Generalizing to Unseen Domains: A Survey on Domain Generalization

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, *Senior Member, IEEE*, Wang Lu, Yiqiang Chen, *Senior Member, IEEE*, Wenjun Zeng, *Fellow, IEEE*, Philip S. Yu, *Fellow, IEEE*

Abstract—Machine learning systems generally assume that the training and testing distributions are the same. To this end, a key requirement is to develop models that can generalize to unseen distributions. Domain generalization (DG), *i.e.*, out-of-distribution generalization, has attracted increasing interests in recent years. Domain generalization deals with a challenging setting where one or several different but related domain(s) are given, and the goal is to learn a model that can generalize to an *unseen* test domain. Great progress has been made in the area of domain generalization for years. This paper presents the first review of recent advances in this area. First, we provide a formal definition of domain generalization and discuss several related fields. We then thoroughly review the theories related to domain generalization and carefully analyze the theory behind generalization. We categorize recent algorithms into three classes: data manipulation, representation learning, and learning strategy, and present several popular algorithms in detail for each category. Third, we introduce the commonly used datasets, applications, and our open-sourced codebase for fair evaluation. Finally, we summarize existing literature and present some potential research topics for the future.

Index Terms—Domain generalization, Domain adaptation, Transfer learning, Out-of-distribution generalization

1 Introduction

Machine learning (ML) has achieved remarkable success in various areas, such as computer vision, natural language processing, and healthcare. The goal of ML is to design a model that can learn general and predictive knowledge from training data, and then apply the model to new (test) data. Traditional ML models are trained based on the *i.i.d.* assumption that training and testing data are identically and independently distributed. However, this assumption does not always hold in reality. When the probability distributions of training data and testing data are different, the performance of ML models often deteriorates due to domain distribution gaps [1]. Collecting the data of all possible domains to train ML models is expensive and even prohibitively impossible. Therefore, enhancing the *generalization* ability of ML models is important in both industry and academic fields.

There are many generalization-related research topics such as domain adaptation, meta-learning, transfer learning, covariate shift, and so on. In recent years, *Domain generalization (DG)* has received much attention. As shown in Fig. 1, the goal of domain generalization is to learn a model from one or several different but related domains (*i.e.*, diverse training datasets) that will generalize well on *unseen* testing domains. For instance, given a training set consisting of images coming from sketches, cartoon images and paintings, domain generalization requires to train a good machine learning model that has minimum prediction error in classifying images coming from natural images or photos, which are clearly having distinguished distributions from the images in training set. Over the past years, domain generalization has made significant progress in various areas such as computer vision and natural language processing. Despite the progress, there has not been a survey in this area that comprehensively introduces and summarizes its main ideas, learning algorithms and other related problems to provide research insights for the future.

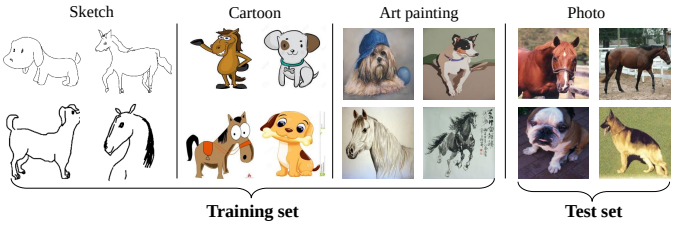


Fig. 1. Examples from the dataset PACS [2] for domain generalization. The training set is composed of images belonging to domains of sketch, cartoon, and art paintings. DG aims to learn a generalized model that performs well on the unseen target domain of photos.

In this paper, we present the first survey on domain generalization to introduce its recent advances, with special focus on its formulations, theories, algorithms, research areas, datasets, applications, and future research directions. We hope that this survey can provide a comprehensive review for interested researchers and inspire more research in this and related areas.

There are several survey papers after the conference version of our paper, and they are significantly different from ours. Concurrently, Zhou et al. [3] also wrote a survey on DG, while their focus is in computer vision field. A more

- *J. Wang, C. Lan, C. Liu, W. Zeng, and T. Qin are with Microsoft Research Asia, Beijing, China. Correspondence to: Jindong Wang. E-mail: {jindong.wang,culan,changliu,wezeng,taoqin}@microsoft.com.*
- *Y. Ouyang is with School of Data Science, Chinese University of Hong Kong, Shenzhen. Email: yidongouyang@link.cuhk.edu.cn.*
- *W. Lu and Y. Chen are with Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.*
- *P. Yu is with University of Illinois at Chicago, and Institute for Data Science, Tsinghua University. Email: psyu@uic.edu.*

recent survey paper is on out-of-distribution (OOD) generalization by Shen et al. [4]. Their work focused on causality and stable neural networks. A related survey paper [5] is for OOD detection instead of building a working algorithm that can be applied to any unseen environments.

This paper is a heavily extended version of our previously accepted short paper at IJCAI-21 survey track (6 pages, included in the appendix file). Compared to the short paper, this version makes the following extensions:

- We present the theory analysis on domain generalization and the related domain adaptation.
- We substantially extend the methodology by adding new categories: *e.g.*, causality-inspired methods, generative modeling for feature disentanglement, invariant risk minimization, gradient operation-based methods, and other learning strategies to comprehensively summarize these DG methods.
- For all the categories, we broaden the analysis of methods by including more related algorithms, comparisons, and discussion. And we also include more recent papers (over 30% of new work).
- We extend the scope of datasets and applications, and we also explore evaluation standards to domain generalization. Finally, we build an open-sourced codebase for DG research named *DeepDG*[1] and conduct some analysis of the results on public datasets.

This paper is organized as follows. We formulate the problem of domain generalization and discuss its relationship with existing research areas in Section 2. Section 3 presents the related theories in domain generalization. In Section 4, we describe some representative DG methods in detail. In Section 5, we show some new DG research areas extended from the traditional setting. Section 6 presents the applications and Section 7 introduces the benchmark datasets for DG. We summarize the insights from existing work and present some possible future directions in Section 8. Finally, we conclude this paper in Section 9.

2 Background

2.1 Formalization of Domain Generalization

In this section, we introduce the notations and definitions used in this paper.

Definition 1 (Domain)**.** *Let X denote a nonempty input space and Y an output space. A domain is composed of data that are sampled from a distribution. We denote it as $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ P_{XY}, where $\mathbf{x} \in X \subseteq \mathbb{R}^d$, $y \in Y \subseteq \mathbb{R}$ denotes the label, and P_{XY} denotes the joint distribution of the input sample and output label. X and Y denote the corresponding random variables.*

Definition 2 (Domain generalization)**.** *As shown in Fig. 2, in domain generalization, we are given M training (source) domains $S_{train} = \{S^i \mid i = 1, \cdots, M\}$ where $S^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i-th domain. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq M$. The goal of domain generalization is to learn a robust and generalizable predictive function $h : X \to Y$ from the M training domains to achieve a minimum prediction error on an* unseen *test domain S_{test} (i.e., S_{test} cannot be accessed in training and $P_{XY}^{test} \neq P_{XY}^i$ for $i \in \{1, \cdots, M\}$):*

$$\min_h \mathbb{E}_{(\mathbf{x},y)\in S_{test}}[\ell(h(\mathbf{x}), y)], \tag{1}$$

where $\mathbb{E}$ is the expectation and $\ell(\cdot, \cdot)$ is the loss function.

We list the frequently used notations in TABLE 1.

TABLE 1
Notations used in this paper.

Notation	Description	Notation	Description
$\mathbf{x}; y$	Instance/label	$\ell(\cdot,\cdot)$	Loss function
$\mathcal{X}; \mathcal{Y}$	Feature/label space	h	Predictive function
$\mathcal{S}$	Domain	$g; f$	Feature extractor/classifier
$P(\cdot)$	Distribution		Error (risk)
$\mathbb{E}[\cdot]$	Expectation		Model parameter
M	Number of source domain	n_i	Data size of source domain i

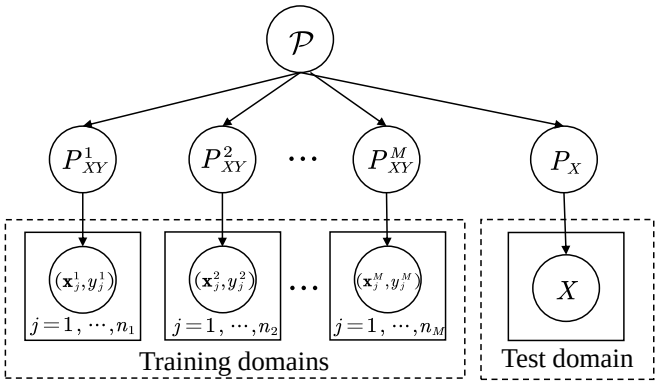


Fig. 2. Illustration of domain generalization. Adapted from [6].

2.2 Related Research Areas

There are several research fields closely related to domain generalization, including but not limited to: transfer learning, domain adaptation, multi-task learning, multiple domain learning, meta-learning, lifelong learning, and zero-shot learning. We summarize their differences with domain generalization in TABLE 2 and briefly describe them in the following.

Multi-task learning [7] jointly optimizes models on several related tasks. By sharing representations between these tasks, we could enable the model to generalize better on the original task. Note that multi-task learning does not aim to enhance the generalization to a new (unseen) task. Particularly, multi-domain learning is a kind of multi-task learning, which trains on multiple related domains to learn good models for each original domain [8] instead of new test domains.

Transfer learning [9, 10, 11] trains a model on a source task and aims to enhance the performance of the model on a different but related target domain/task. Pretraining-finetuning is the commonly used strategy for transfer learning where the source and target domains have different tasks and target domain is accessed in training. In DG, the target domain cannot be accessed and the training and test tasks are often the same while they have different distributions.

1. https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG

TABLE 2
Comparison between domain generalization and some related learning paradigms.

Learning paradigm	Training data	Test data	Condition	Test access
Multi-task learning	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{Y}^i \neq \mathcal{Y}^j; 1 \leq i \neq j \leq n$	✓
Transfer learning	$\mathcal{S}^{src}, \mathcal{S}^{tar}$	$\mathcal{S}^{tar}$	$\mathcal{Y}^{src} \neq \mathcal{Y}^{tar}$	✓
Domain adaptation	$\mathcal{S}^{src}; \mathcal{S}^{tar}$	$\mathcal{S}^{tar}$	$P(\mathcal{X}^{src}) \neq P(\mathcal{X}^{tar})$	✓
Meta-learning	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^{n+1}$	$\mathcal{Y}^i \neq \mathcal{Y}^j; 1 \leq i \neq j \leq n+1$	✓
Lifelong learning	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^i$ arrives sequentially	✓
Zero-shot learning	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^{n+1}$	$\mathcal{Y}^{n+1} \neq \mathcal{Y}^i; 1 \leq i \leq n$	×
Domain generalization	$\mathcal{S}^1;\cdots;\mathcal{S}^n$	$\mathcal{S}^{n+1}$	$P(\mathcal{S}^i) \neq P(\mathcal{S}^j); 1 \leq i \neq j \leq n+1$	×

Domain adaptation (DA) [12, 13] is also popular in recent years. DA aims to maximize the performance on a given target domain using existing training source domain(s). The difference between DA and DG is that DA has access to the target domain data while DG cannot see them during training. This makes DG more challenging than DA but more realistic and favorable in practical applications.

Meta-learning [14, 15, 16] aims to learn the learning algorithm itself by learning from previous experience or tasks, i.e., learning-to-learn. While the learning tasks are different in meta-learning, the learning tasks are the same in domain generalization. Meta-learning is a general learning strategy that can be used for DG [17, 18, 19, 20] by simulating the meta-train and meta-test tasks in training domains to enhance the performance of DG.

Lifelong Learning [21], or continual learning, cares about the learning ability among multiple sequential domains/tasks. It requires the model to continually learn over time by accommodating new knowledge while retaining previously learned experiences. This is also different from DG since it can access the target domain in each time step, and it does not explicitly handle the different distributions across domains.

Zero-shot learning [22, 23] aims at learning models from seen classes and classify samples whose categories are unseen in training. In contrast, domain generalization in general studies the problem where training and testing data are from the same classes but with different distributions.

3 THEORY

In this section, we review some theories related to domain generalization. Since domain adaptation is closely related to DG, we begin with the theory of domain adaptation.

3.1 Domain Adaptation

For a binary classification problem, we denote the true labeling functions on the source domain as $h^s : X \to [0;1]$ [2] and that on the target domain as h^t. Let $h : X \to [0;1]$ be any classifier from a hypothesis space H. The classification difference on the source domain between two classifiers h and h' can then be measured by

$$\epsilon^s(h;h') = \mathbb{E}_{\mathbf{x} \sim P_X^s}[h(\mathbf{x}) \neq h'(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim P_X^s}[|h(\mathbf{x}) - h'(\mathbf{x})|]; \quad (2)$$

and similarly we can define ϵ^t when taking $\mathbf{x} \sim P_X^t$ in the expectation. We define $\epsilon^s(h) := \epsilon^s(h;h^s)$ and $\epsilon^t(h) := \epsilon^t(h;h^t)$ as the risk of classifier h on the source and target domains, respectively.

