NICO^{++}: Towards Better Benchmarking for Domain Generalization

Xingxuan Zhang[†], Yue He[†], Renzhe Xu, Han Yu, Zheyan Shen, Peng Cui*
Department of Computer Science, Tsinghua University

xingxuanzhang@hotmail.com, heyue18@mails.tsinghua.edu.cn, xrz199721@gmail.com,
yuh21@mails.tsinghua.edu.cn, shenzy17@mails.tsinghua.edu.cn, cuip@tsinghua.edu.cn

Abstract

Despite the remarkable performance that modern deep neural networks have achieved on independent and identically distributed (I.I.D.) data, they can crash under distribution shifts. Most current evaluation methods for domain generalization (DG) adopt the leave-one-out strategy as a compromise on the limited number of domains. We propose a large-scale benchmark with extensive labeled domains named NICO^{++} along with more rational evaluation methods for comprehensively evaluating DG algorithms. To evaluate DG datasets, we propose two metrics to quantify covariate shift and concept shift, respectively. Two novel generalization bounds from the perspective of data construction are proposed to prove that limited concept shift and significant covariate shift favor the evaluation capability for generalization. Through extensive experiments, NICO^{++} shows its superior evaluation capability compared with current DG datasets and its contribution in alleviating unfairness caused by the leak of oracle knowledge in model selection. The data and code for the benchmark based on NICO^{++} are available at https://github.com/xxgege/NICO-plus.

1. Introduction

Machine learning has illustrated its excellent capability in a wide range of areas [37, 65, 82]. Most current algorithms minimize the empirical risk in training data relying on the assumption that training and test data are independent and identically distributed (I.I.D.). However, this ideal hypothesis is hardly satisfied in real applications, especially those high-stake applications such as healthcare [10, 49], autonomous driving [1, 13, 39] and security systems [6], owing to the limitation of data collection and intricacy of the scenarios. Distribution shifts between training and test data may lead to the unreliable performance of current approaches in practice. Hence, instead of generalization within the training distribution, the ability to generalize under distribution shift, domain generalization (DG) [75, 94], is of more critical significance in realistic scenarios.

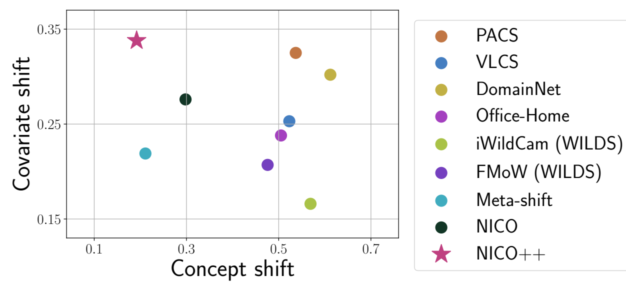

Figure 1. Covariate shift ($\mathcal{M}_{\rm cov}$ in Equation (1)) and concept shift ($\mathcal{M}_{\rm cpt}^{\max}$ in Equation (2)) of NICO^{++} and current DG datasets. NICO^{++} has the lowest concept shift and highest covariate shift, showing the superiority in evaluation capability.

In the field of computer vision, benchmarks that provide the common ground for competing approaches often play a role of catalyzer promoting the advance of research [14]. An advanced DG benchmark should provide sufficient diversity in distributions for both training and evaluating DG algorithms [74, 78] while ensuring essential common knowledge of categories for inductive inference across domains [33, 34, 93]. The first property drives generalization challenging, and the second ensures the solvability [81]. This requires adequate distinct domains and instructive features for each category shared among all domains.

Current DG benchmarks, however, either lack sufficient domains (e.g., 4 domains in PACS [40], VLCS [18] and Office-Home [73] and 6 in DomainNet [53]) or too simple or limited to simulating significant distribution shifts in real scenarios [2, 21, 30]. To enrich the diversity and perplexing distribution shifts in training data as much as possible, most of the current evaluation methods for DG adopt the leave-one-out strategy, where one domain is considered as the test domain and the others for training. This is not an ideal evaluation for generalization but a compromise due to the limited number of domains in current datasets, which impairs

[†]Equal contribution
*Corresponding Author

the evaluation capability. To address this issue, **we suggest testing DG methods on multiple test domains instead of one specific domain in each evaluation after training**.

To benchmark DG methods comprehensively and simulate real scenarios where a trained model may encounter any possible test data while providing sufficient diversity in the training data, we construct a large-scale DG dataset named NICO^{++} with extensive domains and two protocols supported by aligned and flexible domains across categories, respectively, for better evaluation. Our dataset consists of 80 categories, 10 aligned common domains for all categories, 10 unique domains specifically for each category, and more than 230,000 images. Abundant diversity in both domain and category supports flexible assignments for training and test, controllable degree of distribution shifts, and extensive evaluation on multiple target domains. Images collected from real-world photos and consistency within category concepts provide sufficient common knowledge for recognition across domains on NICO^{++}.

To evaluate DG datasets in-depth, we investigate distribution shifts on images (covariate shift) and common knowledge for category discrimination across domains (concept agreement) within them. Formally, we present quantification for covariate shift and the opposite of concept agreement, namely concept shift, via two novel metrics. We propose two novel generalization bounds and analyze them from the perspective of data construction instead of models. Through these bounds, we prove that limited concept shift and significant covariate shift favor the evaluation capability for generalization.

Moreover, a critical yet common problem in DG is the model selection and the potential unfairness in the comparison caused by leveraging the knowledge of target data to choose hyperparameters that favors test performance [3,27]. This issue is exacerbated by the notable variance of test performance with various algorithm irrelevant hyperparameters on current DG datasets. Intuitively, strong and unstable concept shift such as confusing mapping relations from images to labels across domains embarrasses training convergence and enlarges the variance.

We conduct extensive experiments on three levels. First, we evaluate NICO^{++} and current DG datasets with the proposed metrics and show the superiority of NICO^{++} in evaluation capability, as shown in Figure 1. Second, we conduct copious experiments on NICO^{++} to benchmark current representative methods with the proposed protocols. Results show that the room for improvement of generalization methods on NICO^{++} is spacious. Third, we show that NICO^{++} helps alleviate the issue by squeezing the possible improvement space of oracle leaking and contributes as a fairer benchmark to the evaluation of DG methods, which meets the proposed metrics.

