LEARNING CAUSAL SEMANTIC REPRESENTATION FOR OUT-OF-DISTRIBUTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional supervised learning methods, especially deep ones, are found to be sensitive to out-of-distribution (OOD) examples, largely because the learned representation mixes the semantic factor with the variation factor due to their domain-specific correlation, while only the semantic factor *causes* the output. To address the problem, we propose a Causal Semantic Generative model (CSG) based on causality to model the two factors separately, and learn it on a single training domain for prediction without (OOD generalization) or with (domain adaptation) unsupervised data in a test domain. We prove that CSG identifies the semantic factor on the training domain, and the invariance principle of causality subsequently guarantees the boundedness of OOD generalization error and the success of adaptation. We design learning methods for both effective learning and easy prediction, by leveraging the graphical structure of CSG. Empirical study demonstrates the effect of our methods to improve test accuracy for OOD generalization and domain adaptation.

1 INTRODUCTION

Deep learning has initiated a new era of artificial intelligence where the potential of machine learning models is greatly unleashed. Despite the great success, these methods heavily rely on the independently-and-identically-distributed (IID) assumption. This does not always perfectly hold in practice, and the prediction of output (label, response, outcome) y may be saliently affected in out-of-distribution (OOD) cases, even from an essentially irrelevant change to the input (covariate) x, like a position shift or rotation of the object in an image, or a change of background, illumination or style (Shen et al., 2018; He et al., 2019; Arjovsky et al., 2019). These phenomena pose serious concerns on the robustness and trustworthiness of machine learning methods and severely impede them from risk-sensitive scenarios.

Looking into the problem, although deep learning models allow extracting abstract representation for prediction with their powerful approximation capacity, the representation may be overconfident in the correlation between semantic factors s (*e.g.*, shape of an object) and variation factors v (*e.g.*, background, illumination, object position). The correlation may be domain-specific and spurious, and may change drastically in a new environment. So it has become a desire to learn representation that separates semantics s from variations v (Cai et al., 2019; Ilse et al., 2019). Formally, the importance of this goal is that s represents the *cause* of y. Causal relations better reflect the fundamental mechanisms of nature, bringing the merit to machine learning that they tend to be universal and *invariant* across domains (Peters et al., 2017; Schölkopf, 2019), thus providing the most transferable and confident information to unseen domains. Causality has also been shown to lead to proper domain adaptation (Schölkopf et al., 2012; Zhang et al., 2013), and lower the adaptation cost and mitigate catastrophic forgetting (Bengio et al., 2019; Ke et al., 2019).

In this work, we propose a Causal Semantic Generative model (CSG) for robust and proper OOD prediction, including OOD generalization (test data unavailable during training) and domain adaptation (unsupervised test data available during training) based on a single-domain supervised training dataset. **(1)** We explicitly separate the latent variables into a (group of) semantic variable s and a (group of) variation variable v, and prove under proper conditions that CSG can identify the semantic factor as s by fitting the training dataset, even in presence of an s-v correlation. **(2)** By leveraging the causal invariance, we prove that a well-learned CSG is guaranteed to have a bounded OOD generalization error. The bound shows how causal mechanisms affect the error. **(3)** We de-

velop a domain adaptation method using CSG and causal invariance, which suggests to fix the causal generative mechanisms and adapt the prior to the new domain. We prove the identification of the new prior and the benefit of adaptation. **(4)** To learn and adapt the model, we design a novel reformulation of the conventional Evidence Lower BOund (ELBO) objective, so that the learning-stage inference model can also serve for prediction, and the inference model on the training domain can be avoided in domain adaptation. To our best knowledge, our work is the *first* to identify semantic factor and leverage latent causal invariance for OOD prediction. Empirical improvement in OOD performance and adaptation is demonstrated by experiments on multiple tasks including shifted MNIST and ImageCLEF-DA task.

2 Related work

There have been works that aim to leverage the merit of causality for OOD prediction. For OOD generalization, some works ameliorate discriminative models towards a causal behavior. Bahadori et al. (2017) introduce a regularizer that reweights input dimensions based on their approximated causal effects to the output, and Shen et al. (2018) reweight training samples by amortizing causal effects among input dimensions. They are extended to nonlinear cases (Bahadori et al., 2017; He et al., 2019) by learning a representation linear to the output. These methods introduce no additional modeling effort, but may also be limited to capture invariant causal mechanisms (being non-generative) and may only behave quantitatively causal in the training domain.