The goal of DG/DA is to minimize the target risk $\epsilon^t(h)$, but it is not accessible since we do not have any information on h^t. So people seek to bound the target risk $\epsilon^t(h)$ using the tractable source risk $\epsilon^s(h)$. Ben-David et al. [24] (Thm. 1) give a bound relating the two risks:

$$\epsilon^t(h) \leq \epsilon^s(h) + 2d_1(P_X^s;P_X^t) \quad (3)$$
$$+ \min_{P_X \in \{P_X^s;P_X^t\}} \mathbb{E}_{\mathbf{x} \sim P_X}[|h^s(\mathbf{x}) - h^t(\mathbf{x})|]; \quad (4)$$

where $d_1(P_X^s;P_X^t) := \sup_{A \in \mathscr{X}} |P_X^s[A] - P_X^t[A]|$ is the *total variation* between the two distributions, and $\mathscr{X}$ denotes the sigma-field on X. The second term on the r.h.s measures the difference of cross-domain distributions, and the third term represents the difference in the labeling functions (covariate shift is not *a priori* assumed).

However, the total variation is a strong distance (i.e., it tends to be very large) that may loosen the bound (4), and is hard to estimate using finite samples. To address this, Ben-David et al. [24] developed another bound ([24], Thm. 2; [25], Thm. 1):

$$\epsilon^t(h) \leq \epsilon^s(h) + d_{H\Delta H}(P_X^s;P_X^t) + \lambda_H; \quad (5)$$

where the $H\Delta H$-divergence is defined as $d_{H\Delta H}(P_X^s;P_X^t) := \sup_{h;h' \in H} |\epsilon^s(h;h') - \epsilon^t(h;h')|$, replacing the total variation d_1 to measure the distribution difference, and the ideal joint risk $\lambda_H := \inf_{h \in H}[\epsilon^s(h) + \epsilon^t(h)]$ measures the complexity of H for the prediction tasks on the two domains. $H\Delta H$-divergence has a better finite-sample guarantee, leading to a non-asymptotic bound:

Theorem 1 (Domain adaptation error bound (non-asymptotic) [24] (Thm. 2))**.** *Let d be the Vapnik–Chervonenkis (VC) dimension [26] of H, and U^s and U^t be unlabeled samples of size n from the two domains. Then for any $h \in H$ and $\delta \in (0;1)$, the following inequality holds with probability at least $1-\delta$:*

$$\epsilon^t(h) \leq \epsilon^s(h) + \hat{d}_{H\Delta H}(U^s;U^t) + \lambda_H \quad (6)$$
$$+ 4\sqrt{\frac{2d\log(2n) + \log(2/\delta)}{n}}; \quad (7)$$

where $\hat{d}_{H\Delta H}(U^s;U^t)$ is the estimate of $d_{H\Delta H}(P_X^s;P_X^t)$ on the two sets of finite data samples.

In the above bounds, the domain distribution difference $d(P_X^s;P_X^t)$ is not controllable, but one may learn a representation function $g : X \to Z$ that maps the original input data $\mathbf{x}$ to some representation space Z, so that the representation distributions of the two domains become

2. When the output is in $(0;1)$, it means the probability of $y = 1$.

closer. This direction of DA is thus called DA based on domain-invariant representation (DA-DIR). The theory of domain-invariant representations has since inspired many DA/DG methods, which can be seen in Section 4.2.

3.2 Domain Generalization

3.2.1 Average risk estimation error bound

The first line of domain generalization theory considers the case where the target domain is totally unknown (not even unsupervised data), and measures the average risk over all possible target domains. Assume that all possible target distributions follow an underlying hyper-distribution P on $(x;y)$ distributions: $P^t_{XY} \sim P$, and that the source distributions also follow the same hyper-distribution: $P^1_{XY}, \dots, P^M_{XY} \sim P$. For generalization to any possible target domain, the classifier to be learned in this case also includes the domain information P_X into its input, so prediction is in the form $y = h(P_X, x)$ on the domain with distribution P_{XY}. For such a classifier h, its average risk over all possible target domains is then given by:

$$E(h) := \mathbb{E}_{P_{XY} \sim P} \mathbb{E}_{(x,y) \sim P_{XY}} [\ell(h(P_X, x), y)], \tag{8}$$

where ℓ is a loss function on Y. Exactly evaluating the expectations is impossible, but we can estimate it using finite domains/distributions following P, and finite $(x;y)$ samples following each distribution. As we have assumed $P^1_{XY}, \dots, P^M_{XY} \sim P$, the source domains and supervised data could serve for this estimation:

$$\hat{E}(h) := \frac{1}{M} \sum_{i=1}^{M} \frac{1}{n^i} \sum_{j=1}^{n^i} \ell(h(U^i, x^i_j), y^i_j), \tag{9}$$

where we use the supervised dataset $U^i := \{(x^i_j, y^i_j) \in S^i\}$ from domain i as an empirical estimation for P^i_X.

The first problem to consider is how well such an estimate approximates the target $E(h)$. This can be measured by the largest difference between $E(h)$ and $\hat{E}(h)$ on some space of h. To our knowledge, this is first analyzed by Blanchard et al. [6], where the space of h is taken as a reproducing kernel Hilbert space (RKHS). However, different from common treatment, the classifier h here also depends on the distribution P_X, so the kernel defining the RKHS should be in the form $k((P^1_X, x_1), (P^2_X, x_2))$. Blanchard et al. [6] construct such a kernel using kernels k_X, k'_X on X and kernel on the RKHS $H_{k'_X}$ of kernel k'_X: $k((P^1_X, x_1), (P^2_X, x_2)) := (\Phi_{k'_X}(P^1_X), \Phi_{k'_X}(P^2_X)) k_X(x_1, x_2)$, where $\Phi_{k'_X}(P_X) := \mathbb{E}_{x \sim P_X}[k'_X(x, \cdot)] \in H_{k'_X}$ is the kernel embedding of distribution P_X via kernel k'. The result is given in the following theorem, which gives a bound on the largest average risk estimation error within an origin-centered closed ball $B_{H_k}(r)$ of radius r in the RKHS H_k of kernel k, in a slightly simplified case where $n^1 = \dots = n^M =: n$.

Theorem 2 (Average risk estimation error bound for binary classification [6]). Assume that the loss function ℓ is L_ℓ-Lipschitz in its first argument and is bounded by B_ℓ. Assume also that the kernels k_X, k'_X and are bounded by $B^2_k, B^2_{k'} \le 1$ and B^2, respectively, and the canonical feature map $\Phi : v \in H_{k'_X} \mapsto \Phi(v, \cdot) \in H$ of is L_Φ-Hölder of order $\in (0, 1]$ on the closed ball $B_{H_{k'_X}}(B_{k'})$ [3]. Then for any $r > 0$ and $\in (0, 1)$, with probability at least 1, it holds that:

$$\sup_{h \in B_{H_k}(r)} |\hat{E}(h) - E(h)| \le C \left(B_\ell \sqrt{\frac{1}{M} \log \frac{1}{}} \right) \tag{10}$$

$$+ r B_k L_\ell B_{k'} L_\Phi \left(n^{-1} \log(M/) \right)^{/2} + B \sqrt{\frac{}{M}}; \tag{11}$$

where C is a constant.

The bound becomes larger in general if $(M;n)$ is replaced with $(1;Mn)$. It indicates that using domain-wise datasets is better than just pooling them into one mixed dataset, so the domain information plays a role. This result is later extended in [27], and Deshmukh et al. [28] give a bound for multi-class classification in a similar form.

3.2.2 Generalization risk bound

Another line of DG theory considers the risk on a specific target domain, under the assumption of covariate shift (i.e., the labeling function h or $P_{Y|X}$ is the same over all domains). This measurement is similar to what is considered in domain adaptation theory in Section 3.1, so we adopt the same definition for the source risks $\epsilon^1, \dots, \epsilon^M$ and the target risk ϵ^t. With the covariate shift assumption, each domain is characterized by the distribution on X. Albuquerque et al. [102] then consider approximating the target domain distribution P^t_X within the convex hull of source domain distributions: $\Lambda := \{\sum_{i=1}^M \lambda_i P^i_X \mid \lambda \in \Delta_M\}$, where Δ_M is the $(M-1)$-dimensional simplex so that each λ represents a normalized mixing weights. Similar to the domain adaptation case, distribution difference is measured by the H-divergence to include the influence of the classifier class.

Theorem 3 (Domain generalization error bound [102]). Let $\Delta := \min_{\lambda \in \Delta_M} d_H(P^t_X, \sum_{i=1}^M \lambda_i P^i_X)$ with minimizer [4] be the distance of P^t_X from the convex hull Λ, and $P_X := \sum_{i=1}^M \lambda_i P^i_X$ be the best approximator within Λ. Let $\Gamma := \sup_{P'_X, P''_X \in \Lambda} d_H(P'_X, P''_X)$ be the diameter of Λ. Then it holds that

$$\epsilon^t(h) \le \sum_{i=1}^M \lambda_i \epsilon^i(h) + \frac{\Gamma + \Delta}{2} + \epsilon_{H;(P^t_X, P_X)}; \tag{12}$$

where $\epsilon_{H;(P^t_X, P_X)}$ is the ideal joint risk across the target domain and the domain with the best approximator distribution P_X.

The result can be seen as the generalization of domain adaptation bounds in Section 3.1 when there are multiple source domains. Again similar to the domain adaptation case, this bound motivates domain generalization methods based on domain invariant representation, which simultaneously minimize the risks over all source domains corresponding to the first term of the bound, as well as the representation distribution differences among source and target domains in the hope to reduce and on the representation space. To sum up, the theory of generalization

3. This means that for any $u, v \in B_{H_{k'_X}}(B_{k'})$, it holds that $|k(u, \cdot) - (v, \cdot)|_k \le L_\Phi |u - v|^_{k'}$, where the norms are of the respective RKHSs.

4. The original presentation does not mention that the is the minimizer, but the proof indicates so.

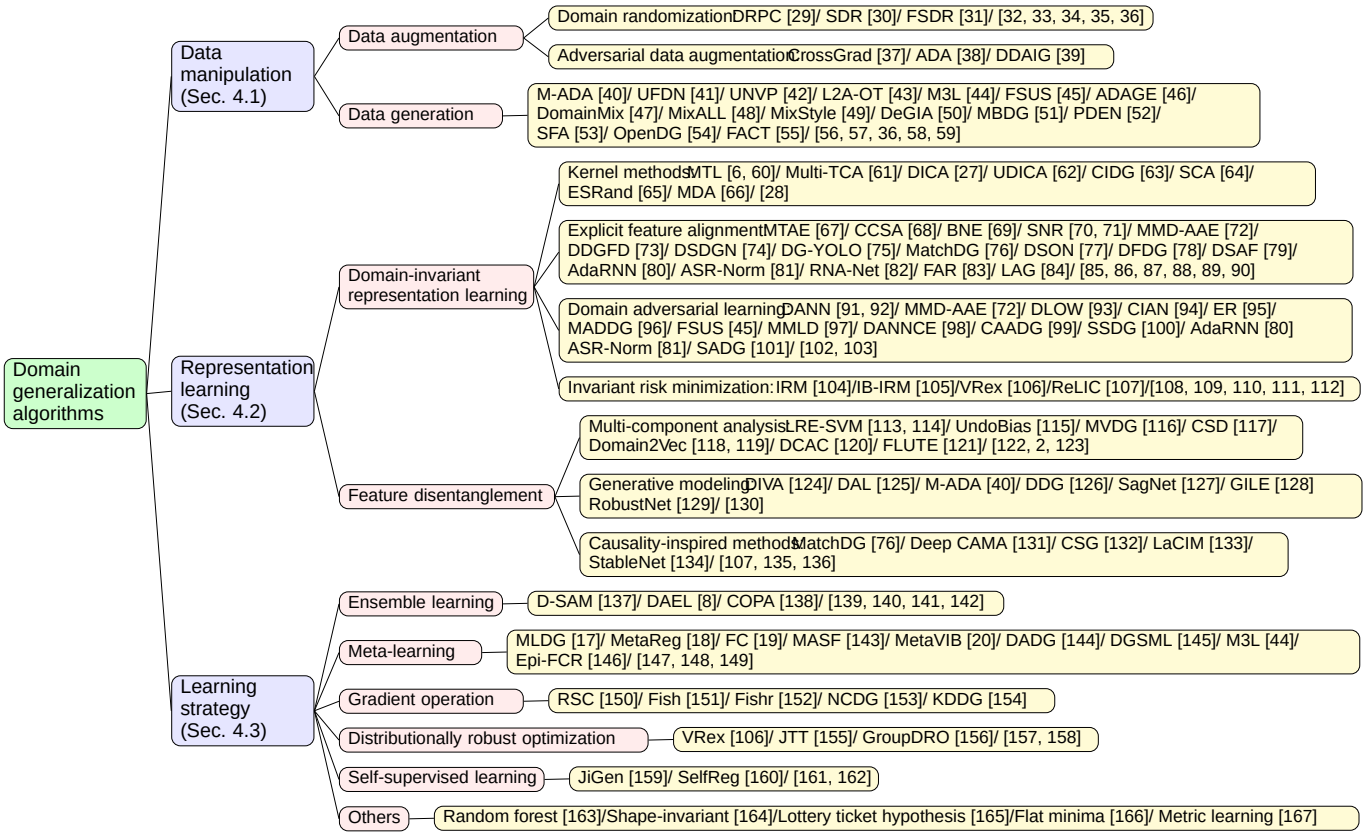


Fig. 3. Taxonomy of domain generalization methods.

is an active research area and other researchers also derived different DG theory bounds using informativeness [168] and adversarial training [102, 98, 168, 28].