2. Related Works

DG Benchmarks. After the high-speed development benefited from the datasets, like PASCAL VOC [17], ImageNet [14] and MSCOCO [45], in IID scenarios, a range of image datasets has been raised for the research of domain generalization in visual recognition. The first branch modifies traditional image datasets with synthetic transformations, typically including the ImageNet variants [29–31], MNIST variants [2, 25], Waterbirds [60], OOD-CV [92], and WILDS [38]. Another branch considers collecting data coming from different source domains, including PACS [40], Office-Home [73], DomainNet [53], Terra Incognita [4], VLCS [18], and NICO [28]. However, these datasets utilize a simple criterion to distinguish distributions, e.g. image style, not enough to cover the complexity in reality. In addition, the domains of most current DG datasets are limited, leading to inadequate diversity in training or test data. Please see the detailed comparison with the last version of NICO [28], other DG datasets, and other benchmarks [27,40] in Appendix B.

Domain Generalization. There are several streams of literature studying the domain generalization problem in vision. With extra information on test domains, domain adaptation methods [5,19,23,62,66,67,69,77,86] show effectiveness in addressing the distribution shift problems. By contrast, domain generalization aims to learn models that generalize well on unseen target domains while only data from several source domains are accessible. According to [64], DG methods can be divided into three branches, including representation learning [7, 8, 20, 24, 26, 32, 35, 50, 51], training strategies [9, 15, 33, 42, 44, 46, 59, 61, 76, 88, 90], and data augmentation methods [36,54,55,63,71,72,74,83,95]. More comprehensive surveys on domain generalization methods can be found in [75,96].

3. NICO^{++}: Domain-Extensive Large Scale Domain Generalization Benchmark

In this section, we introduce a novel large-scale domain generalization benchmark NICO^{++}, which contains extensive domains and categories. Similar to the original version of NICO [28], each image in NICO^{++} consists of two kinds of labels, namely the category label and the domain label. The category labels correspond to the objective concept (*e.g.*, cat and dog) while the domain labels represent other visual information in images, including the background of the image (*e.g.* on grass and in water), the attributes of the foreground (*e.g.* lying or running), and the relationship with other objects (*e.g.*, behind a table). To boost the heterogeneity in the dataset to support the thorough evaluation of generalization ability in domain generalization scenarios, we greatly enrich the types of categories and domains and collect a larger amount of images in NICO^{++}.

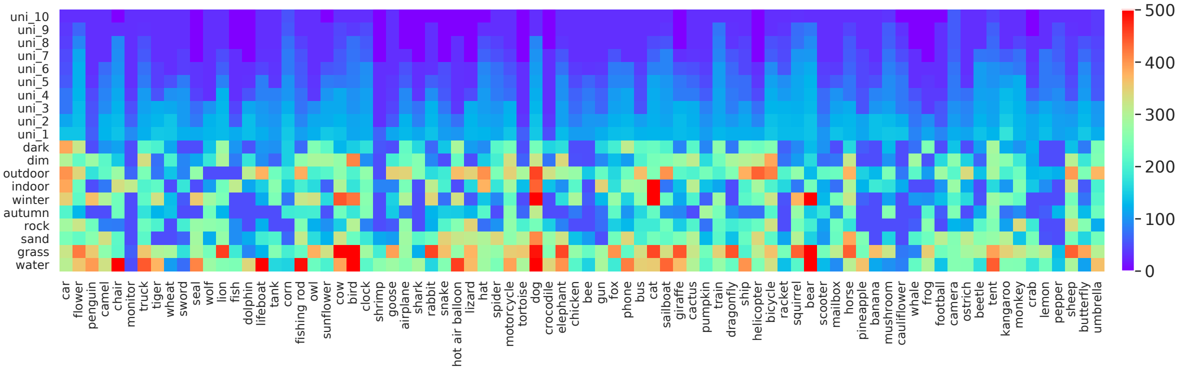

Figure 2. Statistical overview of NICO^{++}. The figure shows the number of instances in each domain and each category. The horizontal axis is for categories and the vertical axis for domains. The color of each bin corresponds to the number of instances in each *(category, domain)* pair. The 10 domains at the bottom are common domains while the 10 at the top are unique domains.

3.1. Constructions of the Category / Domain Labels

We first select 80 categories and then build 10 common and 10 category-specific domains upon them. We provide detailed statistics of the selected categories and domains in Appendix E.

Categories. A total of 80 categories are provided with a hierarchical structure in NICO^{++}. Four broad categories *Animal*, *Plant*, *Vehicle*, and *Substance* lie on the top level. For each of *Animal*, *Plant*, and *Vehicle*, there exist narrow categories derived from it (e.g., *felida* and *insect* belong to *Animal*) in the middle level. Finally, 80 concrete categories are assigned to their super-category respectively. The hierarchical structure ensures the diversity and balance[*] of categories in NICO^{++}, which is vital to simulate realistic domain generalization scenarios in wild environments.

Common domains. Towards the settings of domain generalization or domain adaption, we design 10 common domains that are aligned across all categories. Each of the selected common domains refers to a family of concrete contexts with similar semantics so that they are general and common enough to generate meaningful combinations with all categories. For example, the common domain *water* contains contexts of *swimming*, *in pool*, *in river*, etc. A comparison between common domains in NICO^{++} and domains in current DG datasets is in Appendix B.

Unique domains. To increase the number of domains and support the flexible DG scenarios where the training domains are not aligned with respect to categories, we further attain unique domains specifically for each of the 80 categories. We select the unique domains according to the following conditions: (1) they are different from the common domains; (2) they can include various concepts, such as attributes (e.g. action, color), background, camera shooting angle, and accompanying objects, etc.; (3) different types of them hold a balanced proportion for diversity.

[*]The ratio of the number of categories in *Animal*, *Plant*, *Vehicle* and *Substance* is $40 : 12 : 14 : 14$.

3.2. Data Collection and Statistics

NICO^{++} has 10 common domains, covering nature, season, humanity, and illumination, for a total of 80 categories, and 10 unique domains for each category. The capacity of the most common domains and unique domains is at least 200 and 50, respectively. The images from most domains are collected by searching a combination of a category name and a phrase extended from the domain name (e.g. "dog sitting on grass" for the category *dog* and the domain *grass*) on various search engines. Over 32,000 combinations are adopted for searching images. The downloaded data contain a large portion of outliers that require artificial annotations. Each image is assigned to two annotators to label both the category and domain and passes the selection when agreed upon by both annotators. After the annotation process, 232.4k images are selected from over 1.0 million images downloaded from the search engines. The scale of NICO^{++} is enormous enough to support the training of deep convolutional networks (*e.g.*, ResNet-50) from scratch in types of domain generalization scenarios. A statistical overview of the dataset is shown in Figure 2 and example images are shown in Figure 3.