For domain adaptation/generalization, methods are developed under various causal assumptions (Schölkopf et al., 2012; Zhang et al., 2013) or using learned causal relations (Rojas-Carulla et al., 2018; Magliacane et al., 2018). The considered causality is among directly observed variables, which may not be suitable for general data like image pixels where causality rather lies between data and conceptual latent factors (Lopez-Paz et al., 2017; Besserve et al., 2018; Teshima et al., 2020). Their extensions (Zhang et al., 2013; Gong et al., 2016; 2018) to allow certain ways of mechanism shift are not constrained by causal invariance. Domain-invariant-representation-based domain adaptation (Pan et al., 2010; Baktashmotlagh et al., 2013; Ganin et al., 2016; Long et al., 2015; 2018) and generalization (Muandet et al., 2013; Shankar et al., 2018) learn a latent representation and enforce its distribution to be shared across domains. Remarkable results have been achieved. Nevertheless, Johansson et al. (2019); Zhao et al. (2019) point out that this invariance is neither sufficient nor necessary to identify the true semantics or lower the adaptation error (Supplement C). Moreover, the invariance (including that by Arjovsky et al. (2019) for domain generalization) considers the inference direction (*i.e.*, data $\rightarrow$ representation) opposite to the causal invariance, and may not be as general (Sec. 3.2). For generative methods, Cai et al. (2019); Ilse et al. (2019) assume the independence between the semantics and a domain-specific factor in both generative and inference models (similarly for the output and manipulation variable by Zhang & Li (2020)), which may not meet reality closely. They also do not adapt the prior for domain shift thus degenerate to inference invariance. Teshima et al. (2020) rely on causal invariance and adapt the prior, while also assume latent independence and do not separate the semantic factor. Their deterministic and invertible mechanism also indicates inference invariance. Moreover, most domain generalization methods require multiple training domains, with exceptions (*e.g.*, Qiao et al. (2020)) that still seek to augment domains (where causal invariance is not considered). In contrast, CSG leverages causal invariance that covers more than inference invariance, and has guarantee to identify the semantic factor on a single training domain, even with a correlation to the variation factor.

Generative supervised learning is not new (Mcauliffe & Blei, 2008; Kingma et al., 2014), but most works do not consider the encoded causality. Other works consider solving causality tasks, notably causal/treatment effect estimation (Louizos et al., 2017; Yao et al., 2018; Wang & Blei, 2019). The task does not focus on OOD prediction, and requires labels for both treated and controlled groups.

Disentangling latent representations is also of interest in unsupervised learning. Despite some empirical success (Chen et al., 2016; Higgins et al., 2017; Chen et al., 2018), Locatello et al. (2019) conclude that it is impossible to guarantee the disentanglement in unsupervised settings. Khemakhem et al. (2019; 2020) show an encouraging result that disentangled representation can be identified up to a permutation with a cause of the latent variable observed. But the methods cannot separate the semantic factor from variation for supervised learning, and require sufficient different values of the cause variable, making it hard to leverage labels.

Causality with latent variable has been considered in a rich literature (Verma & Pearl, 1991; Spirtes et al., 2000; Richardson et al., 2002; Hoyer et al., 2008; Shpitser et al., 2014), while most works focus on the consequence on observation-level causality. Others consider identifying the latent variable. Janzing et al. (2009) and Lee et al. (2019) show the identifiability up to a reparameterization under additive noise assumption or other directed causal discovery assumptions. For discrete data, a "simple" latent variable can be identified under various specifications (Janzing et al., 2011; Sgouritsa et al., 2013; Kocaoglu et al., 2018). Other works consider identification via interventional datasets (Romeijn & Williamson, 2018). Over these works, we step further to separate and identify the latent variable as semantic and variation factors, and show the benefit for OOD prediction.

3 THE CAUSAL SEMANTIC GENERATIVE MODEL

To develop the model seriously and soberly based on causality, we require the formal definition of causality: *two variables have a causal relation if externally intervening the* cause *(by changing variables out of the considered system) may change the* effect*, but not* vice versa (Pearl, 2009; Peters et al., 2017). We then follow the logic below to build our model.

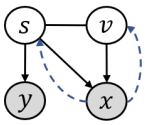

Figure 1: The graphical structure of the proposed Causal Semantic Generative model (CSG) for the semantic s and variation v latent factors and supervised data (x, y). Black solid arrows represent invariant causal generating mechanisms $p(x|s, v)$ and $p(y|s)$, the black undirected edge represent a domain-specific prior $p(s, v)$, and blue dashed bended arrows represent the inference model $q(s, v|x)$ for learning and prediction.

(1) The relation between dimensions of x and y may not be causal generally (*e.g.*, *intervening an image by changing the background during imaging does not change its label (object identity),* which excludes $x \to y$, and *intervening the label by malfunctioning a labeling machine or reforming the labeling policy does not change the input,* which excludes $y \to x$). So we employ a generative model, and introduce a latent variable z to capture factors with causal relations.

(2) The latent variable z as underlying generating factors (*e.g.*, *shape, color and texture of an object, background and illumination in imaging*) is plausible to be the cause of both x (*e.g.*, *the change of object shape or background makes a different image, but intervening an image by breaking a camera sensor unit does not change the shape or background*) and y (*e.g.*, *changing the shape of a bed to that of a desk alters the label, but intervening the label as above does not change the object shape*). So we orient the edges in the generative direction $z \to (x, y)$ (see also Teshima et al. (2020)). This is in contrast to graphs by Cai et al. (2019); Ilse et al. (2019; 2020); Castro et al. (2020) where y is treated as the cause of a semantic factor. When y is also a noisy or summative observation, this would make unreasonable implications (*e.g.*, *intervening the label changes the object shape, and changing the shape does not change* y).