4 METHODOLOGY

In this section, we introduce existing domain generalization methods in detail. As shown in Fig. 3, we categorize them into three groups, namely:

(1) Data manipulation: This category of methods focuses on manipulating the inputs to assist learning general representations. Along this line, there are two kinds of popular techniques: a). Data augmentation, which is mainly based on augmentation, randomization, and transformation of input data; b). Data generation, which generates diverse samples to help generalization.
(2) Representation learning: This category of methods is the most popular in domain generalization. There are two representative techniques: a). Domain-invariant representation learning which performs kernel, adversarial training, explicitly feature alignment between domains, or invariant risk minimization to learn domain-invariant representations; b). Feature disentanglement, which tries to disentangle the features into domain-shared or domain-speci c parts for better generalization.
(3) Learning strategy: This category of methods focuses on exploiting the general learning strategy to promote the generalization capability, which mainly includes several kinds of methods: a). Ensemble learning, which relies on the power of ensemble to learn a uni ed and generalized predictive function; b). Meta-learning which is based on the learning-to-learn mechanism to learn general knowledge by constructing meta-learning tasks to simulate domain shift; c). Gradient operation, which tries to learn generalized representations by directly operating on gradients; d). Distributionally robust optimization, which learns the worst-case distribution scenario of training domains; e). Self-supervised learning, which constructs pretext tasks to learn generalized representations. Additionally, there are other learning strategy that can be used for DG.

These three categories of approaches are conceptually different. They are complementary to each other and can be combined towards higher performance. We will describe the approaches for each category in detail hereafter.

4.1 Data Manipulation

We are always hungry for more training data in machine learning (ML). The generalization performance of a ML model often relies on the quantity and diversity of the training data. Given a limited set of training data, data manipulation is one of the cheapest and simplest way to generate samples so as to enhance the generalization capability of the model. The main objective for data manipulation-based DG is to increase the diversity of existing training data using different data manipulation methods. At the same time, the data quantity is also increased. Although the theoretical insight for why data augmentation or generation techniques can enhance the generalization ability of a model, experiments by Adila and Kang [169] showed that the model

tend to make predictions for both OOD and in-distribution samples based on trivial syntactic heuristics for NLP tasks.

We formulate the general learning objective of data manipulation-based DG as:

$$\min_{h} E_{x,y}[\ell(h(x), y)] + E_{x',y}[\ell(h(x'), y)], \quad (13)$$

where $x' = M(x)$ denotes the manipulated data using a function $M(\cdot)$. Based on the difference on this function, we further categorize existing work into two types: *data augmentation* and *data generation*.

4.1.1 Data augmentation-based DG

Augmentation is one of the most useful techniques for training machine learning models. Typical augmentation operations include flipping, rotation, scaling, cropping, adding noise, and so on. They have been widely used in supervised learning to enhance the generalization performance of a model by reducing overfitting [170, 36]. Without exception, they can also be adopted for DG where $M(\cdot)$ can be instantiated as these data augmentation functions.

4.1.1.1 Domain randomization : Other than typical augmentation, domain randomization is an effective technique for data augmentation. It is commonly done by generating new data that can simulate complex environments based on the limited training samples. Here, the $M(\cdot)$ function is implemented as several manual transformations (commonly used in image data) such as: altering the location and texture of objects, changing the number and shape of objects, modifying the illumination and camera view, and adding different types of random noise to the data. Tobin et al. [32] first used this method to generate more training data from the simulated environment for generalization in the real environment. Similar techniques were also used in [33, 34, 35, 29] to strengthen the generalization capability of the models. Prakash et al. [30] further took into account the structure of the scene when randomly placing objects for data generation, which enables the neural network to learn to utilize context when detecting objects. Peng et al. [58] proposed to not only augment features, but also labels. It is easy to see that by randomization, the diversity of samples can be increased. But randomization is often random, indicating that there could be some useless randomizations that could be further removed to improve the efficiency of the model.

4.1.1.2 Adversarial data augmentation : Adversarial data augmentation aims to guide the augmentation to optimize the generalization capability, by enhancing the diversity of data while assuring their reliability. Shankar et al. [37] used a Bayesian network to model dependence between label, domain and input instance, and proposed CrossGrad, a cautious data augmentation strategy that perturbs the input along the direction of greatest domain change while changing the class label as little as possible. Volpi et al. [38] proposed an iterative procedure that augments the source dataset with examples from a fictitious target domain that is "hard" under the current model, where adversarial examples are appended at each iteration to enable adaptive data augmentation. Zhou et al. [39] adversarially trained a transformation network for data augmentation instead of directly updating the inputs by gradient ascent while they adopted the regularization of weak and strong augmentation in [171, 31]. Adversarial data augmentation often has certain optimization goals that can be used by the network. However, its optimization process often involves adversarial training, thus is difficult.

4.1.2 Data generation-based DG

Data generation is also a popular technique to generate diverse and rich data to boost the generalization capability of a model. Here, the function $M(\cdot)$ can be implemented using some generative models such as Variational Autoencoder (VAE) [172], and Generative Adversarial Networks (GAN) [173]. In addition, it can also be implemented using the Mixup [174] strategy.

Rahman et al. [56] used ComboGAN [175] to generate new data and then applied domain discrepancy measure such as MMD [176] to minimize the distribution divergence between real and generated images to help learn general representations. Qiao et al. [40] leveraged adversarial training to create "fictitious" yet "challenging" populations, where a Wasserstein Auto-Encoder (WAE) [177] was used to help generate samples that preserve the semantic and have large domain transportation. Zhou et al. [43] generated novel distributions under semantic consistency and then maximized the difference between source and the novel distributions. Somavarapu et al. [57] introduced a simple transformation based on image stylization to explore cross-source variability for better generalization, where AdaIN [178] was employed to achieve fast stylization to arbitrary styles. Different from others, Li et al. [52] used adversarial training to generate domains instead of samples. These methods are more complex since different generative models are involved and we should pay attention to the model capacity and computing cost.

In addition to the above generative models, Mixup [174] is also a popular technique for data generation. Mixup generates new data by performing linear interpolation between any two instances and between their labels with a weight sampled from a Beta distribution, which does not require to train generative models. Recently, there are several methods using Mixup for DG, by either performing Mixup in the original space [47, 48, 54] to generate new samples; or in the feature space [49, 148, 55] which does not explicitly generate raw training samples. These methods achieved promising performance on popular benchmarks while remaining conceptually and computationally simple.

4.2 Representation Learning

Representation learning has always been the focus of machine learning for decades [179] and is also one of the keys to the success of domain generalization. We decompose the prediction function h as $h = f \circ g$, where g is a representation learning function and f is the classifier function. The goal of representation learning can be formulated as:

$$\min_{f,g} E_{x,y}\ell(f(g(x)), y) + \ell_{reg}, \quad (14)$$

where ℓ_{reg} denotes some regularization term and λ is the tradeoff parameter. Many methods are designed to better learn the feature extraction function g with corresponding ℓ_{reg}. In this section, we categorize the existing literature on representation learning into two main categories based on different learning principles: *domain-invariant representation learning* and *feature disentanglement*.

4.2.1 Domain-invariant representation-based DG

The work of [180] theoretically proved that if the feature representations remain invariant to different domains, the representations are general and transferable to different domains (also refer to Section 3). Based on this theory, a plethora of algorithms have been proposed for domain adaptation. Similarly, for domain generalization, the goal is to reduce the representation discrepancy between multiple source domains in a speci c feature space to be domain invariant so that the learned model can have a generalizable capability to the unseen domain. Along this line, there are mainly four types of methods: kernel-based methods, domain adversarial learning, explicit feature alignment, and invariant risk minimization.

4.2.1.1 Kernel-based methods : Kernel-based method is one of the most classical learning paradigms in machine learning. Kernel-based machine learning relies on the kernel function to transform the original data into a high-dimensional feature space without ever computing the coordinates of the data in that space, but by simply computing the inner products between the samples of all pairs in the feature space. One of the most representative kernel-based methods is Support Vector Machine (SVM) [181]. For domain generalization, there are plenty of algorithms based on kernel methods, where the representation learning function g is implemented as some feature map $( )$ which is easily computed using kernel function $k(;)$ such as RBF kernel and Laplacian kernel.

Blanchard et al. [6] rst used kernel method for domain generalization and extended it in [60]. They adopted the positive semi-de nite kernel learning to learn a domain-invariant kernel from the training data. Grubinger et al. [61] adapted transfer component analysis (TCA) [182] to bridge the multi-domain distance to be closer for DG. Similar to the core idea of TCA, Domain-Invariant Component Analysis (DICA) [27] is one of the classic methods using kernel for DG. The goal of DICA is to nd a feature transformation kernel $k(;)$ that minimizes the distribution discrepancy between all data in feature space. Gan et al. [62] adopted a similar method as DICA and further added attribute regularization. In contrast to DICA which deals with the marginal distribution, Li et al. [63] learned a feature representation which has domain-invariant class conditional distribution. Scatter component analysis (SCA) [64] adopted Fisher's discriminant analysis to minimize the discrepancy of representations from the same class and the same domain, and maximize the discrepancy of representations from the different classes and different domains. Erfani et al. [65] proposed an Elliptical Summary Randomisation (ESRand) that comprises of a randomised kernel and elliptical data summarization. ESRand projected each domain into an ellipse to represent the domain information and then used some similarity metric to compute the distance. Hu et al. [66] proposed multi-domain discriminant analysis to perform class-wise kernel learning for DG, which is more ne-grained. To sum up, this category of methods is often highly related to other categories to act as their divergence measures or theoretical support.

4.2.1.2 Domain adversarial learning : Domain-adversarial training is widely used for learning domain-invariant features. Ganin and Lempitsky [91] and Ganin et al. [92] proposed Domain-adversarial neural network (DANN) for domain adaptation, which adversarially trains the generator and discriminator. The discriminator is trained to distinguish the domains while the generator is trained to fool the discriminator to learn domain invariant feature representations. Li et al. [72] adopted such idea for DG. Gong et al. [93] used adversarial training by gradually reducing the domain discrepancy in a manifold space. Li et al. [94] proposed a conditional invariant adversarial network (CIAN) to learn class-wise adversarial networks for DG. Similar ideas were also used in [96, 99, 103]. Jia et al. [100] used single-side adversarial learning and asymmetric triplet loss to make sure only the real faces from different domains were indistinguishable, but not for the fake ones. After that, the extracted features of fake faces are more dispersed than before in the feature space and those of real ones are more aggregated, leading to a better generalized class boundary for unseen domains. In addition to adversarial domain classi cation, Zhao et al. [95] introduced additional entropy regularization by minimizing the KL divergence between the conditional distributions of different training domains to push the network to learn domain-invariant features. Some other GAN-based methods [45, 98, 102] were also proposed with theoretically guaranteed generalization bound.

4.2.1.3 Explicit feature alignment : This line of works aligns the features across source domains to learn domain-invariant representations through explicit feature distribution alignment [72, 183, 184, 185], or feature normalization [186, 187, 188, 70]. Motiian et al. [68] introduced a cross-domain contrastive loss for representation learning, where mapped domains are semantically aligned and yet maximally separated. Some methods explicitly minimized the feature distribution divergence by minimizing the maximum mean discrepancy (MMD) [189, 182, 190, 191], second order correlation [192, 193, 194], both mean and variance (moment matching) [184], Wasserstein distance [183], etc, of domains for either domain adaptation or domain generalization. Zhou et al. [183] aligned the marginal distribution of different source domains via optimal transport by minimizing the Wasserstein distance to achieve domain-invariant feature space.

Moreover, there are some works that used feature normalization techniques to enhance domain generalization capability [186, 187]. Pan et al. [186] introduced Instance Normalization (IN) layers to CNNs to improve the generalization capability of models. IN has been extensively investigated in the eld of image style transfer [195, 196, 178], where the style of an image is re ected by the IN parameters, i.e., mean and variance of each feature channel. Thus, IN layers [197] could be used to eliminate instance-speci c style discrepancy to enhance generalization [186]. However, IN is task agnostic and may remove some discriminative information. In IBNNet, IN and Batch Normalization (BN) are utilized in parallel to preserve some discriminative information [186]. In [188], BN layers are replaced by Batch-Instance Normalization (BIN) layers, which adaptively balance BN and IN for each channel by selectively using BN and IN. Jin et al. [70, 71] proposed a Style Normalization and Restitution (SNR) module to simultaneously ensure both high generalization and discrimination capability of the net-

works. After the style normalization by IN, a restitution step is performed to distill task-relevant discriminative features from the residual (i.e., the difference between the original feature and the style normalized feature) and add them back to the network to ensure high discrimination. The idea of restitution is extended to other alignment-based method to restorate helpful discriminative information dropped by alignment [83]. Recently, Qi et al. [79] applied IN to unsupervised DG where there are no labels in the training domains to acquire invariant and transferable features. A combination of different normalization techniques is presented in [81] to show that adaptively learning the normalization technique can improve DG. This category of methods is more flexible and can be applied to other kind of categories.

