Learning Causal Semantic Representation for Out-of-Distribution Prediction

Chang Liu[1]*, **Xinwei Sun**[1], **Jindong Wang**[1], **Haoyue Tang**[2]†, **Tao Li**[3]†,
Tao Qin[1], **Wei Chen**[1], **Tie-Yan Liu**[1]
[1] Microsoft Research Asia, Beijing, 100080.
[2] Tsinghua University, Beijing, 100084. [3] Peking University, Beijing, 100871.

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 a causal reasoning so that the two factors are modeled separately, and develop methods for OOD prediction from a *single* training domain, which is common and challenging. The methods are based on the causal invariance principle, with a novel design in variational Bayes for both efficient learning and easy prediction. Theoretically, we prove that under certain conditions, CSG can identify the semantic factor by fitting training data, and this semantic-identification guarantees the boundedness of OOD generalization error and the success of adaptation. Empirical study shows improved OOD performance over prevailing baselines.

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 assumption that data from training and test domains follow the same distribution (*i.e.*, the IID assumption), while in practice the test domain is often out-of-distribution (OOD), meaning that the test data distribute differently from the training data. Popular models for predicting the output (or label, response, outcome) y from the input (or covariate) x have been found erroneous when confronted with a distribution change, even from an essentially irrelevant perturbation like a position shift or background change for images [91, 6, 102, 41, 2, 27]. 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 unconsciously mix up semantic factors s (*e.g.*, shape of an object) and variation factors v (*e.g.*, background, object position) due to a correlation between them (*e.g.*, desks often appear in a workspace background and beds in bedrooms), so the model also relies on the variation factors v for prediction via this correlation. However, this correlation tends to be superficial and spurious (*e.g.*, a desk can also appear in a bedroom, but this does not make it a bed), and may change drastically in a new domain, making the effect from v misleading. So it is desired to learn a representation that identifies s against v.

Formally, the essence of this goal is to leverage *causal relations* for prediction, since the fundamental distinction between s and v is that only s is the cause of y. Causal relations better reflect basic

*Correspondence to: Chang Liu <changliu@microsoft.com>.
†Work done during an internship at Microsoft Research Asia.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

mechanisms of nature. They bring the merit to machine learning that they tend to be universal and *invariant* across domains [97, 87, 93, 77, 16, 96, 98], thus provide the most transferable and reliable information to unseen domains. This causal invariance has been shown to lead to proper domain adaptation [97, 123], lower adaptation cost and lighter catastrophic forgetting [87, 9, 56].

In this work, we propose a Causal Semantic Generative model (CSG) following a causal consideration to separately model the semantic (cause of prediction) and variation latent factors, and develop OOD prediction methods with theoretical guarantees on identifiability and the boundedness of OOD prediction error. Addressing the complaint that OOD prediction and causality methods often require multi-domain or intervention data, we focus on the most common and also challenging tasks where only one *single* training domain is available, including *OOD generalization* and *domain adaptation*, where in the latter, unsupervised test-domain data are additionally available for training. The methods and theory are based on the causal invariance principle, which suggests to share generative mechanisms across domains, while the latent factor distribution (*i.e.*, the prior $p(s, v)$) changes. We argue that this causal invariance is more reliable than *inference invariance* in the other direction adopted by many existing methods [33, 101, 2, 66, 79]. For our method, we design novel and delicate reformulations of the ELBO objective so that we avoid the cost to build and learn two inference models. Theoretically, we prove that under certain conditions, CSG *can identify* the semantic factor on the single training domain, even in presence of an s-v correlation. We further prove the merits from this identification: prediction error is bounded for OOD generalization, and for domain adaptation, the test-domain prior is identifiable which leads to an accurate prediction. To sum up our contributions,

- Up to our knowledge, we are the first to show a theoretical guarantee (under appropriate conditions) to identify the latent cause of prediction (*i.e.*, the semantic factor) on a single training domain, and also the first to show the theoretical benefits of this identification for OOD prediction. The results also contribute to generative representation learning for revealing what is learned.
- We develop effective methods for OOD generalization and domain adaptation, and achieve mostly better performance than prevailing methods on real-world image classification tasks.