4. Covariate Shift and Concept Shift

Consider a dataset with data points sampled from a joint distribution $P(X, Y) = P(Y|X)P(X)$. Distribution shift within the dataset can be caused by the shift on $P(X)$ (*i.e.*, covariate shift) and shift on $P(Y|X)$ (*i.e.*, concept shift) [64]. We give quantification for these two shifts in any labeled dataset and analyze the preference of them from a perspective of the DG benckmark via presenting two generalization bounds for multi-class classification. Then we evaluate NICO^{++} and current DG datasets empirically with the proposed metrics and show the superiority of NICO^{++}.

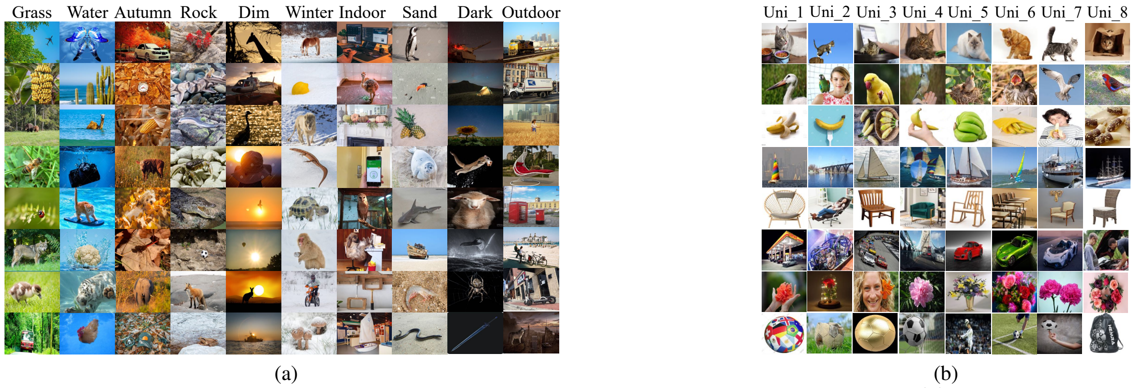

Figure 3. Example images of common (3a) and unique (3b) domains in NICO^{++}.

Notations We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the space of input X and outcome Y, respectively. We use $\Delta_{\mathcal{Y}}$ to denote a distribution on $\mathcal{Y}$. A domain d corresponds to a distribution $\mathcal{D}_d$ on $\mathcal{X}$ and a labeling function[†] $f_d : \mathcal{X} \to \Delta_{\mathcal{Y}}$. The training and test domains are specified by $(\mathcal{D}_{\text{tr}}, f_{\text{tr}})$ and $(\mathcal{D}_{\text{te}}, f_{\text{te}})$, respectively. We use $p_{\text{tr}}(x)$ and $p_{\text{te}}(x)$ to denote the probability density function on training and test domains. Let $\ell : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ define a loss function over $\Delta_{\mathcal{Y}}$ and $\mathcal{H}$ define a function class mapping $\mathcal{X}$ to $\Delta_{\mathcal{Y}}$. For any hypotheses $h_1, h_2 \in \mathcal{H}$, the expected loss $\mathcal{L}_{\mathcal{D}}(h_1, h_2)$ for distribution $\mathcal{D}$ is given as $\mathcal{L}_{\mathcal{D}}(h_1, h_2) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(h_1(x), h_2(x))]$. To simplify the notations, we use $\mathcal{L}_{\text{tr}}$ and $\mathcal{L}_{\text{te}}$ to denote the expected loss $\mathcal{L}_{\mathcal{D}_{\text{tr}}}$ and $\mathcal{L}_{\mathcal{D}_{\text{te}}}$ in training and test domain, respectively. In addition, we use $\varepsilon_{\text{tr}}(h) = \mathcal{L}_{\text{tr}}(h, f_{\text{tr}})$ and $\varepsilon_{\text{te}}(h) = \mathcal{L}_{\text{te}}(h, f_{\text{te}})$ to denote the loss of a function $h \in \mathcal{H}$ *w.r.t.* to the true labeling function f_{tr} and f_{te}, respectively.

4.1. Metrics for Covariate shift and Concept shift

The distribution shift between the training domain $(\mathcal{D}_{\text{tr}}, f_{\text{tr}})$ and test domain $(\mathcal{D}_{\text{te}}, f_{\text{te}})$ can be decomposed into covariate shift (*i.e.*, shift between $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$) and concept shift (*i.e.*, shift between f_{tr} and f_{te}). We propose the following metrics to measure them.

Definition 4.1 (Metrics for covariate shift and concept shift)**.** Let $\mathcal{H}$ be a set of functions mapping $\mathcal{X}$ to $\Delta_{\mathcal{Y}}$ and let $\ell : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ define a loss function over $\Delta_{\mathcal{Y}}$. For the two domains $(\mathcal{D}_{\text{tr}}, f_{\text{tr}})$ and $(\mathcal{D}_{\text{te}}, f_{\text{te}})$, then

- the covariate shift is measured as the discrepancy distance [47] (provided in Definition 4.2) between $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$ under $\mathcal{H}$ and ℓ, *i.e.*,

$$\mathcal{M}_{\text{cov}}\left(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}; \mathcal{H}, \ell\right) \triangleq \operatorname{disc}\left(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}; \mathcal{H}, \ell\right), \quad (1)$$

- the concept shift is measured as the maximum / minimum loss when using f_{tr} on the test domain or using f_{te} on the training domain, *i.e.*,

$$\begin{cases} \mathcal{M}_{\text{cpt}}^{\min}\left(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}, f_{\text{tr}}, f_{\text{te}}; \ell\right) \triangleq \min\left\{\mathcal{L}_{\text{tr}}(f_{\text{tr}}, f_{\text{te}}), \mathcal{L}_{\text{te}}(f_{\text{tr}}, f_{\text{te}})\right\}, \\ \mathcal{M}_{\text{cpt}}^{\max}\left(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}, f_{\text{tr}}, f_{\text{te}}; \ell\right) \triangleq \max\left\{\mathcal{L}_{\text{tr}}(f_{\text{tr}}, f_{\text{te}}), \mathcal{L}_{\text{te}}(f_{\text{tr}}, f_{\text{te}})\right\}. \end{cases} \quad (2)$$

Remark. We introduce two metrics for concept shift terms in Equation (2) because they both provide meaningful characterizations of the concept shift. In addition, both $\mathcal{M}_{\text{cpt}}^{\min}$ and $\mathcal{M}_{\text{cpt}}^{\max}$ have close connections with DG performance as shown in Theorem 4.2 and Theorem 4.3 in Section 4.2. The covariate shift is widely discussed in recent literature [16, 58, 64] yet none of them give the quantification with function discrepancy, which favors the analysis of DG performance and shows remarkable properties when $\mathcal{H}$ is large (such as the function space supported by current deep models). The concept shift can be considered as the discrepancy between the labeling rule f_{tr} on the training data and the labeling rule f_{te} on the test data. Intuitively, consider that a circle in the training data is labeled as class *A* in training domains and class *B* in test domains, models can hardly learn the labeling function on the test data (mapping the circle to class *B*) without knowledge about test domains. The discrepancy distance mentioned above is defined as follows.