4.2.1.4 Invariant risk minimization (IRM) : Arjovsky et al. [104] considered another perspective on the domain-invariance of representation for domain generalization. They did not seek to match the representation distribution of all domains, but to enforce the optimal classifier on top of the representation space to be the same across all domains. The intuition is that the ideal representation for prediction is the cause of y, and the causal mechanism should not be affected by other factors/mechanisms, thus is domain-invariant. Formally, IRM can be formulated as:

$$\min_{g\in\mathcal{G};\ f\in\bigcap_{i=1}^{M}\arg\min_{f'\in\mathcal{F}}\ell_i(f'\circ g)} \sum_{i=1}^{M}\ell_i(f\circ g) \tag{15}$$

for some function classes $\mathcal{F}$ of g and $\mathcal{F}$ of f. The constraint for f embodies the desideratum that all domains share the same representation-level classifier, and the objective function encourages f and g to achieve a low source domain risk. However, this problem is hard to solve as it involves an inner-level optimization problem in its constraint. The authors then develop a surrogate problem to learn the feature extractor g that is much more practical:

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{M}\ell_i(g) + \left\|\nabla_{f=1.0}\,\ell_i(f\circ g)\right\|_{f=1}^{2}; \tag{16}$$

where a dummy representation-level classifier $f = 1$ is considered, and the gradient norm term measures the optimality of this classifier. The work also presents a generalization theory under a perhaps strong linear assumption, that for plenty enough source domains, the ground-truth invariant classifier can be identified.

IRM has gain notable visibility recently. There are some further theoretical analyses on the success [112] and failure cases of IRM [111], and IRM has been extended to other tasks including text classification [110] and reinforcement learning [198]. The idea to pursue the invariance of optimal representation-level classifier is also extended. Krueger et al. [106] promote this invariance by minimizing the extrapolated risk among source domains, which essentially minimizes the variance of source-domain risks. Mitrovic et al. [107] aim to learn such a representation in a self-supervised setup, where the second domain is constructed by data augmentation showing various semantic-irrelevant variations. Recently, Ahuja et al. [105] found the invariance of f alone is not sufficient. They found IRM still fails if g captures "fully informative invariant features", which makes y independent of x on all domains. This is particularly the case for classification (vs. regression) tasks. An information bottleneck regularization is hence introduced to maintain only partially informative features.

4.2.2 Feature disentanglement-based DG

Disentangled representation learning aims to learn a function that maps a sample to a feature vector, which contains all the information about different factors of variation and each dimension (or a subset of dimensions) contains information about only some factor(s). Disentanglement based DG approaches in general decompose a feature representation into understandable compositions/sub-features, with one feature being domain-shared/invariant feature and the other domain-specific feature. The optimization objective of disentanglement-based DG can be summarized as:

$$\min_{g_c;g_s;f} E_{x;y}\,\ell(f(g_c(x)),y) + \ell_{reg} + \ell_{recon}([g_c(x);g_s(x)];x), \tag{17}$$

where g_c and g_s denote the domain-shared and domain-specific feature representations, respectively. and are tradeoff parameters. The loss ℓ_{reg} is a regularization term that explicitly encourages the separation of the domain shared and specific features and ℓ_{recon} denotes a reconstruction loss that prevents information loss. Note that $[g_c(x);g_s(x)]$ denotes the combination of two kinds of features (which is not limited to concatenation operation).

Based on the choice of network structures and implementation mechanisms, disentanglement-based DG can mainly be categorized into three types: multi-component analysis, generative modeling, and causality-inspired methods

4.2.2.1 Multi-component analysis : In multi-component analysis, the domain-shared and domain-specific features are in general extracted using the domain-shared and domain-specific network parameters. The method of UndoBias [115] started from a SVM model to maximize interval classification on all training data for domain generalization. They represented the parameters of the i-th domain as $w_i = w_0 + {}_i$, where w_0 denotes the domain-shared parameters and ${}_i$ denotes the domain-specific parameters. Some other methods extented the idea of UndoBias from different aspects. Niu et al. [116] proposed to use multi-view learning for domain generalization. They proposed Multi-view DG (MVDG) to learn the combination of exemplar SVMs under different views for robust generalization. Ding and Fu [122] designed domain-specific networks for each domain and one shared domain-invariant network for all domains to learn disentangled representations, where low-rank reconstruction is adopted to align two types of networks in structured low-rank fashion. Li et al. [2] extended the idea of UndoBias into the neural network context and developed a low-rank parameterized CNN model for end-to-end training. Zunino et al. [123] learned disentangled representations through manually comparing the attention heat maps for certain areas from different domains. There are also other works that adopt multi-component analysis for disentanglement [118, 113, 114, 117, 121, 199, 200, 201, 119]. In general, multi-component analysis can be implemented in different architectures and remains effective for representation disentanglement.

4.2.2.2 Generative modeling : Generative models can be used for disentanglement from the perspective of data generation process. This kind of methods tries to formulate the generative mechanism of the samples from the domain-level, sample-level, and label-level. Some works further disentangle the input into class-irrelevant features, which contain the information related to speci c instance [202]. The Domain-invariant variational autoencoder (DIVA) [124] disentangled the features into domain information, category information, and other information, which is learned in the VAE framework. Peng et al. [125] disentangled the ne-grained domain information and category information that are learned in VAEs. Qiao et al. [40] also used VAE for disentanglement, where they proposed a Uni ed Feature Disentanglement Network (UFDN) that treated both data domains and image attributes of interest as latent factors to be disentangled. Similarly, Zhang et al. [126] disentangled the semantic and variational part of the samples. Similar spirits also include [130, 129]. Nam et al. [127] proposed to disentangle the style and other information using generative models that their method is both for domain adaptation and domain generalization. Generative models can not only improve OOD performance, but can also be used for generation tasks, which we believe is useful for many potential applications.

4.2.2.3 Causality-inspired methods : Causality is a ner description of variable relations beyond statistics (joint distribution). Causal relation gives information of how the system will behave under intervention, so it is naturally suitable for transfer learning tasks, since domain shift can be seen as an intervention. Particularly, under the causal consideration, the desired representation is the true cause of the label (e.g., object shape), so that prediction will not be affected by intervention on correlated but semantically irrelevant features (e.g., background, color, style). There are a number of works [203, 204, 205, 206] that exploited causality for domain adaptation.

For domain generalization, He et al. [136] reweighted input samples in a way to make the weighted correlation re ect causal effect. Zhang et al. [134] took Fourier features as the causing factors of images, and enforce the independence among these features. Using the additional data of object identity (it is a more detailed label than the class label), [207] enforced the conditional independence of the representation from domain index given the same object. When such object label is unavailable, [76] further learned an object feature based on labels in a separate stage. For single-source domain generalization, [107, 135] used data augmentation to present information of the causal factor. The augmentation operation is seen as producing outcomes under intervention on irrelevant features, which is implemented based on speci c domain knowledge. There are also generative methods under the causal consideration. Zhang et al. [131] explicitly modeled a manipulation variable that causes domain shift, which may be unobserved. Liu et al. [132] leveraged causal invariance for single-source generalization, i.e., the invariance of the process of generating $(x; y)$ data based on the factors, which is explained more general than inference invariance that existing methods implicitly rely on. The two factors are allowed correlated which is more realistic. They theoretically prove the identi ability of the causal factor is possible and the identi cation bene ts generalization. Sun et al. [133] extended the method and theory to multiple source domains. With more informative data, the irrelevant factor is also identi able.

4.3 Learning Strategy

In addition to data manipulation and representation learning, DG was also studied in general machine learning paradigms, which is divided into several categories: ensemble learning-based DG, meta-learning-based DG, gradient operation-based DG, distributionally robust optimization-based DG, self-supervised learning-based DG, and other strategies.

4.3.1 Ensemble learning-based DG

Ensemble learning usually combines multiple models, such as classi ers or experts, to enhance the power of models. For domain generalization, ensemble learning exploits the relationship between multiple source domains by using speci c network architecture designs and training strategies to improve generalization. They assume that any sample can be regarded as an integrated sample of the multiple source domains, so the overall prediction result can be seen as the superposition of the multiple domain networks.

Mancini et al. [139] proposed to use learnable weights for aggregating the predictions from different source speci c classi ers, where a domain predictor is used to predict the probability that a sample belongs to each domain (weights). Seg u et al. [69] maintained domain-dependent batch normalization (BN) statistics and BN parameters for different source domains while all the other parameters were shared. In inference, the nal prediction was a linear combination of the domain-dependent models with the combination weights inferred by measuring the distances between the instance normalization statistics of the test sample and the accumulated population statistics of each domain. The work of [137] proposed domain-speci c layers of different source domains and learning the linear aggregation of these layers to represent a test sample. Zhou et al. [8] proposed Domain Adaptive Ensemble Learning (DAEL), where a DAEL model is composed of a CNN feature extractor shared across domains and multiple domain-speci c classi er heads. Each classi er is an expert to its own domain and a non-expert to others. DAEL aims to learn these experts collaboratively, by teaching the non-experts with the expert so as to encourage the ensemble to learn how to handle data from unseen domains. There are also other works [138, 142]. Ensemble learning remains a powerful tool for DG since ensemble allows more diversity of models and features. However, one drawback of ensemble learning-based DG is maybe its computational resources as we need more space and computations for training and saving different models.

4.3.2 Meta-learning-based DG

The key idea of meta-learning is to learn a general model from multiple tasks by either optimization-based methods [208], metric-based learning [209], or model-based methods [210]. The idea of meta-learning has been exploited for domain generalization. They divide the data form multi-source domains into meta-train and meta-test sets to sim-

ulate domain shift. Denote the model parameters to be learned, meta-learning can be formulated as:

$$\begin{aligned} &= \text{Learn}(S_{mte}; \quad) \\ &= \text{Learn}(S_{mte}; \text{MetaLearn}(S_{mtrn})); \end{aligned} \tag{18}$$

where = MetaLearn(S_{mtrn}) denotes the meta-learned parameters from the meta-train set S_{mtrn} which is then used to learn the model parameters on the meta-test set S_{mte}. The two functions Learn( ) and MetaLearn( ) are to be designed and implemented by different meta-learning algorithms, which corresponds to a bi-level optimization problem. The gradient update can be formulated as:

$$= \frac{@(`(S_{mte}; ) + `(S_{mtrn};))}{@}; \tag{19}$$

where and are learning rates for outer and inner loops, respectively.

Finn et al. [208] proposed Model-agnostic meta-learning (MAML). Inspired by MAML, Li et al. [17] proposed MLDG (meta-learning for domain generalization) to use the meta-learning strategy for DG. MLDG splits the data from the source domains into meta-train and meta-test to simulate the domain shift situation to learn general representations. Balaji et al. [18] proposed to learn a meta regularizer (MetaReg) for the classi er. [19] proposed feature-critic training for the feature extractor by designing a meta optimizer. Dou et al. [143] used the similar idea of MLDG and additionally introduced two complementary losses to explicitly regularize the semantic structure of feature space. Du et al. [20] proposed an extended version of information bottleneck named Meta Variational Information Bottleneck (MetaVIB). They regularize the Kullback–Leibler (KL) divergence between distributions of latent encoding of the samples that have the same category from different domains and learn to generate weights by using stochastic neural networks. Recently, some works also adopted meta-learning for semi-supervised DG or discriminative DG [144, 145, 147, 44, 211]. Meta-learning is widely adopted in DG research and it can be incorporated into several paradigms such as disentanglement [199]. Meta-learning performs well on massive domains since meta-learning can seek transferable knowledge from multiple tasks.

4.3.3 Gradient operation-based DG

Other than meta-learning and ensemble learning, several recent works consider using gradient information to force the network to learn generalized representations. Huang et al. [150] proposed a self-challenging training algorithm that aims to learn general representations by manipulating gradients. They iteratively discarded the dominant features activated on the training data, and forced the network to activate remaining features that correlate with labels. In this way, network can be forced to learn from more bad cases which will improve generalization ability. Shi et al. [151] proposed a gradient-matching scheme, where their assumption is that the gradient direction of two domains should be the same to enhance common representation learning. To this end, they proposed to maximize the gradient inner product (GIP) to align the gradient direction across domains. With this operation, the network can nd weights such that the input-output correspondence is as close as possible across domains. GIP can be formulated as:

$$L = L_{cls}(S_{train};) \quad \frac{2}{M(M \quad 1)} \sum_{i;j}^{i \in j} G_i \quad G_j; \tag{20}$$

where G_i and G_j are gradient for two domains that can be calulated as $G = E\frac{@(x;y;)}{@}$. The gradient-invariance was achieved by adding CORAL [193] loss between gradients in [152], while Tian et al. [153] maximized the neuron coverage of DNN with gradient similarity regularization between the original and the augmented samples. Additionally, Wang et al. [154] designed a knowledge distillation approach for based on gradient ltering.