(3) We attribute all x-y relation to the existence of some latent factors ((Janzing et al., 2009); "purely common cause" (Lee et al., 2019)), and exclude x-y edges. This can be achieved as long as z holds sufficient information of data (*e.g.*, *breaking a sensor unit or intervening the label while locking shape, background* etc. *does not change the label or the image*). Promoting this restriction reduces arbitrariness in explaining x-y relation and enforces the effectiveness of z. This is in contrast to graphs by Kingma et al. (2014); Zhang & Li (2020) or assumptions by Schölkopf et al. (2012); Zhang et al. (2013); Castro et al. (2020) where y is treated as the cause of x since no latent variable is introduced between.

(4) Not all latent factors are the causes of y (*e.g.*, *changing object shape may alter the label, while changing the background does not*). We thus split the latent variable as $z = (s, v)$ and remove the edge $v \to y$, where s represents the *semantic factor* of x that causes y, and v describes the *variation* or diversity of x samples with the same semantics.

(5) The variation v often has a relation to the semantic factor s, which is often a spurious correlation (*e.g.*, *desks prefer a workspace background, but they can also appear in bedrooms and beds can also appear in workspace*). So we keep the undirected s-v edge. Although v is not a cause of y, modeling it explicitly is worth the effort since otherwise it would still be implicitly incorporated in s anyway through the s-v correlation. We summarize these conclusions in the following definition.

Definition 3.1 (CSG)**.** A *Causal Semantic Generative Model* (CSG) $p = (p(x|s, v), p(y|s), p(s, v))$ is a generative model on data variables $x \in \mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ and $y \in \mathcal{Y}$ with semantic $s \in \mathcal{S} \subset \mathbb{R}^{d_\mathcal{S}}$ and variation $v \in \mathcal{V} \subset \mathbb{R}^{d_\mathcal{V}}$ latent variables, following the graphical structure shown in Fig. 1.

3.1 THE CAUSAL INVARIANCE PRINCIPLE

The domain-invariance of causal relations translates to the following principle for CSG:

Principle 3.2 (causal invariance)**.** The causal generative mechanisms $p(x|s,v)$ and $p(y|s)$ in CSG are invariant across domains, and the change of prior $p(s,v)$ is the only source of domain shift.

The invariance is supported by the invariance of basic laws of nature (Besserve et al., 2018; Bühlmann, 2018; Schölkopf, 2019; Teshima et al., 2020). Other works instead introduce domain index (Cai et al., 2019; Ilse et al., 2019; 2020; Castro et al., 2020) or manipulation variables (Zhang & Li, 2020; Khemakhem et al., 2019; 2020) to model distribution change explicitly. They require multiple training domains or additional observations, and such changes can also be explained under causal invariance as long as the latent variable includes all shifted factors (*e.g.*, *domain change of images can be attributed to a different preference of shape, style, texture, background,* etc. *and their correlations, while the processes generating image and label from them remain the same*).

3.2 COMPARISON WITH INFERENCE INVARIANCE

Domain-invariant-representation-based domain adaptation and generalization[1], and invariant risk minimization (Arjovsky et al., 2019) for domain generalization, assume the invariance of the process in the other direction, *i.e.* inferring the latent representation from data. We note that in its supportive examples (*e.g.*, *inferring the object position from an image or extracting the fundamental frequency from a vocal audio*), generating mechanisms are nearly deterministic and invertible, so that the posterior is almost determined by the inverse function, and causal invariance implies inference invariance. For noisy or degenerate mechanisms, ambiguity occurs during inference (Fig. 2), and the inferred result notably relies on the prior. Since the prior may change across domains, the inference process may not be invariant anymore, and causal invariance is more reliable.

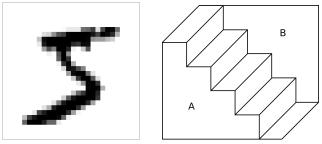

Figure 2: Examples of noisy (left) or degenerate (right) generating mechanisms that lead to ambiguity in inference. Left: handwritten digit that may be "3" or "5". Right: Schröder's stairs that may have A or B being the nearer surface.

4 METHOD

We develop methods for learning, adapting and using a CSG based on the causal invariance principle 3.2, and devise implementation using variational Bayes.

4.1 METHOD FOR OOD GENERALIZATION

When test data come sequentially or adaptation cost is unaffordable, a CSG p gives prediction directly using $\mathbb{E}[y|x]$ derived from causal mechanisms $p(x|s,v), p(y|s)$ and prior $p(s,v)$, which are learned by fitting supervised training data from an underlying data distribution $p^*(x,y)$. Fitting data by maximizing likelihood $\mathbb{E}_{p^*(x,y)}[\log p(x,y)]$ is intractable since $p(x,y)$ is hard to estimate effectively. We thus adopt the conventional Evidence Lower BOund (ELBO) $\mathcal{L}_{q,p}(x,y) := \mathbb{E}_{q(s,v|x,y)}[\log \frac{p(s,v,x,y)}{q(s,v|x,y)}]$ (Jordan et al., 1999; Wainwright et al., 2008) as a tractable surrogate, which requires an auxiliary inference model $q(s,v|x,y)$ to estimate the expectation effectively. Maximizing $\mathcal{L}_{q,p}$ w.r.t q drives q towards the posterior $p(s,v|x,y)$ and meanwhile makes $\mathcal{L}_{q,p}$ a tighter lower bound of $\log p(x,y)$. The expected ELBO $\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{q,p}(x,y)]$ then drives $p(x,y)$ towards $p^*(x,y)$.