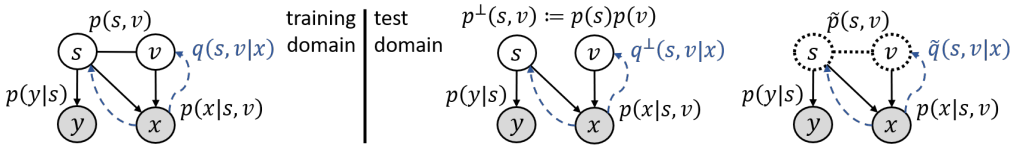
2 Related Work

OOD generalization with causality. There are trials that ameliorate discriminative models towards a causal behavior. Bahadori et al. [4] introduce a regularizer that reweights input dimensions based on their approximated causal effects to the output, and Shen et al. [102] reweight training samples by amortizing causal effects among input samples. Their linear input-output assumption is then extended [4, 41] by learning a representation. Some recent works require identity data (finer than label) and enforce inference invariance via variance minimization [42], or leverage a strong domain knowledge to augment images as an independent intervention on variation factors [79]. These methods introduce no additional generative modeling efforts, at the cost of limited capacity for invariant causal mechanisms.

Domain adaptation/generalization with causality. There are methods developed under various causal assumptions [97, 123] or using learned causal relations [93, 77]. Zhang et al. [123], Gong et al. [35, 36] also consider certain ways of mechanism change. The considered causality is among directly observed variables, which may not well suit general data like image pixels where causality rather lies in the conceptual latent level [75, 10, 59].

To consider latent factors, there are domain adaptation [83, 5, 33, 73, 74] and generalization methods [80, 101, 113] that learn a representation with a domain-invariant marginal distribution. Remarkable results have been achieved. Nevertheless, it is found that this invariance is neither sufficient nor necessary to identify the true semantics or lower the adaptation error ([54, 125]; see also Appx. E). Moreover, these methods are based on inference invariance, which may not be as reliable as causal invariance (see Sec. 3.2).

There are also generative methods for domain adaptation/generalization that model latent factors. Cai et al. [18] and Ilse et al. [49] introduce a semantic factor and a domain-feature factor. They assume the two latent factors are independent in both generative and inference models, which is unrealistic. Correlated factors are then considered [3]. But all these works do not adapt the prior for domain change thus resort to inference invariance. Zhang et al. [121] consider a partially observed manipulation variable, while still assuming its independence from the output in both the joint and posterior, and the adaptation is inconsistent with causal invariance. The above methods also do not show guarantees to identify their latent factors. Teshima et al. [108] leverage causal invariance and



(a) CSG (b) CSG-ind (c) CSG-DA

Figure 1: **(a)** Graphical structure of the proposed CSG. Solid arrows represent causal mechanisms $p(x|s,v)$ and $p(y|s)$, the undirected s-v clique represents a domain-specific prior $p(s,v)$, and the dashed bended arrows represent the inference model $q(s,v|x)$ for learning. **(b, c)** Graphical structures of CSG-ind and CSG-DA for prediction on the *test domain*. An independent prior $p^{\perp\!\!\!\perp}(s,v)$ (constructed from $p(s,v)$) and a new prior $\tilde{p}(s,v)$ (the dotted s-v clique) are introduced reflecting the intervention on the test domain. Respective inference models $q^{\perp\!\!\!\perp}(s,v|x)$ and $\tilde{q}(s,v|x)$ are also shown. All three models share the same causal mechanisms $p(x|s,v)$ and $p(y|s)$.

adapt the prior, yet also assume latent independence and do not separate the semantic factor. They require some supervised test-domain data, and their deterministic and invertible mechanism also indicates inference invariance. In addition, most domain generalization methods require *multiple* training domains, with exceptions [89] that still seek to augment domains. In contrast, CSG leverages causal invariance, and has *guarantee* to identify the semantic factor from a *single training domain*, even with a *correlation* to the variation factor.