Definition 4.2 (Discrepancy Distance [47])**.** Let $\mathcal{H}$ be a set of functions mapping $\mathcal{X}$ to $\Delta_{\mathcal{Y}}$ and let $\ell : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ define a loss function over $\Delta_{\mathcal{Y}}$. The discrepancy distance $\operatorname{disc}(\mathcal{D}_1, \mathcal{D}_2; \mathcal{H}, \ell)$ between two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{X}$ is $\operatorname{disc}(\mathcal{D}_1, \mathcal{D}_2; \mathcal{H}, \ell) \triangleq \sup_{h_1, h_2 \in \mathcal{H}} |\mathcal{L}_{\mathcal{D}_1}(h_1, h_2) - \mathcal{L}_{\mathcal{D}_2}(h_1, h_2)|$.

We give formal analysis of metrics for covariate shift ($\mathcal{M}_{\text{cov}}$) and concept shift ($\mathcal{M}_{\text{cpt}}^{\min}$/$\mathcal{M}_{\text{cpt}}^{\max}$) below and the graphical explanation is shown in Figure 4.

The covariate shift term $\mathcal{M}_{\text{cov}}$. When the capacity of function class $\mathcal{H}$ is large enough and ℓ is bounded, $\mathcal{M}_{\text{cov}}$ is in terms of the ℓ_1 distance between two distributions, given by the following proposition.

Proposition 4.1. *Let $\mathcal{H}$ be the set of all functions mapping $\mathcal{X}$ to $\Delta_{\mathcal{Y}}$ and the range of the loss function*

[†]We use $\Delta_{\mathcal{Y}}$ here to denote that the labeling function may not be deterministic. This formulation also includes deterministic labeling function cases.

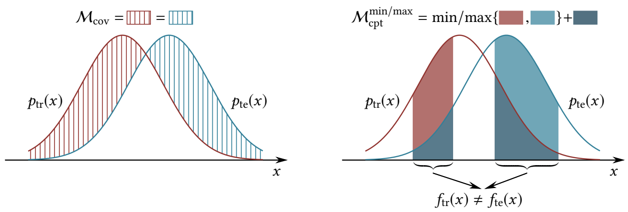

Figure 4. Graphical explanations of our proposed metric $\mathcal{M}_{\text{cov}}$ and $\mathcal{M}_{\text{cpt}}^{\text{min}}/\mathcal{M}_{\text{cpt}}^{\text{max}}$ when $\mathcal{H}$ is the set of all functions mapping $\mathcal{X}$ to $\Delta_{\mathcal{Y}}$ and ℓ is the 0-1 loss.

is $[0, M]$, *then for any two distributions* $\mathcal{D}_{tr}$ *and* $\mathcal{D}_{te}$ *on* $\mathcal{X}$ *with probability density function* p_{tr} *and* p_{te} *respectively,* $\mathcal{M}_{\text{cov}}(\mathcal{D}_{tr}, \mathcal{D}_{te}; \mathcal{H}, \ell) = \frac{M}{2}\ell_1(\mathcal{D}_{tr}, \mathcal{D}_{te}) = \frac{M}{2}\int_{\mathcal{X}} |p_{tr}(x) - p_{te}(x)| \, \mathrm{d}x$.

It is clear that the covariate shift metric $\mathcal{M}_{\text{cov}}$ is determined by the accumulated bias between the distribution $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$ defined on $\mathcal{X}$ and without contribution from $\mathcal{Y}$, which meets the definition of covariate shift.

The concept shift term $\mathcal{M}_{\text{cpt}}^{\text{min}}$ and $\mathcal{M}_{\text{cpt}}^{\text{max}}$. When ℓ is set as the 0-1 loss, *i.e.*, the loss $\ell(f_{\text{tr}}(x), f_{\text{te}}(x))$ is 0 if and only if $f_{\text{tr}}(x) = f_{\text{te}}(x)$, $\mathcal{M}_{\text{cpt}}^{\text{min}}$ and $\mathcal{M}_{\text{cpt}}^{\text{max}}$ can be written as $\mathcal{M}_{\text{cpt}}^{\text{min}}/\mathcal{M}_{\text{cpt}}^{\text{max}} = \min/\max\{\int_{\mathcal{X}} \mathbb{I}[f_{\text{tr}}(x) \neq f_{\text{te}}(x)] p_{\text{tr}}(x) \mathrm{d}x, \int_{\mathcal{X}} \mathbb{I}[f_{\text{tr}}(x) \neq f_{\text{te}}(x)] p_{\text{te}}(x) \mathrm{d}x\}$. Here $\mathbb{I}[f_{\text{tr}}(x) \neq f_{\text{te}}(x)]$ is an indicator function on whether $f_{\text{tr}}(x) \neq f_{\text{te}}(x)$.

Intuitively, the two terms in the $\min/\max$ functions represent the probabilities of inconsistent labeling functions in training and test domains. $\mathcal{M}_{\text{cpt}}^{\text{min}}$ and $\mathcal{M}_{\text{cpt}}^{\text{max}}$ further take the minimal and maximal value of the two probabilities, respectively. It is rational that the concept shift is actually the integral of $p_{\text{tr}}(x)$ (or $p_{\text{te}}(x)$) over any points x where its corresponding label on training data differs from that on test data. In practice, we estimate f_{tr} and f_{te} with models trained on source domains and target domains, respectively. More discussion and comparison of discrepancy distance and other metrics for distribution distance are in Appendix A.