However, the subtlety with supervised learning is that after fitting data, evaluating $p(y|x)$ for prediction is still hard. We thus propose to employ a model for $q(s,v,y|x)$ instead. The required inference model can be then expressed as $q(s,v|x,y) = q(s,v,y|x)/q(y|x)$ where $q(y|x) = \int q(s,v,y|x)\,\mathrm{d}s\mathrm{d}v$. It reformulates the expected ELBO as:

$$\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{q,p}(x,y)] = \mathbb{E}_{p^*(x)}\mathbb{E}_{p^*(y|x)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\mathbb{E}_{q(s,v,y|x)}\Big[\frac{p^*(y|x)}{q(y|x)}\log\frac{p(s,v,x,y)}{q(s,v,y|x)}\Big]. \quad (1)$$

The first term is the common cross entropy loss (negative) driving $q(y|x)$ towards $p^*(y|x)$. Once this is achieved, the second term becomes the expected ELBO $\mathbb{E}_{p^*(x)}[\mathcal{L}_{q(s,v,y|x),p}(x)]$ that drives $q(s,v,y|x)$ towards $p(s,v,y|x)$ (and $p(x)$ towards $p^*(x)$). Since the target $p(s,v,y|x)$ admits the factorization $p(s,v|x)p(y|s)$ (since $(v,x) \perp\!\!\!\perp y|s$) where $p(y|s)$ is already given by the CSG, we can

[1] In addition to the invariance of the marginal distribution of representation that they require.

further ease the modeling of $q(s, v, y|x)$ as $q(s, v|x)p(y|s)$. The objective then becomes:

$$\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{q,p}(x, y)] = \mathbb{E}_{p^*(x,y)}\left[\log q(y|x) + \frac{1}{q(y|x)}\mathbb{E}_{q(s,v|x)}\Big[p(y|s)\log\frac{p(s, v, x)}{q(s, v|x)}\Big]\right], \quad (2)$$

where $q(y|x) = \mathbb{E}_{q(s,v|x)}[p(y|s)]$. The expectations can be estimated by Monte Carlo, and their gradients can be estimated using the reparameterization trick (Kingma & Welling, 2014). When well optimized, $q(s, v|x)$ well approximates $p(s, v|x)$, so $q(y|x)$ then well approximates $p(y|x) = \mathbb{E}_{p(s,v|x)}[p(y|s)]$ for prediction.

CSG-ind To mitigate the spurious s-v correlation in training and let the model stay neutral in test, we also consider using a CSG with an independent prior $p^{\perp\!\!\!\perp}(s, v) := p(s)p(v)$ for prediction, where $p(s)$ and $p(v)$ are the marginals of the training domain prior $p(s, v)$. The corresponding inference model $q^{\perp\!\!\!\perp}(s, v|x)$ is used for prediction, but $q(s, v|x)$ for the correlated prior $p(s, v)$ is also required for learning. We propose to save the cost of learning two inference models, and use $q^{\perp\!\!\!\perp}(s, v|x)$ to represent $q(s, v|x)$. Note that their targets are related by $p(s, v|x) = \frac{p(v|s)}{p(v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p^{\perp\!\!\!\perp}(s, v|x)$, we thus formulate $q(s, v|x) = \frac{p(v|s)}{p(v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}q^{\perp\!\!\!\perp}(s, v|x)$, so that this $q(s, v|x)$ achieves its goal once $q^{\perp\!\!\!\perp}(s, v|x)$ does. The objective is turned into:

$$\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{q,p}(x, y)] = \mathbb{E}_{p^*(x,y)}\left[\log \pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\Big[\frac{p(v|s)}{p(v)}p(y|s)\log\frac{p^{\perp\!\!\!\perp}(s, v, x)}{q^{\perp\!\!\!\perp}(s, v|x)}\Big]\right], \quad (3)$$

where $\pi(y|x) := \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\big[\frac{p(v|s)}{p(v)}p(y|s)\big]$. Prediction is given by $\mathbb{E}^{\perp\!\!\!\perp}[y|x] \approx \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}[p(y|s)]$.

4.2 Method for Domain Adaptation

When unsupervised test data is available from an underlying data distribution $\tilde{p}^*(x)$, we can leverage it for adaptation. According to the causal invariance principle 3.2, we only need to adapt the prior $\tilde{p}(s, v)$ and the corresponding inference model $\tilde{q}(s, v|x)$ to the test domain, while the causal mechanisms $p(x|s, v), p(y|s)$ are shared in both domains. Adaptation is done by fitting the test data, which can be achieved by maximizing the conventional ELBO:

$$\mathbb{E}_{\tilde{p}^*(x)}[\mathcal{L}_{\tilde{q},\tilde{p}}(x)] = \mathbb{E}_{\tilde{p}^*(x)}\mathbb{E}_{\tilde{q}(s,v|x)}\left[\log\Big(\tilde{p}(s, v)p(x|s, v)/\tilde{q}(s, v|x)\Big)\right]. \quad (4)$$