Disentangled latent representations is also of interest in unsupervised learning. Despite empirical success [22, 43, 21], Locatello et al. [70] conclude that it is impossible to guarantee the disentanglement in unsupervised settings. Subsequent works then introduce ways of supervision like a few latent variable observations [71] or sample similarity [20, 72, 104]. Identifiable VAE [57] and extensions [58, 117] leverage the data of a cause variable of the latent variables and have established theoretical guarantees under a diversity condition. But the works do not depict domain change thus not suitable for OOD prediction. Instead of disentangling latent factors, we focus on identifying the semantic factor s (Sec. 5.1) and its benefit for OOD prediction. Appx. D shows more related work.

3 The Causal Semantic Generative Model

To develop the model soberly based on causality, we require its formal definition: *two variables have a causal relation, denoted as "*cause $\rightarrow$ effect*", if intervening the* cause *(by changing external variables out of the considered system) may change the* effect*, but not* vice versa [85, 88]. We follow this definition to build our model (Fig. 1a) by analyzing the example that an photographer takes a photo in a scene as x and labels it as y. Appx. C provides more explanations under other perspectives.

(1) It is likely that neither $y \rightarrow x$ (*e.g.*, intervening the label with noise by distracting the photographer does not change the image) nor $x \rightarrow y$ holds (*e.g.*, intervening an image by breaking a camera sensor unit does not change how the photographer labels it), as also argued in [88, Sec. 1.4; 59]. So we introduce a latent variable z to capture factors with causal relations. Also for this reason, we need a generative model (vs. discriminative model that only learns $x \rightarrow y$).

(2) The latent variable z as underlying generating factors (*e.g.*, object shape and texture, background and illumination during imaging) is plausible to cause both x (*e.g.*, changing object shape or background makes a different image, but breaking the camera does not change the shape or background) and y (*e.g.*, the photographer would give a different label if the object shape had been different, but noise-corrupting the label does not change the shape). So we orient the edges in the generative direction $z \rightarrow (x,y)$, as also adopted in [78, 88, 108]. This is in contrast to prior works [18, 49, 48, 19] that treat y as the cause of a semantic factor, which, when y is also a noisy observation, makes unreasonable implications (*e.g.*, adding noise to the labels in a dataset automatically changes object features and consequently the images, and changing the object features does not change the label). This difference is also discussed in [88, Sec. 1.4; 59].

(3) We attribute all x-y relation to the existence of some latent factor [68, "purely common cause"; 51] and exclude x-y edges. This can be achieved as long as z holds sufficient information of data (*e.g.*, with shape, background *etc.* fixed, breaking the camera does not change the label, and noise-corrupting the label does not change the image). Promoting this structure reduces arbitrariness in explaining x-y relation thus helps identify (part of) z. This is in contrast to prior works [63, 121, 19] that treat y as a cause of x as no latent variable is introduced between.

(4) Not all latent factors are the causes of y (*e.g.*, changing the shape may alter the label, while changing the background does not). We thus split the latent variable as $z = (s, v)$ and remove the $v \to y$ edge, where s represents the *semantic* factor that causes y, and v describes the *variation* or diversity in generating x. This formalizes the intuition on the concepts in Introduction (Sec. 1).

(5) The two factors s and v often have a relation (*e.g.*, a desk/bed shape tends to appear with a workspace/bedroom background), but it is usually a spurious correlation (*e.g.*, putting a desk in a bedroom does not automatically change the room as a workspace, nor does it turn the desk into a bed). So we keep the undirected s-v edge. This is in contrast to prior works [18, 49, 121, 108, 79] which assume independent latent variables. Although v is not a cause of y, modeling it explicitly is worth the effort since otherwise it would still be implicitly mixed into s anyway through the s-v correlation. We summarize these conclusions in the following definition.