4.2. Dataset Evaluation with the Metrics

To use the covariate shift metric $\mathcal{M}_{\text{cov}}$ and concept shift metrics $\mathcal{M}_{\text{cpt}}^{\text{min}}, \mathcal{M}_{\text{cpt}}^{\text{max}}$ for dataset evaluation, we show that larger covariate shift and smaller concept shift favors a discriminative domain generalization benchmark. Intuitively, the critical point of datasets for domain generalization lies in 1) significant covariate shift between domains that drives generalization challenging [56] and 2) common knowledge about categories across domains on which models can rely on to conduct valid predictions on unseen domains [34, 93]. The common knowledge requires the alignment between labeling functions of source domains and target domains, *i.e.*, a moderate concept shift. When there is a strong inconsistency between labeling rules on training and test data, the classification loss instructing biased connections between visual features and concepts is misleading for generalization to test data. Thus models can hardly learn strong predictors for test data without knowledge of test domains.

To analyze the intuitions theoretically, we first propose an upper bound for the expected loss in the test domain for any hypothesis $h \in \mathcal{H}$.

Theorem 4.2. *Suppose the loss function ℓ is symmetric and obeys the triangle inequality. Suppose $f_{tr}, f_{te} \in \mathcal{H}$. Then for any hypothesis $h \in \mathcal{H}$, the following holds*

$$\varepsilon_{te}(h) \leq \varepsilon_{tr}(h) + \mathcal{M}_{\text{cov}}(\mathcal{D}_{tr}, \mathcal{D}_{te}; \mathcal{H}, \ell) + \mathcal{M}_{\text{cpt}}^{\text{min}}(\mathcal{D}_{tr}, \mathcal{D}_{te}, f_{tr}, f_{te}; \ell). \tag{3}$$

Remark. Theorem 4.2 is closely related to generalization bounds in domain adaptation (DA) literature [5, 89, 91, 93]. In detail, [5] first studied the generalization bound from a source domain to a target domain in binary classification problems and [89, 91] further extended the results to multi-class classification problems. However, the bounds in their results depend on a specific term $\lambda^* \triangleq \min_{h \in \mathcal{H}} \varepsilon_{\text{tr}}(h) + \varepsilon_{\text{te}}(h)$, which is conservative and relatively loose and can not be measured as concept shift directly [93]. As a result, [93] developed a bound which explicitly takes concept shift (termed as conditional shift by them) into account. However, their results are only applied to binary classifications and ℓ_1 loss function. By contrast, Theorem 4.2 can be applied to multi-class classifications problems and any loss functions that are symmetric and obeys the triangle inequality.

Theorem 4.2 quantitatively gives an estimation of the biggest gap between the performance of a model on training and test data. If we consider $\mathcal{H}$ as a set of deep models trained on training data with different learning strategies, the estimation indicates the upper bound of the range in which their performance varies. If we consider h as a model that fits training data, the bound gives an estimation of how much the distribution shift of the dataset contributes to the performance drop between training and test data.

Furthermore, we propose a lower bound for the expected loss in the test domain for any hypothesis $h \in \mathcal{H}$ to better understand how the proposed metrics affect the discrimination ability of datasets.

Theorem 4.3. *Suppose the loss function ℓ is symmetric and obeys the triangle inequality. Suppose $f_{tr}, f_{te} \in \mathcal{H}$. Then for any hypothesis $h \in \mathcal{H}$, the following holds*

$$\varepsilon_{te}(h) \geq \mathcal{M}_{\text{cpt}}^{\text{max}}(\mathcal{D}_{tr}, \mathcal{D}_{te}, f_{tr}, f_{te}; \ell) - \mathcal{M}_{\text{cov}}(\mathcal{D}_{tr}, \mathcal{D}_{te}; \mathcal{H}, \ell) - \varepsilon_{tr}(h). \tag{4}$$

As shown in Theorem 4.3, for any hypothesis $h \in \mathcal{H}$, the term $(\mathcal{M}_{\text{cpt}} - \mathcal{M}_{\text{cov}})$ determines the lower bound of the test loss and further determines the upper bound of the test performance of h. The bound is critical to evaluate a dataset since the performance of any well-trained model on test data is upper bounded by the properties (concept shift and covariate shift) of the dataset, disregarding how the model is designed or learned. Specifically, consider the stop training condition of a any possible model h is that the loss on the training data is smaller than γ, which is rational with most of current training strategies, the performance of the model on test data is upper bounded by $\gamma - \mathcal{M}_{\text{cpt}} + \mathcal{M}_{\text{cov}}$, which is irrelevant to the choice of h and the learning protocol. Intuitively, when the discrepancy between labeling functions between training and test data, the better the model fits training data, the worse it generalizes to test domains. Conversely, with more aligned labeling functions, the common knowledge between training and test data is richer and more instructive, so that the ceiling of generalization is higher. Moreover, the covariate shift $\mathcal{M}_{\text{cov}}$ contributes positively to the upper bound of the test performance, given that the concept shift $\mathcal{M}_{\text{cpt}}$ can be considered as integral of probability density $p_{\text{tr}}(x)$ (or $p_{\text{te}}(x)$) over points with unaligned labeling functions, where the covariate shift $\mathcal{M}_{\text{cov}}$ helps to counteract the impact of labeling mismatch.

As a result, the drop given by Theorem 4.3 is unsolvable for algorithms but modifiable by suppressing the concept shift or enhancing the covariate shift. To better evaluate generalization ability, an DG benchmark requires small concept shift and large covariate shift.

4.3. Empirical Evaluation

We compare NICO^{++} with current DG datasets in both covariate shift and concept shift. Please see details of the implementation in Appendix B.

Results are shown in Table 1. Concept shift on NICO^{++} is significantly lower than other datasets, indicating more aligned labeling rules across domains and more instructive common knowledge of categories can be learned by models. The covariate shifts of NICO^{++}, PACS, and DomainNet are comparable, which demonstrates that the distribution shift on images caused by the background can be as strong as style shifts. It is worth noticing that the term $\mathcal{M}_{\text{cpt}} - \mathcal{M}_{\text{cov}}$ in Theorem 4.3 is larger than 0 on current DG datasets while lower than 0 on NICO^{++}, indicating that the drop caused by a shift of labeling function across domains is significant enough to damage the upper generalization bound while the common knowledge across domains in NICO^{++} is sufficient for models to approach the oracle performance.

