Generalizing to Unseen Domains: A Survey on Domain Generalization

Jindong Wang[1*], **Cuiling Lan**[1], **Chang Liu**[1], **Yidong Ouyang**[2], **Tao Qin**[1]
[1]Microsoft Research, Beijing, China
[2]Central University of Finance and Economics, Beijing, China
{jindong.wang,culan,changliu,taoqin}@microsoft.com, 2017312311@email.cufe.edu.cn

Abstract

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 over the years. This paper presents the first review of recent advances in domain generalization. First, we provide a formal definition of domain generalization and discuss several related fields. We then categorize recent algorithms into three classes: data manipulation, representation learning, and learning strategy, and present some algorithms in detail for each category. Third, we introduce the commonly used datasets and applications. Finally, we summarize existing literature and present some potential research topics for the future.

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. 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 with the names of domain adaptation, meta-learning, transfer learning, covariate shift, and so on. In recent years, *Domain generalization (DG)* has received much attention. 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. 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 insights into the future research.

In this paper, we present the first survey on domain generalization to introduce its recent advances, with special focus on its formulations, algorithms, 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 some related areas.

2 Domain Generalization and Related Areas

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

Definition 2 (Domain generalization)**.** *In domain generalization problems, we are given M training (source) domains $\mathcal{S}_{train} = \{\mathcal{S}^i \mid i = 1, \cdots, M\}$ where $\mathcal{S}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i-th domain. The joint distributions between each two 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 $h : \mathcal{X} \to \mathcal{Y}$ from the M training domains to achieve a minimum prediction error on an* unseen *test domain $\mathcal{S}_{test}$ (i.e., $\mathcal{S}_{test}$ cannot be accessed in training and $P_{XY}^{test} \neq P_{XY}^i$ for $i \in \{1, \cdots, M\}$): $\min_h \mathbb{E}_{(\mathbf{x},y)\in\mathcal{S}_{test}}[\ell(h(\mathbf{x}), y)]$, where $\mathbb{E}$ is the expectation and $\ell(\cdot, \cdot)$ is the loss function.*

We briefly describe the related research areas.

Multi-task learning [Caruana, 1997] 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)

[*] Contact author. Corresponding to Tao Qin. Full version of this survey is at: https://arxiv.org/abs/2103.03097.

task. Particularly, multi-domain learning is a kind of multi-task learning, which trains on multiple domains to learn models for each original domain instead of new test domains.

Transfer learning [Pan and Yang, 2010; Zhuang *et al.*, 2020; Wang, 2018] 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.

Domain adaptation (DA) [Patel *et al.*, 2015] 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 in practical applications.

Meta-learning [Vilalta and Drissi, 2002] aims to learn the learning algorithm itself by learning from previous experience or tasks, i.e., learning-to-learn. While the tasks are different in meta-learning, in domain generalization, the learning tasks are the same. Meta-learning is a general learning strategy that can be used for DG [Li *et al.*, 2018a; Balaji *et al.*, 2018; Li *et al.*, 2019b; Du and others, 2020] by simulating the meta-train and meta-test tasks in training domains to enhance the performance for DG.

Lifelong Learning [Biesialska *et al.*, 2020], 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 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 Methodology

In this section, we introduce existing domain generalization methods in detail. As shown in Figure 1, 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, or explicitly feature alignment between domains to learn domain-invariant representations; b). *Feature disentanglement*, which tries to disentangle the features into domain-shared or domain-specific parts for better generalization.

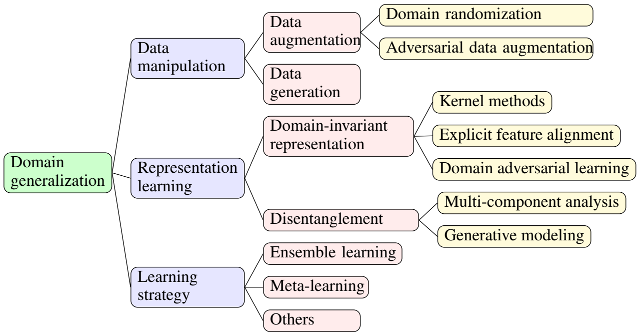

Figure 1: Taxonomy of domain generalization methods.