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

3.1 The Causal Invariance Principle

Through the above process, we see that the s-v correlation embodied in the prior $p(s, v)$ tends to change across domains. Under a causal view, this means that the domain change comes from a (soft) intervention on s or v or both, leading to a different prior. On the other hand, the generative processes are likely causal mechanisms, so they enjoy the celebrated Independent Causal Mechanisms principle [88, 98] indicating that they are unaffected under the intervention on prior. This leads to the following causal invariance principle for CSG.

Principle 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 change.

This invariance reflects the universality of basic laws of nature and is considered in some prior works [97, 88, 10, 16]. Other works instead introduce domain index [18, 49, 48, 19] or manipulation variables [121, 57, 58] to model distribution change explicitly. They then require multiple training domains or additional observations, while such changes can also be explained under causal invariance as long as the latent variables include all changing factors.

3.2 Comparison with Inference Invariance

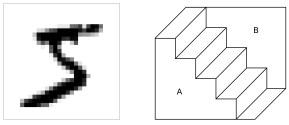


Figure 2: Examples of noisy (left) or degenerate (right) generating mechanisms that lead to ambiguity in inference. Left: handwritten digit that may be generated as either "3" or "5". Right: Schröder's stairs that may be generated with either A or B being the nearer surface. Inference results notably rely on the prior on the digits/surfaces, which is domain-specific.

Most domain adaptation and generalization methods (incl. domain-invariant-representation based [33, 101], invariant-latent-predictor based [2, 66, 79]) use a shared representation extractor across domains. This effectively assumes the invariance in the other direction, *i.e.* inferring latent factors z from observed data x. We note in its supportive examples (*e.g.*, inferring object position from image, extracting the fundamental frequency from audio), the causal mechanism $p(x|z)$ is nearly deterministic and invertible such that it preserves the information of z. Formally, for a given x, only one single z value achieves a positive $p(x|z)$ while all other values lead to zero. The inferred representation given by the posterior via the Bayes rule $p(z|x) \propto p(z)p(x|z)$ then concentrates on this z value, which is determined by the causal mechanism $p(x|z)$ alone, regardless of the domain-specific prior $p(z)$. Causal invariance then implies inference invariance.

In more general cases, the causal mechanism may be noisy or degenerate (Fig. 2), such that there are multiple z values that give a positive $p(x|z)$, *i.e.* they all could generate the same x. Inference is then ambiguous, and the posterior relies on the prior to choose from these z values. Since the prior changes across domains (*e.g.*, different labelers have different mindset), the inference rule then *changes by nature* and is not invariant,[3] while the causal invariance is rather more fundamental and reliable. To leverage causal invariance, we use a different prior for the test domain (CSG-ind and CSG-DA), which gives a different and more reliable prediction than following inference invariance.

[3]Particularly, although Mitrovic et al. [79] consider a similar causal structure and promote the invariance of $p(y|s)$, s actually depends on v for a given x, even when they are independent in the prior. So $p(s|x)$ must depend on the domain-specific $p(v)$, and a domain-invariant representation extractor does not exist.

4 Method

We now develop methods based on variational Bayes [55, 62] for OOD generalization and domain adaptation using CSG. Appx. F.1 shows all details.