5. Experiments

Inspired by [87], we present two evaluation settings, namely *classic domain generalization* and *flexible domain generalization*, and perform extensive experiments on both settings. We design experimental settings to evaluate current DG methods and illustrate how NICO^{++} contributes to filling in the evaluation on generalization to multiple unseen domains. Due to space limitations, we only report major results, and more experimental details are in Appendix D.

5.1. Evaluation Metrics for Algorithms

Despite the fact that the widely adopted evaluation methods in DG effectively show the generalization ability of models to the unseen target domain, they fail to sufficiently simulate real application scenarios. For example, the most popular evaluation method, namely leave-one-out evaluation [40, 64], tests models on a single target domain for each training process, while in real applications, a trained model is required to be reliable under any possible scenarios with various data distributions. The compromise on the limitation of domain numbers in current benchmarks, including PACS, VLCS, DomainNet, Office-Home, can be addressed by NICO^{++} with sufficient aligned and unique domains. The superiority supports designing more realistic evaluation metrics to evaluate generalizability comprehensively.

We consider three simple metrics to evaluate DG algorithm, namely average accuracy, overall accuracy, and the standard deviation of accuracy across domains. The metrics are defined as follows.

$$\begin{aligned}&\text{Average} = \frac{1}{K}\sum_{k=1}^{K}\text{acc}_k, \text{Overall} = \frac{1}{\sum_{k=1}^{K} N_k}\sum_{k=1}^{K} N_k\text{acc}_k,\\ &\text{Std} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\text{acc}_k - \text{Average})^2}.\end{aligned} \tag{5}$$

Here K is the number of domains in the test data, N_k is the number of samples in the k-th domain, and acc_k is the prediction accuracy in the k-th domain. The metric Average is widely used in DG literature, where both training and test domains for different categories are aligned. The metric Overall is more reasonable when the domains can be various for different categories or the test data are a mixture of unknown domains, and thus the accuracy for each domain is incalculable. The metric Std indicates the standard deviation of the performance across different domains. Since learning models that are consistently reliable in any possible environment is the target of DG and many methods are designed to learn invariant representations [22], Std is rational and instructive. Please note that Std is insignificant in the leave-one-out evaluation method where models tested on different target domains are trained on different combinations of source domains, while domains of NICO^{++} are rich enough to evaluate models on various target domains with fixed source domains.

5.2. Benchmark for Standard DG

The common domains in NICO^{++} are rich and consistent for all categories, which supports multiple test domains

Table 1. Results of estimated covariate shift and concept shift of NICO^{++} and current DG datasets. $\uparrow$ donates that the higher the metric is, the better and $\downarrow$ is the opposite. The best results of all datasets are highlighted in bold font.

	I.I.D.	PACS	DomainNet	VLCS	Office-Home	MNIST-M	NICO^{++}
$\mathcal{M}_{\text{cov}} \uparrow$	0	0.325(±0.053)	0.302(±0.039)	0.256(±0.041)	0.238(±0.049)	0.225(±0.034)	**0.338**(±0.031)
$\mathcal{M}_{\text{cpt}}^{\text{min}} \downarrow$	0	0.434(±0.023)	0.247(±0.055)	0.303(±0.064)	0.353(±0.086)	0.243(±0.048)	**0.152**(±0.034)
$\mathcal{M}_{\text{cpt}}^{\text{max}} \downarrow$	0	0.537(±0.054)	0.612(±0.057)	0.523(±0.044)	0.505(±0.084)	0.449(±0.030)	**0.192**(±0.040)

Table 2. Results of the DG setting on NICO^{++}. Oracle donates the model trained with data sampled from the target distribution (yet none of the test images is seen in the training). Ova. and Avg. indicate the overall accuracy of all the test data and the arithmetic mean of the accuracy of 6 domains, respectively. They are different because the capacities of different domains are not equal. The reported results are the average over three repetitions of each run. The best results are highlighted with bold font and the second best with underline.

Method	Training: Di, G, O, Wa		Training: A, R, O, Wa		Training: A, R, Di, G				
	A	R	Di	G	O	Wa	Ova.	Avg.	Std
ERM	81.89	79.76	72.42	82.31	76.80	71.01	77.08	77.36	4.39
SWAD [11]	82.98	81.21	74.59	83.50	78.43	72.81	78.65	78.92	4.06
MMLD [48]	80.62	79.63	73.17	81.24	78.08	71.23	77.09	77.33	**3.80**
RSC [33]	81.26	79.99	71.91	81.67	76.51	70.78	76.73	77.02	4.35
AdaClust [70]	79.25	78.93	71.41	81.48	74.23	70.13	75.71	75.91	4.24
SagNet [52]	**83.12**	81.17	73.72	83.42	78.43	73.03	78.56	78.81	4.18
EoA [3]	82.88	**81.86**	**75.83**	83.29	**78.63**	72.80	**78.88**	**79.22**	3.87
MixStyle [96]	75.83	73.51	65.89	76.69	70.51	63.41	70.66	70.97	4.93
MLDG [41]	82.24	80.57	72.24	**84.14**	77.19	71.33	77.76	77.95	4.84
MMD [43]	81.73	79.26	72.33	82.57	77.24	70.90	77.11	77.34	4.41
CORAL [68]	82.89	80.69	73.77	82.90	78.26	**73.21**	78.38	78.62	3.95
StableNet [87]	82.82	80.30	74.05	83.52	76.91	72.34	78.06	78.32	4.23
FACT [79]	81.55	81.03	74.32	82.16	78.07	71.30	77.74	78.07	4.03
JiGen [9]	82.64	80.36	74.15	83.29	77.14	71.59	77.89	78.19	4.31
GroupDRO [60]	81.81	79.69	72.37	82.11	77.28	71.72	77.26	77.50	4.17
DDG [85]	82.53	79.68	72.42	83.03	77.91	71.86	77.70	77.90	4.42
DNA [12]	82.24	80.62	72.07	82.56	78.00	71.39	77.54	77.81	4.55
Fishr [57]	81.98	79.38	72.62	82.37	77.61	70.91	77.22	77.48	4.37
IRM [2]	81.66	79.82	72.58	82.46	76.83	70.92	77.11	77.38	4.38
Mixup [80, 84]	81.84	80.38	74.02	82.62	78.20	72.36	78.01	78.24	3.85
Oracle	91.18	89.98	89.29	90.27	88.55	86.23	88.99	89.25	1.58

evaluation for domain generalization, as discussed in Section 1. In this section, we give the official split of domains for the standard domain generalization. Currently, 6 out of 10 common domains are publicly available and we select two of them as test domains while others as training domains for each evaluation. We run 3 individual evaluations and cover all 6 domains as test domains. Specifically, in the first evaluation, we select domains [*Autumn*, *Rock*] as test domains and others as training domains. We select domains [*Dim*, *Grass*] and [*Outdoor*, *Water*] as test domains for the second and third evaluations, respectively[‡]. The results of current representative methods with ResNet-50 as the backbone are shown in Table 2. Models generally show better generalization when tested on a single cluster of common domains than the opposite, indicating that generalization to diverse unseen domains is more challenging. Current SOTA methods such as EoA, CORAL, and StableNet show their effectiveness, yet a significant gap between them and oracle shows that the room for improvement is spacious. More splits and implementation details are in Appendix D.