4.3.4 Distributionally robust optimization-based DG

The goal of distributionally robust optimization (DRO) [212] is to learn a model at worst-case distribution scenario to hope it can generalize well to the test data, which shares similar goal as DG. To optimize the worst-case distribution scenario, Sagawa et al. [213] proposed a GroupDRO algorithm that requires explicit group annotation of the samples. Such annotation was later narrowed down to a small fraction of validation set in [155], where they formulated a two-stage weighting framework. Other researchers reduced the variance of training domain risks by risk extrapolation (VRex) [106] or reducing class-conditioned Wasserstein DRO [158]. Recently, Koh et al. [157] proposed the setting of subpopulation shift where they also applied DRO to this problem. Particularly, Du et al. [80] proposed AdaRNN, a similar algorithm to the spirit of DRO that did not require explicit group annotation; instead, they learned the worst-case distribution scenario by solving an optimization problem. To summarize, DRO focuses on the optimization process that can also be leveraged in DG research.

4.3.5 Self-supervised learning-based DG

Self-supervised learning (SSL) is a recently popular learning paradigm that builds self-supervised tasks from the large-scale unlabeled data [214]. Inspired by this, Carlucci et al. [159] introduced a self-supervision task of solving jigsaw puzzles to learn generalized representations. Apart from introducing new pretext tasks, contrastive learning is another popular paradigm of self-supervised learning, which was adopted in several recent works [160, 162, 161]. The core of contrastive learning is to perform unsupervised learning between positive and negative pairs. Note that self-supervised learning is a general paradigm that can be applied to any existing DG methods, especially unsupervised DG where there are no labels in training domains [79]. Another possible application of SSL-based DG is the pretraining of multi-domain data that trains powerful pretraining models while also handling domain shifts. However, a possible limitation of SSL-based DG maybe its computational ef ciency and requirement of computing resources.

4.3.6 Other learning strategy for DG

There are some other learning strategies for domain generalization. For instance, metric learning was adopted in [167] to explore better pair-wise distance for DG. Ryu et al. [163] used random forest to improve the generalization ability of

convolutional neural networks (CNN). They sampled the triplets based on the probability mass function of the split results given by random forest, which is used for updating the CNN parameters by triplet loss. Other works [215, 216] adopted model calibration for DG, where they argued that the calibrated performance has a close relationship with OOD performance. Zhang et al. [165] followed the lottery ticket hyphothesis to design network substructures for DG, while Narayanan et al. [164] focused on the shape-invariant features. Additionally, Cha et al. [166] observed the at minima is important to DG and they designed a simple stochastic weight averaging densely method to nd the at minima. Since DG is a general learning problem, there will be more works that uses other strategy in the future.

5 Other Domain Generalization Research Areas

Most of the existing literature on domain generalization adopt the basic (traditional) de nition of DG in Def. 2. There is some existing literature that extends such setting to new scenarios to push the frontiers of DG (ref. TABLE 3). This section brie y discuss the existing DG research areas to give the readers a brief overview of this problem.

5.1 Single-source Domain Generalization

Setting $M = 1$ in Def. 2 gives single-source DG. Compared to traditional DG ($M > 1$), single-source DG becomes more challenging since there are less diversity in training domains. Thus, the key to this problem is to generate novel domains using data generation techniques to increase the diversity and informativeness of training data. Several methods designed different generation strategies [217, 100, 160, 52, 135, 58, 40, 81, 59] for single-source DG in computer vision tasks. A recent work [80] studied this setting in time series data where there is usually one uni ed dataset by using min-max optimization. We expect more application areas can bene t from single-source DG.

5.2 Semi-supervised Domain Generalization

Compared to traditional DG, semi-supervised DG does not require the full labels of training domains. It is common to apply existing semi-supervised learning algorithms such as FixMatch [221] and FlexMatch [222] to learn pseudo labels for the unlabeled samples. For instance, two recent works adopted the consistency regularization in semi-supervised learning [171, 218] for semi-supervised DG. It can be seen that this setting is more general than traditional DG and we expect there will be more works in this area.

5.3 Federated Learning with Domain Generalization

Privacy and security of machine learning is becoming increasingly critical [223]. Mahajan et al. [224] rstly studied this problem and showed that if the features are stable, then the model is more robust to membership inference attack. In federated DG [225, 226], models do not access the raw training data; instead, they aggregate the parameters from different clients. Under this circumstance, the key is to design better aggregation scheme through generalization techniques [219, 220, 138]. Federated DG is more important in healthcare [227]. On the other hand, decentralized training is another possible solution [228]. However, similar privacy risks emerge when there is a need to update the model. Thus, we hope there could be more research.

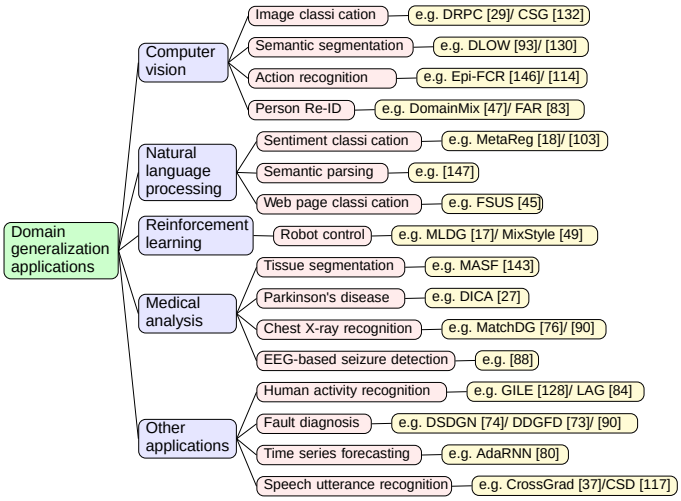


Fig. 4. Several applications of domain generalization.

5.4 Other DG Settings

There are also other settings in DG, such as open domain generalization [54] and unsupervised domain generalization [79]. Open DG shares the similar setting of universal domain adaptation where the training and test label spaces are not the same. Unsupervised DG assumes that all labels in the training domains are not accessible. As the environment gets more general and challenging, there will be other DG research areas aiming at solving certain limitations.

6 Applications

In this section, we discuss the popular tasks/applications for domain generalization (ref. Fig. 4).

High generalization ability is desired in various vision tasks. Many works investigate DG on classi cation. Some works also study DG for semantic segmentation [93], action recognition [114, 146], face anti-spoo ng [96], person Re-ID [47, 83], street view recognition [40], video understanding [116], and image compression [229]. Medical analysis [120] is one of the important application areas for DG due to its nature of data scarcity and existence of domain gaps, with the tasks of tissue segmentation [143], Parkinson's disease recognition [27], activity recognition [65], chest X-ray recognition [76, 90], and EEG-based seizure detection [88].

Apart from those areas, DG is also useful in reinforcement learning of robot control [49, 17] to generalize to an unseen environment. Some work used DG to recognize speech utterance [37, 117], fault diagnosis [90, 74, 73], physics [89], brain-computer interface [86].

In natural language processing, it is also common that training data comes from different domains with different distributions and DG techniques are helpful. Some work used domain generalization for sentiment classi cation on the Amazon Review dataset [103, 18]. Others used DG

TABLE 3
Existing research areas of domain generalization

Setting	De nition	Reference
Traditional domain generalization	Def. 2	Most of this paper
Single-source domain generalization	Set $M = 1$ in Def. 2	[217, 100, 160, 52, 135, 58, 40, 217, 81, 59]
Semi-supervised domain generalization	S_{train} is partially labeled	[171, 218]
Federated domain generalization	S_{train} cannot broadcast to the server	[219, 220, 138]
Open domain generalization	$Y_{train} \neq Y_{test}$	[54]
Unsupervised domain generalization	S_{train} is totally unlabeled	[79]

for semantic parsing [147], web page classi cation [45]. For instance, if we are given natural language data from multiple domains and want to learn a generalized model that predicts well on any new domain, we can use domain generalization to acquire domain-invariant representations.

Moreover, DG techniques favor broad prospects in some applications, such as nancial analysis, weather prediction, and logistics. For instance, Du et al. [80] tried to adopt DG to time series modeling. They rst propose the temporal covariate shift problem that widely exists in time series data, then, they proposed an RNN-based model to solve this problem to align the hidden representations between any pair of training data that are from different domains. Their algorithm, the so-called AdaRNN, was applied to stock price prediction, weather prediction, and electric power consumption. Another example is [128], where they applied domain generalization to sensor-based human activity recognition. In their application, the activity data from different persons are from different distributions, resulting in severe model collapse when applied to new users. To resolve such problem, they developed a variational autoencoder-based network to learn the domain-invariant and domain-speci c modules, thus achieving the disentanglement.

In the future, we hope there can be more DG applications in other areas to tackle with the distributional shift that widely exists in different applications. Another important problem is the evaluation of DG algorithms without accessing the test distribution in reality. While we can use the test data for evaluation in research, we simply cannot do it for real applications. In this case, one possible approach would be performing meta-train and meta-test split for the original data for multiple times. In each time, one split can be regarded as the unseen test data while the other as the training data. We can call it the meta-cross-validation for DG in reality. At the same time, we also hope there could be more evaluation metrics. For more evaluation in research, please refer to the next section.

7 Datasets, Evaluation, and Benchmark

In this section, we summarize the existing common datasets and model selection strategies for domain generalization. Then, we introduce the codebase, DeepDG, and demonstrate some observations from experiment conducted via it.

7.1 Datasets

TABLE 4 offers an overview of several popular datasets. Among them, PACS [2], VLCS [231], and Of ce-Home [232] are three most popular datasets. For large-scale evaluation, DomainNet [234] and Wilds [157] (i.e., a collection of datasets in TABLE 4 with `-wilds') are becoming popular.

Besides the datasets mentioned above, there exists some other datasets for domain generalization with different tasks. The Graft-versus-host disease (GvHD) dataset [246] is also popular and is used to test several methods [27, 6] for the ow cytometry problem. This dataset was collected from 30 patients with sample sizes ranging from 1,000 to 10,000. This is a time series classi cation dataset. Some works [93, 29, 71] applied domain generalization to semantic segmentation, where CityScape [247] and GTA5 [248] datasets were adopted as benchmark datasets. Some works applied DG to object detection, using the datasets of Cityscapes [247], GTA5 [248], Synthia [249] for investigation [71]. Some other works used public datasets or RandPerson [250] for person re-identi cation [187, 251, 70]. Some works [17, 49] used the OpenAI Gym [252] as the testbed to evaluate the performance of algorithms in reinforcement learning problems such as Cart-Pole and mountain car.

In addition to these widely used datasets, there are also other datasets used in existing literature. The Parkinson's telemonitoring dataset [253] is popular for predicting the clinician's motor and total UPDRS scoring of Parkinson's disease symptoms from voice measures. Some methods [27, 60, 61] used the data from several people as the training domains to learn models that generalize to unseen subjects.

It is worth noting that the datasets of domain generalization have some overlaps with domain adaptation. For instance, Of ce-31, Of ce-Caltech, Of ce-Home, and DomainNet are also widely used benchmarks for domain adaptation. Therefore, most domain adaptation datasets can be used for domain generalization benchmark in addition to those we discussed here. For example, Amazon Review dataset [254] is widely used in domain adaptation. It has four different domains on product review (DVDs, Kitchen appliance, Electronics and Books), which can also be used for domain generalization.

7.2 Evaluation

To test domain generalization algorithm on a test domain, three strategies are proposed [87], namely, Test-domain validation set, Leave-one-domain-out cross-validation, and Training-domain validation set. Test-domain validation set utilizes parts of the target domain as validations. Although it can obtain the best performance in most circumstances for that validation and testing share the same distribution, there is often no access to targets when training, which means it cannot be adopted in real applications. Leave-one-domain-out

TABLE 4
Eighteen popular datasets for domain generalization. The last ten datasets are from WILDS [157].