(3) **Learning strategy:** This category of methods focuses on exploiting the general learning strategy to promote the generalization capability, which mainly includes two kinds of methods: a). *Ensemble learning*, which relies on the power of ensemble to learn a unified 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. Additionally, there are other learning strategy that can also be used for DG and we categorized them as other learning strategy.

While these categories are conceptually different, they are complementary to each other and can be combined towards better performance. We will describe each category in detail.

3.1 Data Manipulation

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. Data manipulation-based DG is formulated as:

$$\min_{h} \mathbb{E}_{\mathbf{x},y}[\ell(h(\mathbf{x}),y)] + \mathbb{E}_{\mathbf{x}',y}[\ell(h(\mathbf{x}'),y)], \tag{1}$$

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

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. Without exception, they can also be adopted for DG where $\mathrm{mani}(\cdot)$ can be instantiated as these data augmentation functions.

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. As data becomes more complex and diverse, generalization ability can be increased. Here, the $\mathrm{mani}(\cdot)$ 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. [Khirodkar *et al.*, 2019] used domain randomization to generate more training data in simulated environments in order to generalize in real test environment. [Prakash and others, 2019] further took into account the structure of the scene when randomly placing objects for data generation, which enables the neural network to utilize context information when detecting objects.

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.*, 2018] 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.*, 2018] 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. Other than directly updating the inputs by gradient ascent, [Zhou *et al.*, 2020b] adversarially trained a data augmentation network to generate samples that can fool the feature extractor to eventually learn general representations.

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 $\mathrm{mani}(\cdot)$ can be implemented using some generative models such as Variational Auto-encoder (VAE) [Kingma and Welling, 2013], and Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014]. In addition, it can also be implemented using the Mixup [Zhang *et al.*, 2018] strategy.

[Rahman *et al.*, 2019] leveraged ComboGAN [Anoosheh *et al.*, 2018] to generate new data and then applied domain discrepancy measure such as MMD [Gretton *et al.*, 2012] to minimize the distribution divergence between real and generated images to learn general representations. [Qiao *et al.*, 2020] leveraged adversarial training to create “fictitious” yet “challenging” populations, where a Wasserstein Auto-Encoder [Tolstikhin *et al.*, 2017] was used to generate samples that preserve the semantics while having large domain transportation cost. [Zhou *et al.*, 2020c] generated novel distributions under semantic consistency and then maximized the difference between source and the novel distributions. [Somavarapu *et al.*, 2020] introduced a simple transformation based on image stylization to explore cross-source variability.

In addition to the above generative models, Mixup [Zhang *et al.*, 2018] 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 [Wang *et al.*, 2020d; Wang *et al.*, 2020e] to generate new samples; or in the feature space [Zhou *et al.*, 2021] which does not explicitly generate raw training samples. These methods achieved promising performance on popular benchmarks while remaining conceptually and computationally simple.

3.2 Representation Learning

Representation learning has always been the focus of machine learning for decades 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 is formulated as:

$$\min_{f,g} \mathbb{E}_{\mathbf{x},y} \ell(f(g(\mathbf{x})), y) + \lambda \ell_{\mathrm{reg}}, \tag{2}$$

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*.

Domain-invariant Representation-based DG

[Ben-David *et al.*, 2007] theoretically proved that if the feature representations remain invariant to different domains, the representations are general and transferable to different domains. Motivated by this theory, the learning objective for DG is to minimize the distribution gap within the training domains in order to learn domain-invariant features. Along this line, there are mainly three types of methods: kernel-based methods, domain adversarial learning, and explicit feature alignment.

Kernel-based methods. 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. For domain generalization, there are plenty of algorithms based on kernel methods, where the representation learning function g is implemented as some feature map $\phi(\cdot)$ which is easily computed using kernel function $k(\cdot, \cdot)$ such as RBF kernel and Laplacian kernel.