4.1 Method for OOD Generalization

For OOD generalization, one only has supervised data from the underlying data distribution $p^*(x, y)$ on the *training domain*. Fitting a CSG $p := \langle p(s, v), p(x|s, v), p(y|s)\rangle$ to data by maximizing likelihood $\mathbb{E}_{p^*(x,y)}[\log p(x, y)]$ is intractable, since $p(x, y) := \int p(s, v, x, y)\,\mathrm{d}s\mathrm{d}v$ where $p(s, v, x, y) := p(s, v)p(x|s, v)p(y|s)$, is hard to estimate. The Evidence Lower BOund (ELBO) $\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y) := \mathbb{E}_{q(s,v|x,y)}[\log \frac{p(s,v,x,y)}{q(s,v|x,y)}]$ [55, 112] is a tractable surrogate with the help of an inference model $q(s, v|x, y)$ that enjoys easy sampling and density evaluation. It is known that $\max_{q_{s,v|x,y}} \mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)$ drives $q(s, v|x, y)$ towards the posterior $p(s, v|x, y) := \frac{p(s,v,x,y)}{p(x,y)}$, meanwhile makes $\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)$ a tighter lower bound of $\log p(x, y)$ for optimizing CSG p.

However, the subtlety with supervised learning is that prediction is still hard, as the introduced model $q(s, v|x, y)$ does not help estimate $p(y|x)$. To address this, we propose to employ an auxiliary model $q(s, v, y|x)$ targeting $p(s, v, y|x)$. It allows easy sampling of y given x for prediction, and can also serve as the required inference model: $q(s, v|x, y) = \frac{q(s,v,y|x)}{q(y|x)}$, where $q(y|x) := \int q(s, v, y|x)\,\mathrm{d}s\mathrm{d}v$ is also determined by $q(s, v, y|x)$. The ELBO objective $\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)]$ then becomes:

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

As a functional of $q(s, v, y|x)$ (instead of $q(s, v|x, y)$) and the CSG p, this objective also drives them towards their targets: the first term is the negative of the standard cross entropy (CE) loss which drives $q(y|x)$ towards $p^*(y|x)$, and once this is achieved, the second term becomes the expected ELBO $\mathbb{E}_{p^*(x)}[\mathcal{L}_{p,\,q_{s,v,y|x}}(x)]$ that drives $q(s, v, y|x)$ towards $p(s, v, y|x)$ and $p(x)$ towards $p^*(x)$. Furthermore, as the target of $q(s, v, y|x)$ factorizes as $p(s, v, y|x) = p(s, v|x)p(y|s)$ (due to Fig. 1a) where $p(y|s)$ is already known (part of the CSG), we can instead employ a lighter inference model $q(s, v|x)$ for the minimally intractable component $p(s, v|x)$ therein, and use $q(s, v|x)p(y|s)$ as $q(s, v, y|x)$. This turns the objective Eq. (1) to:

$$\max_{p,\,q_{s,v|x}} \mathbb{E}_{p^*(x,y)}\Big[\log q(y|x) + \frac{1}{q(y|x)}\mathbb{E}_{q(s,v|x)}\Big[p(y|s)\log\frac{p(s, v)p(x|s, v)}{q(s, v|x)}\Big]\Big], \tag{2}$$

where $q(y|x) := \mathbb{E}_{q(s,v|x)}[p(y|s)]$. The expectations can be estimated by Monte Carlo after applying the reparameterization trick [62]. This is the basic CSG method.

CSG-ind To actively improve OOD generalization performance, we consider using an **ind**ependent prior $p^{\perp\!\!\!\perp}(s, v) := p(s)p(v)$ for prediction in the *test domain* (Fig. 1b), where $p(s)$ and $p(v)$ are the marginals of the training-domain prior $p(s, v)$. Intuitively, $p^{\perp\!\!\!\perp}(s, v)$ discards the spurious correlation between s and v on the training domain (*e.g.*, the "desk-workspace", "bed-bedroom" association), and promotes a cautious neutral belief on the unknown test-domain correlation in defence against all possibilities (*e.g.*, a "desk-bedroom", "bed-workspace" association). Formally, $p^{\perp\!\!\!\perp}(s, v)$ has a larger entropy than $p(s, v)$ [24, Thm. 2.6.6], so it reduces training-domain-specific information and encourages reliance on the causal mechanisms for better generalization. It also amounts to applying the do-operator [85] to Fig. 1a, representing a randomized experiment by independently soft-intervening s or v. In this way, causal invariance is properly leveraged, making a different and more reliable prediction than following inference invariance. Our theory below also shows that $p^{\perp\!\!\!\perp}(s, v)$ leads to a smaller generalization error bound (Thm. 6 Remark).