Prediction is given by $\tilde{\mathbb{E}}[y|x] \approx \mathbb{E}_{\tilde{q}(s,v|x)}[p(y|s)]$. Similar to the case of CSG-ind, we need $\tilde{q}(s, v|x)$ for prediction, but $q(s, v|x)$ is also required for learning on the training domain. When data from both domains are available during learning, we can save the effort of learning $q(s, v|x)$ using a similar technique. We formulate $q(s, v|x) = \frac{\tilde{p}(x)}{p(x)}\frac{p(s,v)}{\tilde{p}(s,v)}\tilde{q}(s, v|x)$ using $\tilde{q}(s, v|x)$ following the relation between their targets, and the expected ELBO on the training domain becomes:

$$\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{q,p}(x, y)] = \mathbb{E}_{p^*(x,y)}\left[\log \pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{\tilde{q}(s,v|x)}\Big[\frac{p(s, v)}{\tilde{p}(s, v)}p(y|s)\log\frac{\tilde{p}(s, v)p(x|s, v)}{\tilde{q}(s, v|x)}\Big]\right], \quad (5)$$

where $\pi(y|x) := \mathbb{E}_{\tilde{q}(s,v|x)}\big[\frac{p(s,v)}{\tilde{p}(s,v)}p(y|s)\big]$. The total objective in this case is its sum with Eq. (4).

In each of the three cases, only one inference model is required for the respective prediction task. Supplement D.2 shows its implementation from a general discriminative model (*e.g.*, how to select its hidden nodes as s and v). Supplement D.1 details the deduction of the equations. In practice, the second term of Eq. (1) may be scales larger than the first term, raising difficulty in matching $p^*(y|x)$. So we include an additional cross entropy loss in the objective.

5 Theory

We now establish guarantee for the methods on identifying the semantic factor and the subsequent merits for OOD generalization and domain adaptation. We only consider the infinite-data regime to isolate another source of error from finite data. Supplement A shows all the proofs. Identifiability is hard to achieve for latent variable models (Koopmans & Reiersol, 1950; Murphy, 2012; Yacoby et al., 2019; Locatello et al., 2019), since it is a task beyond modeling observational relations (Janzing et al., 2009; Peters et al., 2017). Assumptions are required to draw definite conclusions.

Assumption 5.1 (additive noise)**.** There exist nonlinear functions f and g with bounded derivatives up to third-order, and independent random variables μ and ν, such that $p(x|s, v) = p_\mu(x - f(s, v))$, and $p(y|s) = p_\nu(y - g(s))$ for continuous y or $p(y|s) = \mathrm{Cat}(g(s))$ for categorical y.

This structure disables describing a bivariate joint distribution in both generating directions (Zhang & Hyvärinen (2009, Thm. 8), Peters et al. (2014, Prop. 23)), and is widely adopted in directed causal discovery (Janzing et al., 2009; Bühlmann et al., 2014). CSG needs this since it should make the causal direction exclusive. It is also easy to implement with deep models (Kingma & Welling, 2014), so does not essentially restrict model capacity.

Assumption 5.2 (bijectivity)**.** Function f is bijective and g is injective.

It is a common (Janzing et al., 2009; Shalit et al., 2017; Khemakhem et al., 2019; Lee et al., 2019) sufficient condition for the fundamental (Peters et al., 2014, Prop. 7) requirement of causal minimality (Peters et al., 2014; 2017) for identifiability. Particularly, s and v would otherwise have dummy dimensions that f and g simply ignore. According to the commonly acknowledged manifold hypothesis (Weinberger & Saul, 2006; Fefferman et al., 2016) that data tends to lie on a lower-dimensional manifold in the data space, we can take $\mathcal{X}$ or $\mathcal{Y}$ as the manifold and such bijections exist as coordinate maps, which are injections to the data space (allowing $d_{\mathcal{S}} + d_{\mathcal{Z}} < d_{\mathcal{X}}$).

5.1 Identifiability Theory

We first formalize the goal of identifying the semantic factor.

Definition 5.3 (semantic-equivalence)**.** We say two CSGs p^* and p are *semantic-equivalent*, if there exists a measurable homeomorphism[2] Φ on $\mathcal{S} \times \mathcal{V}$, such that its output dimensions in $\mathcal{S}$, *i.e.* $\Phi^{\mathcal{S}}(s, v)$, is constant of v, and it converts the two models: $\Phi_{\#}[p^*_{s,v}] = p_{s,v}$, $p^*(x|s, v) = p(x|\Phi(s, v))$ and $p^*(y|s) = p(y|\Phi^{\mathcal{S}}(s))$.

It is an equivalent relation if $\mathcal{V}$ is connected and is either open or closed in $\mathbb{R}^{d_{\mathcal{V}}}$ (Supplement A.1). Here, $\Phi_{\#}[p^*_{s,v}]$ denotes the pushed-forward distribution by Φ, *i.e.* the distribution of $\Phi(s, v)$ when $(s, v) \sim p^*_{s,v}$. Such a Φ acts as a *reparameterization* since the two CSGs can be mutually converted by applying Φ to their latent variables (s, v). At the heart of the definition, the v-constancy implies that one model *does not mix* the other's v into its s, so that the s variables of both models hold equivalent information.