[Blanchard *et al.*, 2011] is the first work to use kernel method for DG which is extended later in [Blanchard *et al.*, 2017]. Their goal is to learn a domain-invariant positive semi-definite kernel for minimum risk control. [Grubinger *et al.*, 2015] adapted transfer component analysis (TCA) [Pan *et al.*, 2011] to bridge the multi-domain distributions gap for DG. Similar to TCA, Domain-Invariant Component Analysis

(DICA) [Muandet *et al.*, 2013] is one of the classic methods using kernel for DG. DICA finds a feature transformation kernel $k(\cdot,\cdot)$ that minimizes the distribution discrepancy between all data in feature space. [Gan *et al.*, 2016] adopted the similar method as DICA and further added attribute regularization. In contrast to DICA which deals with the marginal distribution, [Li *et al.*, 2018c] learned feature representations that remain domain-invariant in class-conditional distribution. [Ghifary *et al.*, 2016] used Fisher's discriminant analysis to minimize intra-class discrepancy while maximizing the inter-class discrepancy from different domains. They proposed Scatter Component Analysis (SCA) to learn domain-invariant representations using MMD [Gretton *et al.*, 2012], where the SCA kernel takes into account inter- and intra-class discrepancies. [Erfani *et al.*, 2016] 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.*, 2019] proposed multi-domain discriminant analysis to perform class-wise kernel learning for DG, which is more fine-grained. [Mahajan *et al.*, 2020] learned disentangled representations using causal matching.

Domain adversarial learning. Domain-adversarial training is widely used for learning domain-invariant features. [Ganin and Lempitsky, 2015; Ganin *et al.*, 2016] proposed Domain-adversarial neural network (DANN) for domain adaptation. In DANN, 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.*, 2018b] adopted such idea for DG. [Gong *et al.*, 2019] used adversarial training by gradually reducing the domain discrepancy in a manifold space. [Li *et al.*, 2018d] proposed a conditional invariant adversarial network (CIAN) to learn class-wise adversarial networks for DG. Similar ideas were also used in [Shao *et al.*, 2019; Rahman *et al.*, 2020; Wang *et al.*, 2020f]. [Jia *et al.*, 2020] 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. In addition to adversarial domain classification, [Zhao *et al.*, 2020a] 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 [Garg *et al.*, 2020] were also proposed with theoretically guaranteed generalization bound.

Explicit feature alignment. This line of works aligns the feature distributions across training domains to learn domain-invariant representations through explicit distribution alignment [Li *et al.*, 2018b; Zhou *et al.*, 2020a], or feature normalization [Jin *et al.*, 2020b]. [Motiian and others, 2017] 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) [Pan *et al.*, 2011; Wang *et al.*, 2018; Zhu *et al.*, 2020], second order correlation [Sun and Saenko, 2016], both mean and variance (moment matching), Wasserstein distance [Zhou *et al.*, 2020a], of domains for either domain adaptation or domain generalization. [Zhou *et al.*, 2020a] aligned the marginal distribution of different source domains via optimal transport by minimizing the Wasserstein distance.

[Jin *et al.*, 2020b; Jin *et al.*, 2021] proposed Style Normalization and Restitution (SNR) to simultaneously ensure both high generalization and discrimination capability of the networks. After the style normalization, 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. Restitution is extended to other alignment-based method to restorate discriminative information dropped by alignment [Jin *et al.*, 2020a].

Feature Disentanglement-based DG

Disentangled representation learning tries to learn a function that maps a sample to a feature vector that contains all the information about different factors of variation, with each dimension (or a subset of dimensions) containing information about only some factor(s). Disentanglement-based DG in general decomposes a feature representation into understandable compositions/sub-features, with one part being domain-shared/invariant part, and the other domain-shared one, which is formulated as:

$$\min_{g_c,g_s,f} \mathbb{E}_{\mathbf{x},y}\ell(f(g_c(\mathbf{x})),y)+\lambda\ell_{\mathrm{reg}}+\mu\ell_{\mathrm{recon}}([g_c(\mathbf{x}),g_s(\mathbf{x})],\mathbf{x}), \tag{3}$$

where g_c and g_s denote the domain-shared and domain-specific feature representations, respectively. λ,μ are trade-off 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 the loss of information. Note that $[g_c(\mathbf{x}),g_s(\mathbf{x})]$ denotes the combination/integration of two kinds of features (which is not limited to concatenation). Based on the choice of architectures and implementation mechanisms, the disentanglement-based DG can be categorized into two types: *multi-component analysis* and *generative modeling*.