Methodologically, we need the test-domain inference model $q^{\perp\!\!\!\perp}(s, v|x)$ for prediction $p^{\perp\!\!\!\perp}(y|x) \approx \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}[p(y|s)]$, but also need $q(s, v|x)$ for learning on the training domain. To save the cost of building and learning two inference models, we propose to use $q^{\perp\!\!\!\perp}(s, v|x)$ to represent $q(s, v|x)$. Noting that their targets are related by $p(s, v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p^{\perp\!\!\!\perp}(s, v|x)$, we formulate $q(s, v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}q^{\perp\!\!\!\perp}(s, v|x)$ accordingly, so that this $q(s, v|x)$ achieves its target if and only if $q^{\perp\!\!\!\perp}(s, v|x)$

does. The objective Eq. (1) then becomes:

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

where $\pi(y|x) := \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\big[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\big]$. (Note $p^{\perp\!\!\!\perp}(s,v)$ is determined by $p(s,v)$ in the CSG p.)

4.2 Method for Domain Adaptation

In domain adaptation, one also has unsupervised data from the underlying data distribution $\tilde{p}^*(x)$ on the *test domain*. We can leverage them for better prediction. According to the causal invariance principle (2), we only need a new prior $\tilde{p}(s,v)$ for the test-domain CSG $\tilde{p} := \langle \tilde{p}(s,v), p(x|s,v), p(y|s)\rangle$ (Fig. 1c). Fitting test-domain data can be done through the standard ELBO objective with the test-domain inference model $\tilde{q}(s,v|x)$:

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

Prediction is given by $\tilde{p}(y|x) \approx \mathbb{E}_{\tilde{q}(s,v|x)}[p(y|s)]$. Similar to the CSG-ind case, we still need $q(s,v|x)$ for fitting training-domain data, and we can also avoid a separate $q(s,v|x)$ model by representing it using $\tilde{q}(s,v|x)$. Following the same relation between their targets, we let $q(s,v|x) = \frac{\tilde{p}(x)}{p(x)}\frac{p(s,v)}{\tilde{p}(s,v)}\tilde{q}(s,v|x)$, which reformulates the same training-domain objective Eq. (1) as:

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

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 resulting method, termed CSG-DA, solves both optimization problems Eqs. (4, 5) simultaneously.

4.3 Implementation and Model Selection

To implement the three CSG methods, we only need one inference model in each. Appx. F.2 shows its construction from a general discriminative model (*e.g.*, how to select its hidden nodes as s and v). In practice x often has a much larger dimension than y, making the first supervision term overwhelmed by the second unsupervised term in Eqs. (2,3,5). So we downscale the second term.

As recently emphasized [39], an OOD method should include a model selection method, since it is nontrivial and significantly affects performance [95, 120]. For our methods, we use a validation set from the *training domain* for model selection. This complies with the OOD setup, and is also suggested by our theory below which gives guarantees based on a good fit to the training-domain data distribution. For CSG-ind/DA, the learned predictor targets the *test* domain, so we *do not* use it directly for evaluating validation accuracy, but by normalizing $\pi(y|x)$. Appx. F.3 shows details.

5 Theory

We now establish theory for the identification of the semantic factor (cause of prediction) and subsequent merits for OOD generalization and domain adaptation. We focus on the distribution-level generalization instead of from finite samples to unseen samples under the same distribution, so we only consider the infinite-data regime. Appx. A shows all the proofs and auxiliary theory.

Latent variable identification is hard [65, 81, 116, 70] as it is beyond observational relations [51, 88]. Assumptions are thus required to draw definite conclusions.