A learned CSG p *identifies* the semantic factor if it is semantic-equivalent to the ground-truth CSG p^*. This identification cannot be characterized by the *statistical independence* between s and v (as in Cai et al. (2019); Ilse et al. (2019); Zhang & Li (2020)), which is not sufficient (Locatello et al., 2019) nor necessary (due to spurious correlation). Also, the identification does not guarantee the learned v to be constant of ground-truth s, while *disentanglement* requires that a semantic transformation on x only changes the learned s (Higgins et al., 2018).

To identify the semantics, the ground-truth model could at most provide its information via the data distribution $p^*(x, y)$. Although semantic-equivalent CSGs induce the same distribution on (x, y), the inverse is nontrivial. The following theorem shows the semantic-identifiability under appropriate conditions.

Theorem 5.4 (semantic-identifiability)**.** *With Assumptions (5.1, 5.2), a well-learned CSG p with $p(x, y) = p^*(x, y)$ is semantic-equivalent to the ground-truth CSG p^*, under the conditions that $\log p(s, v)$ and $\log p^*(s, v)$ have bounded derivatives up to second-order, and that* **(i)**[3] $\frac{1}{\sigma_\mu^2} \to \infty$, *where σ_μ^2 is the variance in the principal component of p_μ, or* **(ii)** *p_μ has an a.e. non-zero characteristic function (*e.g.*, a Gaussian distribution).*

Remarks. **(1)** The requirement on $p^*(s, v)$ excludes *extreme* training data with a deterministic s-v relation, which makes $p^*(s, v)$ unbounded and discontinuous. In that case (*e.g.*, *all desks appear in workspace and all beds in bedrooms*), one cannot tell whether the label is caused by *shape* or *background*.

(2) In condition **(i)**, $\frac{1}{\sigma_\mu^2}$ measures the *intensity* of the causal mechanism $p(x|s, v)$. A strong $p(x|s, v)$ helps disambiguating values of (s, v) in generating a given x. The condition makes $p(x|s, v)$ so strong that it is almost deterministic and invertible, so inference invariance also holds (Sec. 3.2). Supplement A.2 provides a quantitative reference of large intensity for a practical consideration, and Supplement B gives a non-asymptotic extension showing how the intensity trades-off the tolerance

[2] A transformation is a homeomorphism if it is a continuous bijection with continuous inverse.

[3] As an asymptotic statement, the conclusion is that the equalities in Def. 5.3 hold in the limit.

of equalities in Def. 5.3. Condition **(ii)** covers more than inference invariance. It roughly implies that different values of (s, v) a.s. produce different distributions $p(x|s, v)$ on $\mathcal{X}$, so their roles in generating x become clear which helps identification.

(3) The theorem does not contradict the impossibility result by Locatello et al. (2019), which considers disentangling each latent dimension with an unconstrained $(s, v) \to (x, y)$, while we identify s as a whole with the edge $v \to y$ removed which breaks the s-v symmetry.

5.2 OOD Generalization Theory

The causal invariance principle 3.2 forms the ground-truth CSG on the test domain as $\tilde{p}^* = (\tilde{p}^*(s, v), p^*(x|s, v), p^*(y|s))$ with the new ground-truth prior $\tilde{p}^*(s, v)$, which gives the optimal predictor $\tilde{\mathbb{E}}^*[y|x]$. The principle also leads to the invariance of identified causal mechanisms, which shows that the OOD generalization error of CSG is bounded:

Theorem 5.5 (OOD generalization error)**.** *With Assumptions (5.1, 5.2), for a semantically-identified CSG p, we have up to $O(\sigma_\mu^2)$ that for $\tilde{p}^*(x)$-a.e. x,*

$$\left|\mathbb{E}[y|x] - \tilde{\mathbb{E}}^*[y|x]\right| \leqslant \sigma_\mu^2 \left\|J_{f^{-1}}(x)\right\|_2^2 \|\nabla g(s)\|_2 \|\nabla \log(p(s, v)/\tilde{p}(s, v))\|_2 \Big|_{(s,v)=f^{-1}(x)},$$

where $J_{f^{-1}}$ is the Jacobian matrix of f^{-1}, and $\tilde{p}_{s,v} := \Phi_\#[\tilde{p}^*_{s,v}]$ is the test-domain prior under the parameterization of the identified model.

The result shows that when the causal mechanism $p(x|s, v)$ is strong, especially in the extreme case $\sigma_\mu = 0$ where inference invariance also holds, it dominates prediction over the prior and the generalization error diminishes. In more general cases where only causal invariance holds, the prior change deviates the prediction rule. The prior-change term $\|\nabla \log(p(s, v)/\tilde{p}(s, v))\|_2$ measures the *hardness* or severity of OOD. It diminishes in IID cases, and makes the bound ineffective when the two priors do not share their support. Using a CSG to fit training data enforces causal invariance and other assumptions, so its $\mathbb{E}[y|x]$ behaves more faithfully in low $p^*(x)$ area and the boundedness becomes more plausible in practice. CSG-ind actively uses an independent prior whose larger support covers more $\tilde{p}_{s,v}$ candidates.