Multi-component analysis. In multi-component analysis, the domain-shared and domain-specific features are in general extracted using the the corresponding network modules. UndoBias [Khosla *et al.*, 2012] started from a SVM model to maximize interval classification on all training data for domain generalization. It represented the parameters of the i-th domain as $\mathbf{w}_i = \mathbf{w}_0 + \Delta_i$, where $\mathbf{w}_0$ denotes the domain-shared parameters and Δ_i denotes the domain-specific parameters. Some other methods extented the idea of UndoBias from different aspects. [Niu *et al.*, 2015] proposed Multi-view DG (MVDG) using multi-view learning to learn the combination of exemplar SVMs under different views for robust generalization. [Ding and Fu, 2017] designed domain-specific networks for each domain and one shared domain-invariant network for all domains to learn disentangled representations. [Li *et al.*, 2017a] extended the idea of UndoBias into the neural network context and developed a low-rank parameterized CNN model for end-to-end training.

Generative modeling. Generative models can be used for disentanglement from the perspective of data generation process. They formulate the generative mechanism of the samples from the domain-level, sample-level and label-level. The Domain-invariant variational autoencoder (DIVA) [Ilse *et al.*, 2020] disentangled the features into domain information, category information, and other information, which is learned in the VAE framework. [Peng *et al.*, 2019] further disentangled the domain and class information using VAE. [Qiao *et al.*, 2020] also adopted VAE for disentanglement, where they proposed a Unified Feature Disentanglement Network (UFDN) that treated both data domains and image attributes of interest as latent factors to be disentangled. [Liu *et al.*, 2020] proposed a unified solution for both domain adaptation and domain generalization that used causality to disentangle the features with theoretical bounds.

3.3 Learning Strategy

In addition to data manipulation and representation learning, DG was also studied in general machine learning paradigms, which is divided into three categories: *ensemble learning-based DG*, *meta-learning-based DG*, and *others*.

Ensemble Learning-based DG

For domain generalization, ensemble learning exploits the structure or relations between several source domains by using certain learning architectures and training strategies for better generalization. Ensemble learning assumes that any sample can be treated as an integration of existing training domains, so the final prediction results are obtained as the superposition of the training data or models.

[Mancini *et al.*, 2018] proposed to use learnable weights for aggregating the predictions from different source specific classifiers, where a domain predictor is adopted to predict the probability of a sample belonging to each domain (*i.e.*, weights). [Segù *et al.*, 2020] maintained domain-dependent batch normalization (BN) statistics and BN parameters for different source domains while all the other parameters were shared. In inference, the final 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. [D'Innocente and Caputo, 2018] proposed domain-specific layers of different source domains and learning the linear aggregation of these layers to represent a test sample.

Meta-learning-based DG

Meta-learning [Vanschoren, 2018] is to learn a general model from multiple tasks by induction. To use meta-learning for DG, a general strategy is to divide the multi-source domains into meta-train set $\mathcal{S}_{mtrn}$ and meta-test set $\mathcal{S}_{mte}$ to simulate domain shift. Denote θ the model parameters to be learned, meta-learning can be formulated as:

$$\theta = \theta - \alpha \frac{\partial(\ell(\mathcal{S}_{mte};\theta) + \beta\ell(\mathcal{S}_{mtrn};\phi))}{\partial\theta}, \tag{4}$$

where ϕ is the parameters for meta-train task and α, β are learning rates for outer and inner loops, respectively.

Inspired by Model-agnostic meta-learning (MAML) [Finn *et al.*, 2017], [Li *et al.*, 2018a] proposed MLDG (Meta-Learning for DG), an optimization-based meta-learning strategy for domain generalization. MLDG splits the data from the source domains into meta-train and meta-test, where the model can gradually learn to adapt to the simulated test domains. [Balaji *et al.*, 2018] introduced a meta regularizer (MetaReg) in meta-learning for fine-grained regularization. [Li *et al.*, 2019b] proposed feature-critic training for the feature extractor by designing a meta optimizer learned in different domains. [Dou *et al.*, 2019] used the similar idea of MLDG and additionally introduced two complementary losses to explicitly regularize the semantic structure of feature space. [Du and others, 2020] proposed an extended version of information bottleneck named Meta Variational Information Bottleneck (MetaVIB). MetaVIB learns to minimize the Kullback–Leibler (KL) divergence between the latent distributions belonging to the same category. Recently, some works also adopted meta-learning for semi-supervised DG or discriminative DG [Chen *et al.*, 2020; Sharifi-Noghabi *et al.*, 2020; Wang *et al.*, 2020a; Zhao *et al.*, 2020b].

