 1, 3
- [48] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020. 2

- [49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. <https://papers.nips.cc/paper/4824-Imagenet-classification-with-deep-convolutional-neural-networks>. 2
- [50] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 3, 1
- [51] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 1, 2
- [52] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 4, 1
- [53] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2
- [54] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 2
- [55] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 4, 1
- [57] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, page 15, 2020. 3, 1
- [58] Yuzhe Lu, Yilong Qin, Runtian Zhai, Andrew Shen, Ketong Chen, Zhenlin Wang, Soheil Kolouri, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Characterizing out-of-distribution error via optimal transport. *Advances in Neural Information Processing Systems*, 36:17602–17622, 2023. 2, 5
- [59] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In *European Conference on Computer Vision (ECCV)*, 2018. <https://arxiv.org/abs/1807.11164>. 2, 3
- [60] Nathan Huyen Ng, Neha Hulkund, Kyunghyun Cho, and Marzyeh Ghassemi. Predicting out-of-domain generalization with neighborhood invariance. *Transactions on Machine Learning Research*, 2024. 2, 5
- [61] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019. 4, 1
- [62] Ru Peng, Heming Zou, Haobo Wang, Yawen Zeng, Zenan Huang, and Junbo Zhao. Energy-based automated model evaluation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5
- [63] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 4, 1
- [64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4
- [65] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. <https://arxiv.org/abs/2003.13678>. 2
- [66] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 1, 2
- [67] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018. 2, 3, 1
- [68] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 2, 4, 1
- [69] Elan Rosenfeld and Saurabh Garg. (almost) provable error bounds under distribution shift via disagreement discrepancy. *Advances in Neural Information Processing Systems*, 36:28761–28784, 2023. 2
- [70] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 2, 4, 1
- [71] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. <https://arxiv.org/abs/1801.04381>. 2
- [72] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. 2
- [73] Vaishal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. A systematic framework for natural perturbations from videos. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. 4, 1

- [74] Zheyang 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. 1, 2
- [75] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015. <https://arxiv.org/abs/1409.1556>. 2, 3
- [76] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. 1, 2
- [77] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 7
- [78] Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4285–4294, 2023. 4, 1
- [79] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. <https://arxiv.org/abs/1409.4842>. 2, 3
- [80] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. <https://arxiv.org/abs/1512.00567>.
- [81] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. <https://arxiv.org/abs/1602.07261>. 2, 3
- [82] Mingxing Tan and Quoc Le. EfficientNet: rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. <https://arxiv.org/abs/1905.11946>. 2
- [83] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 1
- [84] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2, 4, 1
- [85] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 1
- [86] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 4, 1
- [87] Renchunzi Xie, Hongxin Wei, Lei Feng, Yuzhou Cao, and Bo An. On the importance of feature separability in predicting out-of-distribution error. *Advances in Neural Information Processing Systems*, 36:27783–27800, 2023. 2, 5
- [88] RENCHUNZI XIE, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, and Bo An. Mano: Exploiting matrix norm for unsupervised accuracy estimation under distribution shifts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 5
- [89] RENCHUNZI XIE, Ambroise Odonnat, Vasilii Feofanov, Ievgen Redko, Jianfeng Zhang, and Bo An. Gradient norm as a powerful proxy to out-of-distribution error estimation. 2024. 5
- [90] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. <https://arxiv.org/abs/1611.05431>. 2, 3
- [91] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: a closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39584–39622, 2023. 1
- [92] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10): 2838–2848, 2024. 1
- [93] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 2
- [94] Han Yu, Peng Cui, Yue He, Zheyang Shen, Yong Lin, Renzhe Xu, and Xingxuan Zhang. Stable learning via sparse variable independence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10998–11006, 2023. 1, 2
- [95] Han Yu, Jiashuo Liu, Xingxuan Zhang, Jiayun Wu, and Peng Cui. A survey on evaluation of out-of-distribution generalization. *arXiv preprint arXiv:2403.01874*, 2024. 1, 2
- [96] Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21897–21908, 2024. 4, 1
- [97] Han Yu, Yue He, Renzhe Xu, Dongbai Li, Jiayin Zhang, Wenchao Zou, and Peng Cui. Sample weight averaging for stable prediction. *arXiv preprint arXiv:2502.07414*, 2025. 2
- [98] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, pages 25721–25746. PMLR, 2022. 5
- [99] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 7



- [100] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. <https://arxiv.org/abs/1611.05725>. 2
- [101] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. <https://arxiv.org/abs/1707.01083>. 2, 3
- [102] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021. 1, 2
- [103] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyang Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16036–16047, 2023. 4, 6, 1
- [104] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5189–5202, 2023. 2
- [105] Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20247–20257, 2023. 2
- [106] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 2
- [107] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 3