5.3 Domain Adaptation Theory

In cases of weak causal mechanism or violent prior change, the new ground-truth prior $p^*_{s,v}$ is important for prediction. The domain adaptation method learns a new prior $\tilde{p}_{s,v}$ by fitting unsupervised test-domain data, with causal mechanisms shared. Once the mechanisms are identified, $p^*_{s,v}$ can also be identified under the learned parameterization, and prediction can be made precise.

Theorem 5.6 (domain adaptation error)**.** *Under conditions of Thm. 5.4, for a semantically-identified CSG p on the training domain with reparameterization Φ, if the new prior $\tilde{p}_{s,v}$ is well-learned with $\tilde{p}(x) = \tilde{p}^*(x)$, then $\tilde{p}_{s,v} = \Phi_\#[\tilde{p}^*_{s,v}]$, and $\tilde{\mathbb{E}}[y|x] = \tilde{\mathbb{E}}^*[y|x]$ for $\tilde{p}^*(x)$-a.e. x.*

Different from existing domain adaptation bounds (Supplement C), Theorems (5.5, 5.6) allow different inference models in the two domains, thus go beyond inference invariance.

6 Experiments

For baselines of OOD generalization, apart from the conventional supervised learning optimizing cross entropy (CE), we also consider a causal discriminative method CNBB (He et al., 2019), and a generative method supervised VAE (sVAE) which is a counterpart of CSG that does not separate its latent variable into s and v. For domain adaptation, we consider well-acknowledged DANN (Ganin et al., 2016), DAN (Long et al., 2015), CDAN (Long et al., 2018) and MDD (Zhang et al., 2019) methods implemented in the `dalib` package (Jiang et al., 2020), and also sVAE using a similar method as CSG-DA. All methods share the same optimization setup. We align the scale of the CE term in the objectives of all methods, and tune their parameters to lie on the margin that makes converging training accuracy near 1, so that they wield the most power on the test domain while be faithful to explicit supervision. See Supplement E for detailed settings.

6.1 Shifted MNIST

We consider an OOD prediction task on MNIST to classify digits "0" and "1". In training data, "0"s are horizontally shifted at random by δ pixels with $\delta \sim \mathcal{N}(-5, 1^2)$, and "1"s by $\delta \sim \mathcal{N}(5, 1^2)$ pixels. We consider two test domains where the digits are not moved, or are shifted $\delta \sim \mathcal{N}(0, 2^2)$ pixels. Both domains have balanced classes. We implement all methods using multilayer perceptron which

Table 1: Accuracy (%) of various methods (ours in bold) on OOD generalization (top) and domain adaptation (bottom) for shifted MNIST. Averaged over 10 runs.

shift	CE	CNBB	sVAE	**CSG**	**CSG-ind**
$= 0$	53.3 ± 8.8	74.4 ± 6.2	60.5 ± 13.9	90.9 ± 6.8	$\mathbf{94.5 \pm 4.5}$
$\sim \mathcal{N}(0, 2^2)$	52.7 ± 2.8	58.0 ± 1.7	58.6 ± 5.8	64.8 ± 2.7	$\mathbf{66.5 \pm 4.7}$

shift	DANN	DAN	CDAN	MDD	sVAE-DA	**CSG-DA**
$= 0$	91.6 ± 5.5	11.3 ± 11.5	60.1 ± 41.4	57.4 ± 3.9	50.3 ± 3.9	$\mathbf{95.2 \pm 11.9}$
$\sim \mathcal{N}(0, 2^2)$	47.0 ± 2.9	50.1 ± 2.9	49.2 ± 5.6	52.2 ± 1.6	68.0 ± 7.5	$\mathbf{76.0 \pm 3.4}$

Table 2: Accuracy (%) of various methods (ours in bold) on OOD generalization (left) and domain adaptation (right) on the **C**→**P** task of ImageCLEF-DA (ima). Results of CE, DANN, DAN and CDAN are taken from Long et al. (2018). Averaged over 10 runs.

CE	CNBB	sVAE	**CSG**	**CSG-ind**	DANN	DAN	CDAN	sVAE-DA	**CSG-DA**
65.5±0.3	72.7±1.1	70.4±0.8	73.1±0.5	**73.4±0.7**	74.3±0.5	69.2±0.4	**74.5±0.3**	73.7±0.5	74.1±0.5

does not naturally exhibit shift invariance. We use a larger architecture for discriminative methods (including domain adaptation methods) to compensate the additional generative components of generative methods.

The OOD performance is shown in Table 1. For OOD generalization, CSG gives more genuine predictions in unseen domains, thanks to the identification of the semantic factor. CSG-ind performs even better, demonstrating the merit of approaching a CSG with an independent prior. Other methods are more significantly misled by the position factor from the spurious correlation. CNBB ameliorates the position bias, but not as thoroughly without explicit structures for causal mechanisms. CSG also outperforms sVAE, showing the benefit of separating semantics and variation and modeling the variation explicitly, so the model could consciously drive semantic representation into s. For domain adaptation, existing methods differ a lot, and are hard to perform well on both test domains. When fail to identify, adaptation sometimes even worsens the result, as the misleading representation based on position gets strengthened on the unsupervised test data. CSG is benefited from adaptation by leveraging test data in a proper way that identifies the semantics.