Other Learning Strategy

There are some other learning paradigms for DG. For instance, inspired by self-supervised learning that builds self-supervised tasks from unlabeled data [Jing and Tian, 2020], [Carlucci *et al.*, 2019] constructed a self-supervised task of solving jigsaw puzzles to learn generalized representations. [Li *et al.*, 2019a] learned feature extractors and classifiers using episodic training. First, they fix the classifier to learn the worst-case feature extractor; then, they fix the feature extractor to learn the worst-case classifier. [Huang *et al.*, 2020] used self-challenging to iteratively discarded the dominant features activated on the training data and forced the network to activate remaining features that correlate with labels.

4 Datasets and Applications

In this section, we briefly discuss the popular datasets and applications in DG. Image classification is the most popular benchmark and application for DG such as Rotated MNIST [Ghifary *et al.*, 2015], PACS [Li *et al.*, 2017a], VLCS [Fang *et al.*, 2013] and Office-Home [Venkateswara *et al.*, 2017] datasets.

There are other areas where DG also plays a key role. In computer vision, apart from image classification, satellite image classification [Deshmukh *et al.*, 2019] is also a popular application area for DG. Some works also used DG for semantic segmentation [Gong *et al.*, 2019], action recognition [Li *et al.*, 2017b; Li *et al.*, 2019a], face anti-spoofing [Shao *et al.*, 2019], person ReID [Wang *et al.*, 2020d; Jin *et al.*, 2020a], and street view recognition [Qiao *et al.*, 2020]. Also, there are some work for video understanding [Niu *et al.*, 2015].

In natural language processing, some work used DG for sentiment classification on Amazon Review dataset [Wang *et al.*, 2020f]. Others used DG for semantic parsing [Wang *et al.*, 2020a], web page classification [Garg *et al.*, 2020]. Medical analysis is one of the key application area for DG due to its nature of data scarcity. In this field, tissue segmentation [Dou *et al.*, 2019], Parkinson's disease [Muandet *et al.*, 2013], activity recognition [Erfani *et al.*, 2016], and chest X-ray recognition [Mahajan *et al.*, 2020; Li *et al.*, 2020].

Apart from those areas, DG was also used in reinforcement learning [Zhou *et al.*, 2021; Li *et al.*, 2018a] to generalize to unseen environment. Some work used DG to recognize speech utterance [Shankar *et al.*, 2018; Piratla *et al.*, 2020] and fault diagnosis [Li *et al.*, 2020]. It is clear that DG is being applied to more areas to learn models that can generalize well to unseen data.

5 Discussion

5.1 Summarization

The quantity and diversity of training data are of great importance to model's generalization ability. Many methods aim to enrich the training data using the aforementioned 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 [Ben-David *et al.*, 2007; Blanchard *et al.*, 2011]. 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 significant 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 [Blanchard *et al.*, 2011; Li *et al.*, 2018b] or conditional [Li *et al.*, 2018c] distributions. Thus, it is more feasible to perform dynamic distribution matching [Wang *et al.*, 2018; Wang *et al.*, 2020c].

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 [Ryu *et al.*, 2019] that used random forest for DG, and we hope more diverse methods will come.

5.2 Future Research Challenges

Continuous domain generalization. In many real applications, a system consumes streaming data with non-stationary statistics. In this case, it is critical to perform continuous domain generalization that efficiently updates DG models to overcome catastrophic forgetting and adapt to new data. While there are some continuous domain adaptation methods [Wang *et al.*, 2020b], there is only limited exploration on continuous DG and this is favorable in real scenarios.

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 [Maniyar *et al.*, 2020] proposed zero-shot DG and we expect more work will come in this area.

Interpretable domain generalization. Disentanglement-based DG methods decompose a feature to domain-invariant/shared and domain-specific 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. Causality [Liu *et al.*, 2020] may be one promising tool to understand domain generalization networks and provide interpretations.

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 [Devlin *et al.*, 2018] and GPT-3 [Brown *et al.*, 2020]. Pre-training on large-scale dataset and then finetuning the model to downstream tasks could improve its performance, where pre-training is beneficial to learn general representations. For instance, existing work that used the pre-trained representations on source domain can achieve competitive performance on domain adaptation tasks [Wang *et al.*, 2019]. Therefore, how to design useful and efficient DG methods to help large-scale pre-training/self-learning is worth investigating.