[‡]The official splits (i.e., training and test data) of each domain are given in https://github.com/xxgege/NICO-plus.

5.3. Benchmark for Flexible DG

Compared current DG setting where domains are aligned across categories, a flexible combination of categories and domains in both training and test data can be more realistic and challenging [64, 87]. In such cases, the level of the distribution shifts varies in different classes, requiring a strong ability of generalization to tell common knowledge of categories from various domains. We present two settings, namely *random* and *compositional*. We randomly select two domains out of common domains as dominant ones, 12 out of the remaining domains as minor ones and the other 6 domains as test data for each category for the *random* setting. There can be spurious correlations between domains and labels since a domain can be with class A in training data and class B in test data. For the *compositional* setting, 4 domains are chosen as exclusive training domains and others as sharing domains. Then 2 domains are randomly selected from exclusive training domains as the majority, 12 from sharing domains as the minority, and the remaining 4 in sharing domains for the test. Thus there are no spurious correlations between dominant domains and labels. We select all images from the dominant domains and

Table 3. Results of the flexible DG setting on NICO^{++}.

Method	ERM	SWAD	MMLD	RSC	AdaClust	SagNet	EoA	MixStyle	StableNet	FACT	JiGen	Oracle
Rand.	74.19	75.62	73.25	75.20	73.39	72.79	<u>76.22</u>	73.47	**77.37**	75.34	75.44	84.60
Comp.	78.01	76.97	76.85	75.76	76.64	76.15	**79.62**	77.01	78.19	<u>79.39</u>	78.77	86.18
Avg.	76.10	76.30	75.05	75.48	75.02	74.47	**77.92**	75.24	<u>77.78</u>	77.37	77.11	85.39

Table 4. Standard deviation across epochs and seeds on different datasets.

	PACS			DomainNet			VLCS			OfficeHome			NICO^{++}		
Method	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap
ERM	0.96	0.82	2.66	0.61	0.57	0.46	0.83	0.58	3.59	0.77	0.59	0.81	**0.22**	**0.10**	**0.39**
SWAD	0.41	0.76	1.61	0.35	0.30	0.39	0.74	0.49	0.58	0.31	0.25	0.30	**0.07**	**0.05**	**0.06**
MMLD	1.68	2.02	3.25	1.03	0.50	0.85	2.33	1.12	3.97	1.25	0.47	0.56	**0.25**	**0.10**	**0.15**
RSC	0.76	0.81	0.93	0.55	0.35	0.56	1.02	0.61	0.80	0.85	0.37	0.89	**0.18**	**0.05**	**0.10**
AdaClust	1.06	1.74	1.54	0.98	0.41	0.72	1.32	1.79	1.34	1.36	1.30	0.28	**0.22**	**0.04**	**0.13**
SagNet	0.74	2.44	2.78	0.92	**0.23**	0.54	0.94	1.74	4.19	0.80	0.30	**0.44**	**0.11**	0.31	0.61
EoA	0.11	0.36	0.18	0.22	0.16	**0.02**	0.15	0.45	0.21	0.05	0.29	0.08	**0.02**	**0.04**	0.13
MixStyle	1.53	0.63	1.69	0.60	0.36	0.42	1.27	1.78	3.40	0.72	0.43	0.56	**0.17**	**0.16**	**0.00**
MLDG	0.82	1.02	1.24	0.53	0.25	0.55	1.15	1.01	4.14	1.03	0.09	0.23	**0.10**	**0.08**	**0.12**
MMD	1.13	2.39	0.66	0.82	0.24	0.50	1.98	1.32	3.72	0.61	**0.02**	**1.34**	**0.11**	0.11	**0.16**
CORAL	1.09	1.02	1.18	0.52	0.48	0.47	0.77	0.94	3.18	0.49	0.28	0.50	**0.06**	**0.17**	**0.19**
StableNet	0.90	1.25	1.03	0.34	0.71	0.82	0.86	0.69	0.88	0.44	0.21	0.48	**0.09**	**0.05**	**0.09**
FACT	0.31	0.46	0.52	0.14	**0.16**	**0.37**	0.64	0.85	1.17	0.21	0.27	0.68	**0.06**	0.19	1.09
JiGen	0.33	1.15	0.70	0.16	0.18	0.39	0.51	0.67	1.30	0.20	0.69	0.25	**0.05**	**0.09**	**0.10**
GroupDRO	1.27	0.96	2.09	0.96	0.37	0.54	1.18	0.85	4.93	0.63	0.47	0.55	**0.16**	**0.10**	**0.16**
IRM	3.77	3.02	4.14	2.17	0.89	0.00	6.00	1.74	5.77	2.10	1.59	0.00	**0.90**	**0.54**	**0.00**

50 images from each minor domain for training and 50 images from each test domain for testing. Results are shown in Table 3. Current SOTA algorithms outperform ERM by a noticeable margin, yet the gap to Oracle remains significant. More splits and discussions are in Appendix D.

5.4. Test Variance and Model Selection

Model selection (including the choice of hyperparameters, training checkpoints, and architecture variants) affects DG evaluation considerably [3, 27]. The leak of knowledge of test data in training or model selection phase is criticized yet still usual in current algorithms [3, 27]. This issue is exacerbated by the variance of test performance across random seeds, training iterations and other hyperparameters in that one can choose the best seed or the model from the best epoch under the guidance of the released oracle validation set for a noticeable improvement. NICO^{++} presents a feasible approach by reducing the test variance and thus decreasing the possible improvement by leveraging the leak.