Assumption 3. (**Additive noise**) There exist nonlinear functions f and g with bounded derivatives up to the 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 $p(y|s) = \mathrm{Cat}(y|g(s))$ for categorical y.
(**Bijectivity**) Assume f is bijective and g is injective.

The additive noise assumption is widely adopted in causal discovery [51, 17]. It disables expressing the same joint in the other direction [122, Thm. 8; 86, Prop. 23] so that CSG unnecessarily indicates inference invariance. For this reason, we exclude GAN [37] and flow-based [61] implementations. Bijectivity is a common assumption for identifiability [51, 100, 57, 68]. It is sufficient [86, Prop. 17; 88, Prop. 7.4] for the more fundamental [86, Prop. 7; 88, p.109] requirement of causal minimality [86, p.2012; 88, Def. 6.33]. Particularly, s and v may otherwise have dummy dimensions that f and g simply ignore, raising another ambiguity against identifiability. On the other hand, according to the

commonly acknowledged manifold hypothesis [115, 31], we can take $\mathcal{X}$ as the lower-dimensional data manifold and such a bijection exists as a coordinate map, which is an injection to the original data space and also allows $d_{\mathcal{S}} + d_{\mathcal{V}} < d_{\mathcal{X}}$.

5.1 Identifiability Theory

We first formalize the goal of identifying the semantic factor.

Definition 4 (semantic-identification)**.** We say a learned CSG p is *semantic-identified*, if there exists a homeomorphism[4] Φ on $\mathcal{S} \times \mathcal{V}$, such that **(i)** its output dimensions in $\mathcal{S}$ is constant of v: $\Phi^{\mathcal{S}}(s, v) = \Phi^{\mathcal{S}}(s, v'), \forall v, v' \in \mathcal{V}$ (hence denote $\Phi^{\mathcal{S}}(s, v)$ as $\Phi^{\mathcal{S}}(s)$), and **(ii)** it is a *reparameterization* of the ground-truth CSG p^*: $\Phi_{\#}[p^*_{s,v}] = p_{s,v}$, $p^*(x|s, v) = p(x|\Phi(s, v))$ and $p^*(y|s) = p(y|\Phi^{\mathcal{S}}(s))$.

Here, $\Phi_{\#}[p^*_{s,v}]$ denotes the pushed-forward distribution[5] of $p^*_{s,v}$ by Φ, *i.e.* the distribution of $\Phi(s, v)$ when $(s, v) \sim p^*_{s,v}$. As the ground-truth CSG could at most provide its information via the data distribution $p^*(x, y)$, a well-learned CSG that achieves $p(x, y) = p^*(x, y)$ still has the degree of freedom in parameterizing (s, v). This is described by this reparameterization Φ (Appx. Lemma 9). At the heart of the definition, the v-constancy of $\Phi^{\mathcal{S}}$ implies that Φ is *semantic-preserving*: the learned model *does not mix* the ground-truth v into its s, so that the learned s holds equivalent information to the ground-truth s. The definition can thus be seen as the semantic equivalence (Appx. Def. 10, Prop. 14) to the ground-truth CSG p^*.

For related concepts, this identification cannot be characterized by the *statistical independence* between s and v (vs. [18, 49, 121]), which is not sufficient [70] nor necessary (due to the existence of spurious correlation). It is also weaker than *disentanglement* [44, 11], which additionally requires the learned v to be constant of the ground-truth s. The following theorem shows that semantic-identification can be achieved on a single domain under certain conditions.