Dataset	#Domain	#Class	#Sample	Description	Reference
Of ce-Caltech	4	10	2,533	Caltech, Amazon, Webcam, DSLR	[230]
Of ce-31	3	31	4,110	Amazon, Webcam, DSLR	[230]
PACS	4	7	9,991	Art, Cartoon, Photos, Sketches	[2]
VLCS	4	5	10,729	Caltech101, LabelMe, SUN09, VOC2007	[231]
Of ce-Home	4	65	15,588	Art, Clipart, Product, Real	[232]
Terra Incognita	4	10	24,788	Wild animal images taken at locations L100, L38, L43, L46	[233]
Rotated MNIST	6	10	70,000	Digits rotated from 0 to 90 with an interval of 15	[67]
DomainNet	6	345	586,575	Clipart, Infograph, Painting, Quickdraw, Real, Sketch	[234]
iWildCam2020-wilds	323	182	203,029	Species classi cation across different camera traps	[235]
Camelyon17-wilds	5	2	45,000	Tumor identi cation across ve different hospitals	[236]
RxRx1-wilds	51	1,139	84,898	Genetic perturbation classi cation across experimental batches	[237]
OGB-MolPCBA	120,084	128	400,000	Molecular property prediction across different scaffolds	[238]
GlobalWheat-wilds	47	bounding boxes	6,515	Wheat head detection across regions of the world	[239]
CivilComments-wilds	-	2	450,000	Toxicity classi cation across demographic identities	[240]
FMoW-wilds	80	62	118,886	Land use classi cation across different regions and years	[241]
PovertyMap-wilds	46	real value	19,669	Poverty mapping across different countries	[242]
Amazon-wilds	3920	5	539,502	Sentiment classi cation across different users	[243]
Py150-wilds	8,421	next token	150,000	Code completion across different codebases	[244, 245]

is another strategy to choose the nal model when training data contains multiple sources. It leaves one training source as the validation while treating the others as the training part. Obviously, when only a single source exists in the training data, it is no longer applicable. In addition, due to different distributions among sources and targets, nal results rely heavily on the selections of validation, which makes nal results unstable. The most common strategy for domain generalization is Training-domain validation set which is used in most existing work. In this strategy, each source is split into two parts, the training part and the validation part. All training parts are combined for training while all validation parts are combined for selecting the best model. Since there still exists divergences between the combined validation and the real unseen targets, this simple and most popular strategy cannot achieve the best performance for some time.

We need to mention that there may exist other evaluation protocols for DG such as [168] since designing effective evaluation protocols is often consistent with the OOD performance. Currently, most of the works adopted the train-domain validation strategy which may not always generate good performance since the distribution of validation set is not the same as the new training data. On the other hand, using accuracy alone may not be suf cient to valid the model performance. We are looking forward to new evaluation metrics that can truly re ect the properties test distributions as much as possible in order to obtain better results.

7.3 Benchmark

To test the performance of DG algorithms in a uni ed codebase, in this paper, we develop a new codebase for DG, named DeepDG [255, 256]. Compared to the existing DomainBed [87], DeepDG simpli es the data loading and model selection process, while also makes it possible to run all experiments in a single machine. DeepDG splits the whole process into a data preparation part, a model part, a core algorithm part, a program entry, and some other auxiliary functions. Each part can be freely modi ed by users without affecting other parts. Users can add their own algorithms or datasets to DeepDG and compare with some state-of-the-art methods fairly. The current public version of DeepDG is only for image classi cation and we offer supports for Of ce-31, PACS, VLCS, and Of ce-Home datasets. Currently, nine state-of-the-art methods are implemented under the same environment, and it covers all three groups, including Data manipulation (Mixup [174]), Representation learning (DDC [189], DANN [91], CORAL [192]), and Learning strategy (MLDG [17], RSC [150], GroupDRO [156], ANDMask [257]).

TABLE 5
Benchmark results for PACS and Of ce-Home with DeepDG.

Dataset	PACS					Of ce-Home				
Method	A	C	P	S	AVG	A	C	P	R	AVG
ERM	77.0	74.5	95.5	77.8	81.2	58.6	47.9	72.2	73.0	62.9
DANN [92]	78.7	75.3	94.0	77.8	81.4	57.7	44.4	71.9	72.5	61.6
CORAL [193]	77.7	77.0	92.6	80.5	82.0	58.7	48.7	72.3	73.6	63.3
Mixup [174]	79.1	73.4	94.4	76.7	80.9	55.7	47.9	71.9	72.8	62.1
RSC [150]	79.7	76.1	95.6	76.6	82.0	58.9	49.2	72.5	74.2	63.7
GroupDRO [213]	76.0	76.0	91.2	79.0	80.6	57.6	48.7	71.5	73.1	62.7
ANDMask [257]	76.2	73.8	91.6	78.0	79.9	56.7	45.9	70.6	73.2	61.6

We conduct some experiments on the two most popular image classi cation datasets, PACS and Of ce-Home, with DeepDG and TABLE 5 shows the results. ResNet-18 is used as the base feature network. Training-domain validation set is used for selecting nal models, and 20% of sources are for validation while the others are for training. From TABLE 5, we observe more insightful conclusions. (1) The baseline method, ERM, has achieved acceptable results on both datasets. Some methods, such as DANN and ANDMask, even have worse performance. (2) Simple data augmentation method, Mixup, cannot obtain remarkable results. (3) CORAL has slight improvements on both datasets compared to ERM, which is consistent with results offered by DomainBed [87]. (4) RSC, a learning strategy, achieves the best performance on both datasets, but the improvements are unremarkable compared to ERM. The results indicate the bene ts of domain generalization in different tasks.

8 Discussion

In this section, we summarize existing methods and then present several challenges for future.

8.1 Summary of Existing Literature

The quantity and diversity of training data are critical to a model's generalization ability. Many methods aim to enrich the training data with the data manipulation methods to achieve good performance. However, one issue of the data manipulation methods is that there is a lack of theoretical guarantee of the unbound risk of generalization. Therefore, it is important to develop theories for the manipulation-based methods which could further guide the data generation designs without violating ethical standards.

Compared to data manipulation, representation learning has theoretical support in general [180, 6, 102]. Kernel-based methods are widely used in traditional methods while deep learning-based methods play a leading role in recent years. While domain adversarial training often achieves better performance in domain adaptation, in DG, we did not see signi cant results improvements from these adversarial methods. We think this is probably because the task is relatively easy. For the explicit distribution matching, more and more works tend to match the joint distributions rather than just match the marginal [6, 72] or conditional [63] distributions. Thus, it is more feasible to perform dynamic distribution matching [190, 191]. Both disentanglement and IRM methods have good motivations for generalization, while more ef cient training strategy can be developed. There are several studies [258] that pointed out merely learning domain-invariant features are insuf cient and representation smoothness should also be considered.

For learning strategy, there is a trend that many works used meta-learning for DG, where it requires to design better optimization strategies to utilize the rich information of different domains. In addition to deep networks, there are also some work [163] that used random forest for DG, and we hope more diverse methods will come.

8.2 Future Research Challenges

8.2.1 Continuous domain generalization

For many real applications, a system consumes streaming data with non-stationary statistics. In this case, it is of great importance to perform continuous domain generalization that ef ciently updates DG models to overcome catastrophic forgetting and adapts to new data. While there are some domain adaptation methods focusing on continuous learning [259], there are only very few investigations on continuous DG [260] whenever this is favorable in real scenarios.

8.2.2 Domain generalization to novel categories

The existing DG algorithms usually assume the label space for different domains are the same. A more practical and general setting is to support the generalization on new categories, i.e., both domain and task generalization. This is conceptually similar to the goal of meta-learning and zero-shot learning. Some work [85, 261] proposed zero-shot DG and we expect more work to come in this area. There are some prior work [54, 262] that tried to use the boundary-based learning paradigms or consistency regularization to solve this problem, which are good approaches that future work might build methods upon them.

8.2.3 Interpretable domain generalization

Disentanglement-based DG methods decompose a feature into domain-invariant/shared and domain-speci c parts, which provide some interpretation to DG. For other categories of methods, there is still a lack of deep understanding of the semantics or characteristics of learned features in DG models. For example, how to relate the results of the approach with the input feature space. How close are current methods to provide this level of interpretability? Causality [132] may be one promising tool to understand domain generalization networks and provide interpretations.

8.2.4 Large-scale pre-training/self-learning and DG

In recent years, we have witnessed the rapid development of large-scale pre-training/self-learning, such as BERT [263], GPT-3 [264], and Wav2vec [265]. Pre-training on large-scale dataset and then netuning the model to downstream tasks could improve its performance, where pre-training is bene cial to learn general representations. Therefore, how to design useful and ef cient DG methods to help large-scale pre-training/self-learning is worth investigating.

8.2.5 Test-time Generalization

While DG focuses on the training phase, we can also request test-time generalization in inference phase. This further bridges domain adaptation and domain generalization since we can also use the inference unlabeled data for adaptation. Very few recent works [266, 267] paid attention to this setting. Compared to traditional DG, test-time generalization will allow more exibility in inference time, while it requires less computation and more ef ciency as there are often limited resources in inference-end devices.

8.2.6 Performance evaluation for DG

The recent work [87] pointed out that on several datasets, the performance of some DG methods is almost the same as the baseline (i.e., ERM). We do not take it as the full evidence that DG is not useful in real applications. Instead, we argue that this might be due to the inappropriate evaluation schemes in use today, or the domain gaps not being so large. In more realistic situations such as person ReID where there are obvious domain gaps [71], the improvement of DG is dramatic. Therefore, we stay positive about the value of DG and hope researchers can also nd more suitable settings and datasets for the study.

9 Conclusion

Generalization has always been an important research topic in machine learning research. In this paper, we review the domain generalization areas by providing in-depth analysis of theories, existing methods, datasets, benchmarks, and applications. Then, we thoroughly analyze the methods. Based on our analysis, we provide several potential research challenges that could be the directions of future research. We hope that this survey can provide useful insights to researchers and inspire more progress in the future.

Acknowledgement

This work is supported in part by NSFC (No. 61972383), NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941.

REFERENCES

[1] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, Dataset shift in machine learning. Mit Press, 2009.

[2] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in ICCV, 2017, pp. 5542–5550.

[3] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization in vision: A survey," arXiv:2103.02503, 2021.

[4] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," arXiv preprint arXiv:2108.13624, 2021.

[5] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," arXiv preprint:2110.11334, 2021.

[6] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," J. Mach. Learn. Res., vol. 22, pp. 2–1, 2021.

[7] R. Caruana, "Multitask learning," Machine learning, vol. 28, no. 1, pp. 41–75, 1997.

[8] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," IEEE TIP, 2021.

[9] S. Pan and Q. Yang, "A survey on transfer learning," IEEE TKDE, vol. 22, pp. 1345–1359, 2010.

[10] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big data, vol. 3, no. 1, pp. 1–40, 2016.

[11] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

[12] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," Neurocomputing, vol. 312, pp. 135–153, 2018.

[13] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," IEEE signal processing magazine, vol. 32, no. 3, pp. 53–69, 2015.

[14] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," Artif. Intell. Rev., vol. 18, no. 2, pp. 77–95, 2002.

[15] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," IEEE TPAMI, 2020.

[16] J. Vanschoren, "Meta-learning: A survey," arXiv preprint arXiv:1810.03548, 2018.

[17] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in AAAI, 2018.

[18] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in NeurIPS, 2018, pp. 998–1008.

[19] Y. Li, Y. Yang, W. Zhou, and T. M. Hospedales, "Feature-critic networks for heterogeneous domain generalization," in ICML, 2019.

[20] Y. Du, J. Xu, H. Xiong, Q. Qiu, X. Zhen, C. G. M. Snoek, and L. Shao, "Learning to learn with variational information bottleneck for domain generalization," in ECCV, 2020.

[21] M. Biesialska, K. Biesialska, and M. R. Costa-jussa, "Continual lifelong learning in natural language processing: A survey," in COLING, 2020.

[22] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," ACM Trans. Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–37, 2019.

[23] Z. JI, H. WANG, Y. YU, and Y. PANG, "A decadal survey of zero-shot image classi cation," SCIENTIA SINICA Informationis, vol. 49, no. 10, pp. 1299–1320, 2019.

[24] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," Machine learning, vol. 79, no. 1-2, pp. 151–175, 2010.

[25] F. D. Johansson, D. Sontag, and R. Ranganath, "Support and invertibility in domain-invariant representations," in AISTAS, 2019, pp. 527–536.

[26] V. Vapnik, E. Levin, and Y. L. Cun, "Measuring the vc-dimension of a learning machine," Neural computation, vol. 6, no. 5, pp. 851–876, 1994.

[27] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in ICML, 2013.

[28] A. A. Deshmukh, Y. Lei, S. Sharma, U. Dogan, J. W. Cutler, and C. Scott, "A generalization error bound for multi-class domain generalization," arXiv:1905.10392, 2019.

[29] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in ICCV, 2019, pp. 2100–2110.

[30] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birch eld, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in ICRA. IEEE, 2019, pp. 7249–7255.

[31] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in CVPR, 2021, pp. 6891–6902.

[32] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in IROS, 2017, pp. 23–30.

[33] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in ICRA. IEEE, 2018, pp. 1–8.

[34] R. Khirodkar, D. Yoo, and K. Kitani, "Domain randomization for scene-speci c car detection and pose estimation," in WACV. IEEE, 2019, pp. 1932–1940.

[35] J. Tremblay, A. Prakash, D. Acuna, M. Brophy et al., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in CVPR Workshop, 2018.

[36] N. H. Nazari and A. Kovashka, "Domain generalization using shape representation," in ECCV, 2020, pp. 666–670.

[37] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in ICLR, 2018.

[38] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in NeurIPS, 2018, pp. 5334–5344.