6 Conclusion

Generalization has always been an important research topic in machine learning research. In this article, we review the domain generalization area by providing in-depth analysis of existing methods, datasets and applications. Finally, we thoroughly analyze the methods and provide several potential research challenges. We hope that this survey can provide useful insights to interested researchers and inspire more progress in domain generalization and other research areas.

References

[Anoosheh *et al.*, 2018] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *CVPR Workshop*, pages 783–790, 2018.

[Balaji *et al.*, 2018] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 998–1008, 2018.

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. In *NIPS*, volume 19, page 137, 2007.

[Biesialska *et al.*, 2020] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. Continual lifelong learning in natural language processing: A survey. *arXiv preprint:2012.09823*, 2020.

[Blanchard *et al.*, 2011] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, pages 2178–2186, 2011.

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

[Brown *et al.*, 2020] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Carlucci *et al.*, 2019] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[Chen *et al.*, 2020] Keyu Chen, Di Zhuang, and J Morris Chang. Discriminative adversarial domain generalization with meta-learning based cross-domain validation. *arXiv preprint arXiv:2011.00444*, 2020.

[Deshmukh *et al.*, 2019] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization. *arXiv:1905.10392*, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[Dou *et al.*, 2019] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019.

[Du and others, 2020] Yingjun Du et al. Learning to learn with variational information bottleneck for domain generalization. In *ECCV*, 2020.

[D'Innocente and Caputo, 2018] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, pages 187–198. Springer, 2018.

[Erfani *et al.*, 2016] Sarah Erfani, Mahsa Baktashmotlagh, Masoud Moshtaghi, Vinh Nguyen, Christopher Leckie, James Bailey, and Ramamohanarao Kotagiri. Robust domain generalisation by enforcing distribution invariance. In *AAAI*, pages 1455–1461, 2016.

[Fang *et al.*, 2013] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013.

[Finn *et al.*, 2017] Chelsea Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

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

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[Ganin *et al.*, 2016] Yaroslav Ganin, E. Ustinova, Hana Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[Garg *et al.*, 2020] Vikas K Garg, Adam Kalai, Katrina Ligett, and Zhiwei Steven Wu. Learn to expect the unexpected: Probably approximately correct domain generalization. *arXiv preprint arXiv:2002.05660*, 2020.

[Ghifary *et al.*, 2015] Muhammad Ghifary, W. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *ICCV*, pages 2551–2559, 2015.

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

[Gong *et al.*, 2019] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, pages 2477–2486, 2019.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

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

[Hu *et al.*, 2019] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *UAI*, volume 35, 2019.

[Huang *et al.*, 2020] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, volume 2, 2020.

[Ilse *et al.*, 2020] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, 2020.

[Jia *et al.*, 2020] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8484–8493, 2020.

[Jin *et al.*, 2020a] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. In *NeurIPS*, 2020.

[Jin *et al.*, 2020b] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020.

[Jin *et al.*, 2021] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Style normalization and restitution for domain generalization and adaptation. *arXiv preprint arXiv:2101.00588*, 2021.

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

[Khirodkar *et al.*, 2019] Rawal Khirodkar, Donghyun Yoo, and Kris Kitani. Domain randomization for scene-specific car detection and pose estimation. In *WACV*, 2019.

[Khosla *et al.*, 2012] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[Li *et al.*, 2017a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[Li *et al.*, 2017b] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain generalization and adaptation using low rank exemplar svms. *IEEE TPAMI*, 40(5):1114–1127, 2017.

[Li *et al.*, 2018a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[Li *et al.*, 2018b] Haoliang Li, Sinno Jialin Pan, S. Wang, and A. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2018c] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018.

[Li *et al.*, 2018d] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018.

[Li *et al.*, 2019a] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *CVPR*, pages 1446–1455, 2019.

[Li *et al.*, 2019b] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, 2019.

[Li *et al.*, 2020] Xiang Li, Wei Zhang, Hui Ma, Zhong Luo, and Xu Li. Domain generalization in rotating machinery fault diagnostics using deep neural networks. *Neurocomputing*, 403:409–420, 2020.

[Liu *et al.*, 2020] Chang Liu, Xinwei Sun, Jindong Wang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *arXiv preprint arXiv:2011.01681*, 2020.