6.2 IMAGECLEF-DA

ImageCLEF-DA (ima) is a standard benchmark dataset for the ImageCLEF 2014 domain adaptation challenge. We select an adaptation task from two of its domains: from Caltech-256 (**C**) to Pascal VOC 2012 (**P**). Each domain has 12 classes and 600 images that follows different distributions compared to another domain. We use the ResNet50 structure (He et al., 2016) pretrained on ImageNet as the backbone of the discriminative/inference model. We adopt the same setup as in Long et al. (2018).

Table 2 shows the results. We see that CSG(-ind) achieves the best OOD generalization results, and performs comparable with modern domain adaptation methods. On this task, the underlying causal mechanism may be very noisy (*e.g.*, photos taken from inside and outside both count for the aircraft class), making identification hard. So our methods do not have a salient improvement over existing methods.

7 CONCLUSION AND DISCUSSION

We tackle OOD generalization and domain adaptation tasks by proposing a Causal Semantic Generative model (CSG), which builds upon a causal reasoning, and models semantic and variation factors separately while allowing their correlation. Using the invariance principle of causality, we develop effective and delicate methods for learning, adaptation and prediction, and prove the identification of the semantic factor, the boundedness of OOD generalization error, and the success of adaptation under appropriate conditions. Experiments show the improved performance in both tasks.

The consideration of separating semantics from variation extends to broader examples regarding robustness. Convolutional neural networks are found to change its prediction under a different texture but the same shape (Geirhos et al., 2019; Brendel & Bethge, 2019). Adversarial vulnerability (Szegedy et al., 2014; Goodfellow et al., 2015; Kurakin et al., 2016) extends variation factors to human-imperceptible features, *i.e.* the adversarial noise, which is shown to have a strong spurious correlation with semantics (Ilyas et al., 2019). The separation also matters for fairness when a sensitive variation factor may change prediction due to a spurious correlation. Our methods are potentially beneficial in these examples.

REFERENCES

The imageclef-da challenge 2014. https://www.imageclef.org/2014.

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

Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769–776, 2013.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010b.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Michel Besserve, Naji Shajarisales, Bernhard Schölkopf, and Dominik Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 557–565. PMLR, 2018.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

Peter Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Conference of IJCAI*, volume 2019, pp. 2060. NIH Public Access, 2019.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1050–1059, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pp. 2839–2848, 2016.

Mingming Gong, Kun Zhang, Biwei Huang, Clark Glymour, Dacheng Tao, and Kayhan Batmanghelich. Causal generative domain adaptation networks. *arXiv preprint arXiv:1804.04333*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yue He, Zheyan Shen, and Peng Cui. Towards non-i.i.d. image classification: A dataset and baselines. *arXiv preprint arXiv:1906.02899*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. *arXiv preprint arXiv:1905.10427*, 2019.

Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Designing data augmentation for simulating interventions. *arXiv preprint arXiv:2005.01856*, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

D Janzing, J Peters, JM Mooij, and B Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pp. 249–257. AUAI Press, 2009.

D Janzing, E Sgouritsa, O Stegle, J Peters, and B Schölkopf. Detecting low-complexity unobserved causes. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 383–391. AUAI Press, 2011.

Junguang Jiang, Bo Fu, and Mingsheng Long. Transfer-learning-library. https://github.com/thuml/Transfer-Learning-Library, 2020.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536, 2019.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P Kingma, and Aapo Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models. *arXiv preprint arXiv:2002.11537*, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, Banff, Canada, 2014. ICLR Committee.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Murat Kocaoglu, Sanjay Shakkottai, Alexandros G Dimakis, Constantine Caramanis, and Sriram Vishwanath. Entropic latent variable discovery. *arXiv preprint arXiv:1807.10399*, 2018.

Tjalling C Koopmans and Olav Reiersol. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Ciarán M Lee, Christopher Hart, Jonathan G Richens, and Saurabh Johri. Leveraging directed causal discovery to detect latent common causes. *arXiv preprint arXiv:1910.10174*, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6979–6987, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pp. 10846–10856, 2018.

Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pp. 121–128, Vancouver, Canada, 2008. NIPS Foundation.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Radford M Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12556–12565, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Thomas Richardson, Peter Spirtes, et al. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Jan-Willem Romeijn and Jon Williamson. Intervention and identifiability in latent variable modelling. *Minds and machines*, 28(2):243–264, 2018.

B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *International Conference on Machine Learning (ICML 2012)*, pp. 1255–1262. International Machine Learning Society, 2012.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Eleni Sgouritsa, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pp. 556–575. AUAI Press, 2013.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR.org, 2017.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *Proceedings of the International Conference on Learning Representations (ICLR 2018)*, 2018.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 411–419. ACM, 2018.

Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, 2014.

Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. *arXiv preprint arXiv:2002.03497*, 2020.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

Yaniv Yacoby, Weiwei Pan, and Finale Doshi-Velez. Learning deep bayesian latent variable regression models that generalize: When non-identifiability is a problem. *arXiv preprint arXiv:1911.00569*, 2019.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2633–2643, 2018.

Cheng Zhang and Yingzhen Li. A causal view on robustness of neural networks. *ICLR 2020 submission*, 2020. URL `https://openreview.net/forum?id=HkxvlOEtDH`.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pp. 647–655. AUAI Press, 2009.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

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

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532, 2019.