[39] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep domain-adversarial image generation for domain generalisation," in AAAI, 2020.

[40] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in CVPR, 2020, pp. 12 556–12 565.

[41] A. H. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, "A uni ed feature disentangler for multi-domain image translation and manipulation," in NeurIPS, 2018, pp. 2590–2599.

[42] T.-D. Truong, C. N. Duong, K. Luu, and M.-T. Tran, "Recognition in unseen domains: Domain generalization via universal non-volume preserving models," arXiv preprint:1905.13040, 2019.

[43] K. Zhou, Y. Yang, T. M. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in ECCV, 2020.

[44] Y. Zhao, Z. Zhong, F. Yang, Z. Luo, Y. Lin et al., "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identi cation," in CVPR, 2021.

[45] V. K. Garg, A. Kalai, K. Ligett, and Z. S. Wu, "Learn to expect the unexpected: Probably approximately correct domain generalization," in AISTAS, 2021.

[46] F. M. Carlucci, P. Russo, T. Tommasi, and B. Caputo, "Hallucinating agnostic images to generalize across domains," in CVPR Workshop, 2019, pp. 0–0.

[47] W. Wang, S. Liao, F. Zhao, C. Kang, and L. Shao, "Domainmix: Learning generalizable person re-identi cation without human annotations," in BMCV, 2021.

[48] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in ICASSP, 2020, pp. 3622–3626.

[49] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in ICLR, 2021.

[50] T.-D. Truong, K. Luu, C.-N. Duong, N. Le, and M.-T. Tran, "Image alignment in unseen domains via domain deep generalization," arXiv preprint arXiv:1905.12028, 2019.

[51] A. Robey, G. J. Pappas, and H. Hassani, "Model-based domain generalization," in NeurIPS, 2021.

[52] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia, "Progressive domain expansion network for single domain generalization," in CVPR, 2021, pp. 224–233.

[53] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in ICCV, 2021, pp. 8886–8895.

[54] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in CVPR, 2021, pp. 9624–9633.

[55] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in CVPR, 2021.

[56] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "Multi-component image translation for deep domain generalization," in WACV. IEEE, 2019, pp. 579–588.

[57] N. Somavarapu, C.-Y. Ma, and Z. Kira, "Frustratingly simple

domain generalization via image stylization," arXiv:2006.11207, 2020.
[58] X. Peng, F. Qiao, and L. Zhao, "Out-of-domain generalization from a single source: An uncertainty quantification approach," arXiv preprint arXiv:2108.02888, 2021.
[59] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in ICCV, 2021, pp. 834–843.
[60] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," arXiv preprint arXiv:1711.07910, 2017.
[61] T. Grubinger, A. Birlutiu, H. Schöner, T. Natschläger, and T. Heskes, "Domain generalization based on transfer component analysis," in ICANN, 2015, pp. 325–334.
[62] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in CVPR, 2016, pp. 87–97.
[63] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao, "Domain generalization via conditional invariant representations," in AAAI, 2018.
[64] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," IEEE TPAMI, vol. 39, no. 7, pp. 1414–1430, 2016.
[65] S. Erfani, M. Baktashmotlagh, M. Moshtaghi, V. Nguyen, C. Leckie, J. Bailey, and R. Kotagiri, "Robust domain generalisation by enforcing distribution invariance," in AAAI, 2016.
[66] S. Hu, K. Zhang, Z. Chen, and L. Chan, "Domain generalization via multidomain discriminant analysis," in UAI, vol. 35, 2019.
[67] M. Ghifary, W. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," ICCV, pp. 2551–2559, 2015.
[68] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in ICCV, 2017, pp. 5715–5725.
[69] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," arXiv preprint arXiv:2011.12672, 2020.
[70] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in CVPR, 2020, pp. 3143–3152.
[71] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domain generalization and adaptation," IEEE Trans. Multimedia, 2021.
[72] H. Li, S. J. Pan, S. Wang, and A. Kot, "Domain generalization with adversarial feature learning," in CVPR, 2018, pp. 5400–5409.
[73] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang, and M. Xu, "Deep domain generalization combining a priori diagnosis knowledge toward cross-domain fault diagnosis of rolling bearing," IEEE Trans. on Instrumentation and Measurement, vol. 70, pp. 1–11, 2020.
[74] Y. Liao, R. Huang, J. Li, Z. Chen, and W. Li, "Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed," IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 10, pp. 8064–8075, 2020.
[75] H. Liu, P. Song, and R. Ding, "Towards domain generalization in underwater object detection," in ICIP. IEEE, 2020, pp. 1971–1975.
[76] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in ICML, 2021.
[77] S. Seo, Y. Suh, D. Kim, J. Han, and B. Han, "Learning to optimize domain specific normalization for domain generalization," in ECCV, 2020.
[78] W. Zhang, M. Ragab, and R. Sagarna, "Robust domain-free domain generalization with class-aware alignment," in ICASSP, 2021.
[79] L. Qi, L. Wang, Y. Shi, and X. Geng, "Unsupervised domain generalization for person re-identification: A domain-specific adaptive framework," arXiv preprint arXiv:2111.15077, 2021.
[80] Y. Du, J. Wang, W. Feng, S. Pan, T. Qin, R. Xu, and C. Wang, "Adarnn: Adaptive learning and forecasting of time series," in CIKM, 2021, pp. 402–411.
[81] X. Fan, Q. Wang, J. Ke, F. Yang, B. Gong, and M. Zhou, "Adversarially adaptive normalization for single domain generalization," in CVPR, 2021, pp. 8208–8217.
[82] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in WACV, 2022.
[83] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Feature alignment and restoration for domain generalization and adaptation," in NeurIPS, 2020.
[84] W. Lu, J. Wang, and Y. Chen, "Local and global alignments for generalizable sensor-based human activity recognition," in ICASSP, 2022.
[85] U. Maniyar, A. A. Deshmukh, U. Dogan, V. N. Balasubramanian et al., "Zero shot domain generalization," in BMVC, 2020.
[86] D.-K. Han and J.-H. Jeong, "Domain generalization for session-independent brain-computer interface," in International Winter Conference on Brain-Computer Interface. IEEE, 2021, pp. 1–5.
[87] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in ICLR, 2021.
[88] K. Ayodele, W. Ikezogwo, M. Komolafe, and P. Ogunbona, "Supervised domain generalization for integration of disparate scalp eeg datasets for automatic epileptic seizure detection," Computers in Biology and Medicine, vol. 120, p. 103757, 2020.
[89] E. Chen, T. S. Mathai, V. Sarode, H. Choset, and J. Galeotti, "A study of domain generalization on ultrasound-based multi-class segmentation of arteries, veins, ligaments, and nerves using transfer learning," in Machine Learning for Health (ML4H) at NeurIPS 2020, 2020.
[90] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Domain generalization in rotating machinery fault diagnostics using deep neural networks," Neurocomputing, vol. 403, pp. 409–420, 2020.
[91] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in ICML, 2015, pp. 1180–1189.
[92] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, pp. 59:1–59:35, 2016.
[93] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in CVPR, 2019, pp. 2477–2486.
[94] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in ECCV, 2018, pp. 624–639.
[95] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, "Domain generalization via entropy regularization," in NeurIPS, vol. 33, 2020.
[96] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in CVPR, 2019, pp. 10 023–10 031.
[97] T. Matsuura and T. Harada, "Domain generalization using a mixture of multiple latent domains," in AAAI, 2020.
[98] A. Sicilia, X. Zhao, and S. J. Hwang, "Domain adversarial neural networks for domain generalization: When it works and how to improve," arXiv preprint arXiv:2102.03924, 2021.
[99] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "Correlation-aware adversarial domain adaptation and generalization," Pattern Recognition, vol. 100, p. 107124, 2020.
[100] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in CVPR, 2020.
[101] J. Luo, J. Guo, W. Qiu, Z. Huang, and H. Hui, "Scale invariant domain generalization image recapture detection," in ICONIP, 2021.
[102] I. Albuquerque, J. Monteiro, T. H. Falk, and I. Mitliagkas, "Adversarial target-invariant representation learning for domain generalization," arXiv preprint arXiv:1911.00804, 2019.
[103] Z. Wang, Q. Wang, C. Lv, X. Cao, and G. Fu, "Unseen target stance detection with adversarial domain generalization," in IJCNN, 2020, pp. 1–8.
[104] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv preprint arXiv:1907.02893, 2019.
[105] K. Ahuja, E. Caballero, D. Zhang, Y. Bengio, I. Mitliagkas, and I. Rish, "Invariance principle meets information bottleneck for out-of-distribution generalization," in NeurIPS, 2021.
[106] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in ICML, 2021, pp. 5815–5826.
[107] J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," in ICLR, 2021.
[108] R. Guo, P. Zhang, H. Liu, and E. Kiciman, "Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix," arXiv preprint arXiv:2101.07732, 2021.
[109] K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney, "Empirical or invariant risk minimization? a sample complexity perspective," in ICLR, 2021.
[110] Y. J. Choe, J. Ham, and K. Park, "An empirical study of invariant risk minimization," arXiv preprint arXiv:2004.05007, 2020.

[111] E. Rosenfeld, P. Ravikumar, and A. Risteski, "The risks of invariant risk minimization," in ICLR, 2021.
[112] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, "Invariant risk minimization games," in ICML, 2020, pp. 145–155.
[113] Z. Xu, W. Li, L. Niu, and D. Xu, "Exploiting low-rank structure from latent domains for domain generalization," in ECCV. Springer, 2014, pp. 628–643.
[114] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, "Domain generalization and adaptation using low rank exemplar svms," TPAMI, vol. 40, no. 5, pp. 1114–1127, 2017.
[115] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in ECCV. Springer, 2012, pp. 158–171.
[116] L. Niu, W. Li, and D. Xu, "Multi-view domain generalization for visual recognition," in ICCV, 2015, pp. 4193–4201.
[117] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in ICML, 2020, pp. 7728–7738.
[118] A. A. Deshmukh, A. Bansal, and A. Rastogi, "Domain2vec: Deep domain generalization," arXiv preprint arXiv:1807.02919, 2018.
[119] X. Peng, Y. Li, and K. Saenko, "Domain2vec: Domain embedding for unsupervised domain adaptation," in ECCV, 2020, pp. 756–774.
[120] S. Hu, Z. Liao, J. Zhang, and Y. Xia, "Domain and content adaptive convolution for domain generalization in medical image segmentation," arXiv preprint arXiv:2109.05676, 2021.
[121] E. Triantafillou, H. Larochelle, R. Zemel, and V. Dumoulin, "Learning a universal template for few-shot dataset generalization," in ICML, 2021.
[122] Z. Ding and Y. Fu, "Deep domain generalization with structured low-rank constraint," IEEE TIP, vol. 27, no. 1, pp. 304–313, 2017.
[123] A. Zunino, S. A. Bargal, R. Volpi, M. Sameki, J. Zhang, S. Sclaroff, V. Murino, and K. Saenko, "Explainable deep classification models for domain generalization," in CVPR, 2021.
[124] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "Diva: Domain invariant variational autoencoders," in Proceedings of the Third Conference on Medical Imaging with Deep Learning, 2020.
[125] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in ICML, 2019.
[126] H. Zhang, Y.-F. Zhang, W. Liu, A. Weller, B. Schölkopf, and E. P. Xing, "Towards principled disentanglement for domain generalization," in ICML2021 Machine Learning for Data Workshop, 2021.
[127] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in CVPR, 2021, pp. 8690–8699.
[128] H. Qian, S. J. Pan, C. Miao, H. Qian, S. Pan, and C. Miao, "Latent independent excitation for generalizable sensor-based cross-person activity recognition," in AAAI, vol. 35, no. 13, 2021, pp. 11 921–11 929.
[129] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in CVPR, 2021, pp. 11 580–11 590.
[130] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in CVPR, 2021, pp. 8300–8311.
[131] C. Zhang, K. Zhang, and Y. Li, "A causal view on robustness of neural networks," in NeurIPS, 2020.
[132] C. Liu, X. Sun, J. Wang, T. Li, T. Qin, W. Chen, and T.-Y. Liu, "Learning causal semantic representation for out-of-distribution prediction," in NeurIPS, 2021.
[133] X. Sun, B. Wu, X. Zheng, C. Liu, W. Chen, T. Qin, and T.-Y. Liu, "Recovering latent causal factor for generalization to distributional shifts," in NeurIPS, 2021.
[134] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in CVPR, 2021, pp. 5372–5382.
[135] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert, "Causality-inspired single-source domain generalization for medical image segmentation," arXiv preprint arXiv:2111.12525, 2021.
[136] Y. He, Z. Shen, and P. Cui, "Towards non-i.i.d. image classification: A dataset and baselines," Pattern Recognition, 2019.
[137] A. D'Innocente and B. Caputo, "Domain generalization with domain-specific aggregation modules," in German Conference on Pattern Recognition. Springer, 2018, pp. 187–198.
[138] G. Wu and S. Gong, "Collaborative optimization and aggregation for decentralized domain generalization and adaptation," in ICCV, 2021, pp. 6484–6493.
[139] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "Best sources forward: domain generalization through source-specific nets," in ICIP, 2018, pp. 1353–1357.
[140] W. He, H. Zheng, and J. Lai, "Domain attention model for domain generalization in object detection," in PRCV, 2018, pp. 27–39.
[141] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "Robust place categorization with deep domain generalization," IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 2093–2100, 2018.
[142] A. Dubey, V. Ramanathan, A. Pentland, and D. Mahajan, "Adaptive methods for real-world domain generalization," in CVPR, 2021, pp. 14 340–14 349.
[143] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in NeurIPS, 2019.
[144] K. Chen, D. Zhuang, and J. M. Chang, "Discriminative adversarial domain generalization with meta-learning based cross-domain validation," Neurocomputing, vol. 467, pp. 418–426, 2022.
[145] H. Shahi-Noghabi, H. Asghari, N. Mehrasa, and M. Ester, "Domain generalization via semi-supervised meta learning," arXiv preprint arXiv:2009.12658, 2020.
[146] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in CVPR, 2019, pp. 1446–1455.
[147] B. Wang, M. Lapata, and I. Titov, "Meta-learning for domain generalization in semantic parsing," in NAACL, 2021.
[148] F. Qiao and X. Peng, "Uncertainty-guided model generalization to unseen domains," in CVPR, 2021, pp. 6790–6800.
[149] J. Kim, J. Lee, J. Park, D. Min, and K. Sohn, "Self-balanced learning for domain generalization," in ICIP. IEEE, 2021, pp. 779–783.
[150] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in ECCV, vol. 2, 2020.
[151] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in ICLR, 2022.
[152] A. Rame, C. Dancette, and M. Cord, "Fishr: Invariant gradient variances for out-of-distribution generalization," arXiv preprint arXiv:2109.02934, 2021.
[153] C. X. Tian, H. Li, X. Xie, Y. Liu, and S. Wang, "Neuron coverage-guided domain generalization," IEEE TPAMI, 2022.
[154] Y. Wang, H. Li, L.-p. Chau, and A. C. Kot, "Embracing the dark knowledge: Domain generalization using regularized knowledge distillation," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2595–2604.
[155] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in ICML, 2021, pp. 6781–6792.
[156] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in ICLR, 2020.
[157] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, and other, "Wilds: A benchmark of in-the-wild distribution shifts," in ICML, 2021, pp. 5637–5664.
[158] J. Wang, Y. Li, L. Xie, and Y. Xie, "Class-conditioned domain generalization via wasserstein distributional robust optimization," in RobustML workshop at ICLR2021.
[159] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in CVPR, 2019, pp. 2229–2238.
[160] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," in ICCV, 2021, pp. 9619–9628.
[161] S. Jeon, K. Hong, P. Lee, J. Lee, and H. Byun, "Feature stylization and domain-aware contrastive learning for domain generalization," in ACMMM, 2021, pp. 22–31.
[162] Z. Li, Z. Cui, S. Wang, Y. Qi, X. Ouyang, Q. Chen, Y. Yang, Z. Xue, D. Shen, and J.-Z. Cheng, "Domain generalization for mammography detection via multi-style and multi-view contrastive learning," in MICCAI, 2021, pp. 98–108.
[163] J. Ryu, G. Kwon, M.-H. Yang, and J. Lim, "Generalized convolutional forest networks for domain generalization and visual recognition," in ICLR, 2019.