[Mahajan *et al.*, 2020] Divyat Mahajan, S. Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML Workshop*, 2020.

[Mancini *et al.*, 2018] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *ICIP*, pages 1353–1357, 2018.

[Maniyar *et al.*, 2020] Udit Maniyar, Aniket Anand Deshmukh, Urun Dogan, Vineeth N Balasubramanian, et al. Zero shot domain generalization. *arXiv preprint arXiv:2008.07443*, 2020.

[Motiian and others, 2017] Saeid Motiian et al. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[Niu *et al.*, 2015] Li Niu, Wen Li, and Dong Xu. Multi-view domain generalization for visual recognition. In *ICCV*, pages 4193–4201, 2015.

[Pan and Yang, 2010] S. Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22:1345–1359, 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, I. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE TNN*, 22:199–210, 2011.

[Patel *et al.*, 2015] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal process mag*, 32(3):53–69, 2015.

[Peng *et al.*, 2019] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *ICML*, 2019.

[Piratla *et al.*, 2020] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020.

[Prakash and others, 2019] Aayush Prakash et al. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *ICRA*, pages 7249–7255, 2019.

[Qiao *et al.*, 2020] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, pages 12556–12565, 2020.

[Rahman *et al.*, 2019] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *WACV*, pages 579–588. IEEE, 2019.

[Rahman *et al.*, 2020] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan.

Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020.

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

[Segù *et al.*, 2020] Mattia Segù, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv:2011.12672*, 2020.

[Shankar *et al.*, 2018] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[Shao *et al.*, 2019] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019.

[Sharifi-Noghabi *et al.*, 2020] Hossein Sharifi-Noghabi, Hossein Asghari, Nazanin Mehrasa, and Martin Ester. Domain generalization via semi-supervised meta learning. *arXiv:2009.12658*, 2020.

[Somavarapu *et al.*, 2020] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv:2006.11207*, 2020.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[Tolstikhin *et al.*, 2017] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[Vanschoren, 2018] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. *CVPR*, pages 5385–5394, 2017.

[Vilalta and Drissi, 2002] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

[Volpi *et al.*, 2018] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5334–5344, 2018.

[Wang *et al.*, 2018] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *ACMMM*, pages 402–410, 2018.

[Wang *et al.*, 2019] Jindong Wang, Yiqiang Chen, Han Yu, Meiyu Huang, and Qiang Yang. Easy transfer learning by exploiting intra-domain structures. In *ICME*, pages 1210–1215, 2019.

[Wang *et al.*, 2020a] Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. *arXiv:2010.11988*, 2020.

[Wang *et al.*, 2020b] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. *arXiv preprint arXiv:2007.01807*, 2020.

[Wang *et al.*, 2020c] Jindong Wang, Yiqiang Chen, Wenjie Feng, Han Yu, Meiyu Huang, and Qiang Yang. Transfer learning with dynamic distribution adaptation. *ACM TIST*, 11(1):1–25, 2020.

[Wang *et al.*, 2020d] Wenhao Wang, Shengcai Liao, Fang Zhao, Cuicui Kang, and Ling Shao. Domainmix: Learning generalizable person re-identification without human annotations. *arXiv:2011.11953*, 2020.

[Wang *et al.*, 2020e] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP*, pages 3622–3626, 2020.

[Wang *et al.*, 2020f] Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. Unseen target stance detection with adversarial domain generalization. In *IJCNN*, pages 1–8, 2020.

[Wang, 2018] Jindong Wang. Everything about transfer learning and domain adapation. https://github.com/jindongwang/transferlearning, 2018.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[Zhao *et al.*, 2020a] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In *NeurIPS*, volume 33, 2020.

[Zhao *et al.*, 2020b] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. *arXiv preprint arXiv:2012.00417*, 2020.

[Zhou *et al.*, 2020a] Fan Zhou, Zhuqing Jiang, Changjian Shui, B. Wang, and B. Chaib-draa. Domain generalization with optimal transport and metric learning. *ArXiv*, abs/2007.10573, 2020.

[Zhou *et al.*, 2020b] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020.

[Zhou *et al.*, 2020c] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020.

[Zhou *et al.*, 2021] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.

[Zhu *et al.*, 2020] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE TNNLS*, 2020.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.