As shown in Section 4, the gap between the performance of a model on training and test data is bounded by the sum of covariant shift and concept shift between source and target domains. Intuitively, test variance on NICO^{++} is lower than other current DG datasets given that NICO^{++} guarantees a significantly lower concept shift. Strong concept shift between source domains introduces confusing mapping relations between input X and output Y, harming the convergence and enlarging the variance. Since most current deep models are optimized by stochastic gradient descent (SGD), the test accuracy is prone to jitter as the input sequence determined by random seeds varies. Moreover, concept shift also grows the mismatch between the performance on validation data and test data, further widening the gap between target-guided and source-guided model selection.

Empirically, we compare the test variance and the improvement of leveraging oracle knowledge on NICO^{++} with other datasets across various seeds and training epochs in Table 4. For the test variance across random seeds, we train 3 models for each method with 3 random seeds and calculate the test variance among them. For the test variance across epochs, we calculate the test variance of the models saved on the last 10 epochs for each random seed and show the mean value of 3 random seeds. NICO^{++} shows a lower test variance compared with other datasets across both various random seeds and training epochs, indicating a more stable estimation of generalization ability robust to the choice of algorithm-irrelevant hyperparameters. As a result, NICO^{++} alleviates the oracle leaking issue by significantly squeezing the possible improvement space, leading to a fairer comparison for DG methods.

6. Conclusion

In this paper, we propose a context-extensive large-scale benchmark named NICO^{++} along with more rational evaluation methods for comprehensively evaluating DG methods. Two metrics on covariate shift and concept shift are proposed to evaluate DG datasets upon two novel generalization bounds. Extensive experiments showed the superiority of NICO^{++} over current datasets and benchmarked DG algorithms comprehensively.

Acknowledgement This work is supported partly by National Key R&D Program of China (No.2018AAA0102004, 2020AAA0106300), National Natural Science Foundation of China (No.U1936219, 62141607), Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 1

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 7

[3] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint arXiv:2110.10832*, 2021. 2, 7, 8

[4] 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. 2

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 2, 5

[6] Daniel S Berman, Anna L Buczak, Jeffrey S Chavis, and Cherita L Corbett. A survey of deep learning methods for cyber security. *Information*, 10(4):122, 2019. 1

[7] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017. 2

[8] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *NeurIPS*, 24:2178–2186, 2011. 2

[9] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2, 7

[10] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 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, 2021. 7

[12] Xu Chu, Yujie Jin, Wenwu Zhu, Yasha Wang, Xin Wang, Shanghang Zhang, and Hong Mei. Dna: Domain generalization with diversified neural averaging. In *International Conference on Machine Learning*, pages 4010–4034. PMLR, 2022. 7

[13] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 1

[14] 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. 1, 2

[15] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017. 2

[16] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020. 4

[17] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2

[18] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 1, 2

[19] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 33:11996–12007, 2020. 2

[20] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, 2016. 2

[21] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 6

[23] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017. 2

[24] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE TPAMI*, 39(7):1414–1430, 2016. 2

[25] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2

[26] Thomas Grubinger, Adriana Birlutiu, Holger Schöner, Thomas Natschläger, and Tom Heskes. Domain generalization based on transfer component analysis. In *International Work-Conference on Artificial Neural Networks*, pages 325–334. Springer, 2015. 2

[27] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 2, 8

[28] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021. 2

[29] 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. 2

[30] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2

[31] 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

[32] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020. 2

[33] 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, 2, 7

[34] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 1, 5

[35] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Style normalization and restitution for domaingeneralization and adaptation. *arXiv preprint arXiv:2101.00588*, 2021. 2

[36] Rawal Khirodkar, Donghyun Yoo, and Kris Kitani. Domain randomization for scene-specific car detection and pose estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1932–1940. IEEE, 2019. 2

[37] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[38] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2

[39] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011. 1

[40] 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. 1, 2, 6

[41] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 7

[42] 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

[43] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 7

[44] Yixiao Liao, Ruyi Huang, Jipu Li, Zhuyun Chen, and Weihua Li. Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed. *IEEE Transactions on Instrumentation and Measurement*, 69(10):8064–8075, 2020. 2

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[46] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE, 2018. 2

[47] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009. 4

[48] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756, 2020. 7

[49] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018. 1

[50] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18. PMLR, 2013. 2

[51] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. *arXiv preprint arXiv:1805.07925*, 2018. 2

[52] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 7

[53] 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. 1, 2

[54] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018. 2

[55] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE, 2019. 2

[56] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008. 5

[57] 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. 7

[58] Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021. 4

[59] Jongbin Ryu, Gitaek Kwon, Ming-Hsuan Yang, and Jongwoo Lim. Generalized convolutional forest networks for domain generalization and visual recognition. In *ICLR*, 2019. 2

[60] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2, 7

[61] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv preprint arXiv:2011.12672*, 2020. 2

[62] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. *Advances in neural information processing systems*, 29, 2016. 2

[63] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 2

[64] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 2, 3, 4, 6, 7

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[66] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007. 2

[67] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007. 2

[68] 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

[69] Jafar Tahmoresnezhad and Sattar Hashemi. Visual domain adaptation via transfer feature learning. *Knowledge and information systems*, 50(2):585–605, 2017. 2

[70] Xavier Thomas, Dhruv Mahajan, Alex Pentland, and Abhimanyu Dubey. Adaptive methods for aggregated domain generalization. *arXiv preprint arXiv:2112.04766*, 2021. 7

[71] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 2

[72] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPR workshops*, pages 969–977, 2018. 2

[73] 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. 1, 2

[74] 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, 2

[75] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021. 1, 2

[76] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE TMI*, 39(12):4237–4248, 2020. 2

[77] Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. Gradual fine-tuning for low-resource domain adaptation. *arXiv preprint arXiv:2103.02205*, 2021. 2

[78] Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Kenichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020. 1

[79] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 7

[80] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 7

[81] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Oodbench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021. 1

[82] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018. 1

[83] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, pages 2100–2110, 2019. 2

[84] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 7

[85] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8024–8034, 2022. 7

[86] Lei Zhang, Wangmeng Zuo, and David Zhang. Lsdt: Latent sparse domain transfer learning for visual adaptation. *IEEE Transactions on Image Processing*, 25(3):1177–1191, 2016. 2

[87] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan 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. 6, 7

[88] Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Domain-irrelevant representation learning for unsupervised domain generalization. *arXiv preprint arXiv:2107.06219*, 2021. 2

[89] Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[90] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei 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*, pages 8035–8045, 2022. 2

[91] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 5

[92] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *arXiv preprint arXiv:2111.14341*, 2021. 2

[93] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019. 1, 5

[94] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv e-prints*, pages arXiv–2103, 2021. 1

[95] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 2

[96] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 2, 7