[164] M. Narayanan, V. Rajendran, and B. Kimia, "Shape-biased domain generalization via shock graph embeddings," in *ICCV*, 2021, pp. 1315–1325.

[165] D. Zhang, K. Ahuja, Y. Xu, Y. Wang, and A. Courville, "Can subnetwork structure be the key to out-of-distribution generalization?" in *ICML*, 2021.

[166] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," in *NeurIPS*, 2021.

[167] M. Faraki, X. Yu, Y.-H. Tsai, Y. Suh, and M. Chandraker, "Cross-domain similarity learning for face recognition in unseen domains," in *CVPR*, 2021, pp. 15 292–15 301.

[168] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," in *NeurIPS*, 2021.

[169] D. Adila and D. Kang, "Understanding out-of-distribution: A perspective of data dynamics," in *NeurIPS workshop*, 2021.

[170] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[171] K. Zhou, C. C. Loy, and Z. Liu, "Semi-supervised domain generalization with stochastic stylematch," in *NeurIPS workshop*, 2021.

[172] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.

[173] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *NIPS*, 2014.

[174] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[175] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *CVPR Workshop*, 2018, pp. 783–790.

[176] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[177] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," *arXiv preprint arXiv:1711.01558*, 2017.

[178] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017.

[179] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[180] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," in *NIPS*, vol. 19, 2007.

[181] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[182] S. J. Pan, I. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE TNN*, vol. 22, pp. 199–210, 2011.

[183] F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-draa, "Domain generalization with optimal transport and metric learning," *ArXiv*, vol. abs/2007.10573, 2020.

[184] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019, pp. 1406–1415.

[185] R. Zhu and S. Li, "Self-supervised universal domain adaptation with adaptive memory separation," in *ICDM*, 2021.

[186] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018, pp. 464–479.

[187] J. Jia, Q. Ruan, and T. M. Hospedales, "Frustratingly easy person re-identification: Generalizing person re-id in practice," *BMVC*, 2019.

[188] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," in *NeurIPS*, 2018.

[189] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[190] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *ACMMM*, 2018, pp. 402–410.

[191] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM TIST*, vol. 11, no. 1, pp. 1–25, 2020.

[192] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016.

[193] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV*, 2016, pp. 443–450.

[194] X. Peng and K. Saenko, "Synthetic to real adaptation with generative correlation alignment networks," in *WACV*. IEEE, 2018, pp. 1982–1991.

[195] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *CVPR*, 2017, pp. 6924–6932.

[196] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *ICLR*, 2017.

[197] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[198] A. Sonar, V. Pacelli, and A. Majumdar, "Invariant policy optimization: Towards stronger generalization in reinforcement learning," in *Learning for Dynamics and Control*, 2021, pp. 21–33.

[199] H. Bui, T. Tran, A. T. Tran, and D. Phung, "Exploiting domain-specific features to enhance domain generalization," in *NeurIPS*, 2021.

[200] C. Kang and K. Nandakumar, "Dynamically decoding source domain knowledge for unseen domain generalization," *arXiv preprint arXiv:2110.03027*, 2021.

[201] C. Liu, L. Wang, K. Li, and Y. Fu, "Domain generalization via feature variation decorrelation," in *ACMMM*, 2021, pp. 1683–1691.

[202] Y. Wang, H. Li, L.-P. Chau, and A. C. Kot, "Variational disentanglement for domain generalization," *arXiv preprint arXiv:2109.05826*, 2021.

[203] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij, "On causal and anticausal learning," in *ICML*, 2012, pp. 1255–1262.

[204] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *ICML*, 2013, pp. 819–827.

[205] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *AAAI*, 2015.

[206] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich, "Causal generative domain adaptation networks," *arXiv preprint arXiv:1804.04333*, 2018.

[207] C. Heinze-Deml and N. Meinshausen, "Conditional variance penalties and domain shift robustness," *stat*, vol. 1050, p. 13, 2019.

[208] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[209] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.

[210] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *ICML*, 2016, pp. 1842–1850.

[211] Y. Zhao, Z. Zhong, F. Yang, Z. Luo, Y. Lin, S. Li, and N. Sebe, "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification," in *CVPR*, 2021, pp. 6277–6286.

[212] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[213] S. Sagawa, P. W. Koh, T. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *ICLR*, 2020.

[214] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE TPAMI*, 2020.

[215] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit, "On calibration and out-of-domain generalization," in *NeurIPS*, 2021.

[216] Y. Gong, X. Lin, Y. Yao, T. G. Dietterich, A. Divakaran, and M. Gervasio, "Confidence calibration for domain generalization under covariate shift," in *ICCV*, 2021.

[217] T. Duboudin, E. Dellandréa, C. Abgrall *et al.*, "Encouraging intra-class diversity through a reverse contrastive loss for better single-source domain generalization," in *ICCV*, 2021.

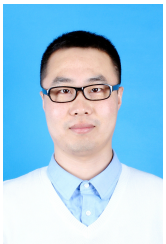
[218] L. Lin, H. Xie, Z. Yang, Z. Sun, W. Liu, Y. Yu, W. Chen, S. Yang, and D. Xie, "Semi-supervised domain generalization in real world: New benchmark and strong baseline," *arXiv preprint arXiv:2111.10221*, 2021.

[219] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning with domain generalization," *arXiv:2111.10487*, 2021.

[220] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *CVPR*, 2021, pp. 1013–1023.

[221] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D.

Cubuk, A. Kurakin *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020.
[222] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *NeurIPS*, 2021.
[223] Z. Zhang, Y. Li, J. Wang, B. Liu, D. Li, X. Chen, Y. Guo, and Y. Liu, "Remos: Reducing defect inheritance in transfer learning via relevant model slicing," in *44th International Conference on Software Engineering (ICSE)*, 2022.
[224] D. Mahajan, S. Tople, and A. Sharma, "The connection between out-of-distribution generalization and privacy of ml models," in *Workshop on Privacy Preserving Machine Learning at NeurIPS*, 2021.
[225] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.
[226] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.
[227] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020.
[228] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.
[229] M. Zhang, A. Zhang, and S. McDonagh, "On the out-of-distribution generalization of probabilistic image modelling," in *NeurIPS*, 2021.
[230] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *ECCV*, 2010, pp. 213–226.
[231] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *ICCV*, 2013, pp. 1657–1664.
[232] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," *CVPR*, pp. 5385–5394, 2017.
[233] S. Beery, G. V. Horn, and P. Perona, "Recognition in terra incognita," in *ECCV*, 2018.
[234] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," *ICCV*, pp. 1406–1415, 2019.
[235] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020.
[236] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. 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*, 2018.
[237] J. Taylor, B. Earnshaw, B. Mabey, M. Victors, and J. Yosinski, "Rxrx1: An image set for cellular morphological variation across many experimental batches." in *ICLR*, 2019.
[238] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," in *NeurIPS*, 2020.
[239] E. David, S. Madec, P. Sadeghi-Tehran *et al.*, "Global wheat head detection (gwhd) dataset: a large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods," *Plant Phenomics*, vol. 2020, 2020.
[240] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," in *WWW*, 2019.
[241] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *CVPR*, 2018.
[242] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature Communications*, 2020.
[243] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *EMNLP-IJCNLP*, 2019.
[244] S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang *et al.*, "Codexglue: A machine learning benchmark dataset for code understanding and generation," *NeurIPS*, 2021.
[245] V. Raychev, P. Bielik, and M. Vechev, "Probabilistic model for code with decision trees," *ACM SIGPLAN Notices*, 2016.
[246] R. R. Brinkman, M. Gasparetto, S.-J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith, "High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease," *Biology of Blood and Marrow Transplantation*, vol. 13, no. 6, pp. 691–700, 2007.
[247] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.
[248] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *ECCV*, 2016.
[249] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *CVPR*, 2016.
[250] Y. Wang, S. Liao, and L. Shao, "Surpassing real-world source training data: Random 3d characters for generalizable person re-identification," in *ACMMM*, 2020, pp. 3422–3430.
[251] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *CVPR*, 2019.
[252] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
[253] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.
[254] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *EMNLP*, 2006.
[255] J. Wang and W. Lu, "Deepdg: Deep domain generalization toolkit," https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG.
[256] J. Wang *et al.*, "Everything about transfer learning and domain adapation," http://transferlearning.xyz.
[257] G. Parascandolo, A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf, "Learning explanations that are hard to vary," in *ICLR*, 2021.
[258] C. Shui, B. Wang, and C. Gagné, "On the benefits of representation regularization in invariance based domain generalization," *Machine Learning*, 2021.
[259] H. Wang, H. He, and D. Katabi, "Continuously indexed domain adaptation," in *ICML*, 2020.
[260] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Sequential learning for domain generalization," in *ECCV*, 2020, pp. 603–619.
[261] M. Mancini, Z. Akata, E. Ricci, and B. Caputo, "Towards recognizing unseen categories in unseen domains," in *ECCV*, 2020.
[262] R. Zhu and S. Li, "Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization," in *ICLR*, 2021.
[263] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2018.
[264] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020.
[265] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
[266] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," *NeurIPS*, 2021.
[267] P. Pandey, M. Raman, S. Varambally, and P. AP, "Domain generalization via inference-time label-preserving target projections," in *CVPR*, 2021.



Jindong Wang is currently a researcher at Microsoft Research Asia, Beijing, China. He received his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2019. He was a visiting student in Hong Kong University of Science and Technology (HKUST) in 2018. His research interest mainly includes transfer learning, machine learning, data mining, and ubiquitous computing. He serves as the publicity co-chair of IJCAI'19 and session chair of ICDM'19. He is the reviewer or PC member of several leading journals and conferences such as TPAMI, TKDE, ICLR, ICML, NeurIPS, CVPR, etc.