Theorem 5 (semantic-identifiability)**.** *With Assumption 3, a CSG p is semantic-identified, if it is well-learned such that $p(x, y) = p^*(x, y)$, under the conditions that $\log p(s, v)$ and $\log p^*(s, v)$ are bounded up to the second-order, and that*[6] **(i)** *$1/\sigma_\mu^2 \to \infty$ where $\sigma_\mu^2 := \mathbb{E}[\mu^\top \mu]$,* ***or*** **(ii)** *$p_\mu$ (*e.g.*, a Gaussian) has an a.e. non-zero characteristic function.*

Remarks. (1) (**Condition and Intuition**) Compared with the multi-domain case [87, 93, 2], identifiability on a single training domain comes at a cost and requires certain conditions. One may imagine that in some extreme cases *e.g.*, all desks appear in workspace and all beds in bedrooms, it is impossible to distinguish whether y labels the object or the background (unlearnable OOD problem [119]). The theorem finds an *appropriate condition* that excludes such cases: when $\log p^*(s, v)$ is bounded, deterministic s-v relations are not allowed as they concentrate $p^*(s, v)$ on a lower-dimensional subspace in $\mathcal{S} \times \mathcal{V}$ thus make it unbounded.

It also leads to the *intuition of identifiability*: a bounded $\log p^*(s, v)$ indicates a stochastic s-v relation, so mixing the ground-truth v into the learned s makes the inference of s more noisy due to the intrinsic diversity/uncertainty of this v. As prediction is made via the inferred s, this worsens prediction accuracy thus violates the "well-learned" requirement. Compared with discriminative models, CSG makes more faithful inference, and its causal structure leads to a proper description of domain change.

(2) In condition **(i)**, $1/\sigma_\mu^2$ measures the *intensity* of the causal mechanism $p(x|s, v)$. When it is large, the "strong" $p(x|s, v)$ helps disambiguating values of (s, v) in generating a given x. The formal version in Appx. Thm. 5' shows a quantitative reference for large enough intensity, and Appx. B gives a non-asymptotic extension showing how the intensity trades-off the tolerance of equalities in Def. 4. Condition **(ii)** goes beyond inference invariance. It roughly implies that different (s, v) values a.s. produce different $p(x|s, v)$, so their roles in generating x become clear which helps identification.

(3) The theorem does not contradict the impossibility result by Locatello et al. [70], which considers disentangling each latent dimension with an unconstrained $(s, v) \to (x, y)$, while we only identify s as a whole, with the $v \to y$ edge removed which breaks the s-v symmetry.

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

[5] The definition of $\Phi_{\#}[p^*_{s,v}]$ requires Φ to be measurable. This is satisfied by the continuity of Φ as a homeomorphism (as long as the Borel σ-field is considered) [13, Thm. 13.2].

[6] To be precise, the conclusions are that the equalities in Def. 4 hold asymptotically in the limit $1/\sigma_\mu^2 \to \infty$ for condition **(i)**, and hold a.e. for condition **(ii)**.

5.2 OOD Generalization Theory

Now we show the benefit of semantic-identification for OOD generalization that the prediction error is bounded. Note the optimal predictor $\tilde{\mathbb{E}}^*[y|x]$ [7] on the test domain is defined by the corresponding ground-truth CSG $\tilde{p}^*$, which differs from p^* only in the test-domain prior $\tilde{p}^*(s,v)$ (Principle 2).

Theorem 6 (OOD generalization error)**.** [8] *With Assumption 3, for a semantic-identified CSG p on the training domain with semantic-preserving reparameterization Φ, we have up to $O(\sigma_\mu^4)$,*

$$\mathbb{E}_{\tilde{p}^*(x)} \|\mathbb{E}[y|x] - \tilde{\mathbb{E}}^*[y|x]\|_2^2 \leqslant \sigma_\mu^4 B_{f^{-1}}'^4 B_g'^2 \, \mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log(\tilde{p}_{s,v}/p_{s,v}) \right\|_2^2, \tag{6}$$

where $B'_{f^{-1}}$ and B'_g bound the 2-norms[9] of the Jacobians of f^{-1} and g, respectively, and $\tilde{p}_{s,v} := \Phi_\#[\tilde{p}^*_{s,v}]$ is the test-domain prior under the parameterization of the CSG p.

In the bound, the term $\mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log(\tilde{p}_{s,v}/p_{s,v}) \right\|_2^2$ is the Fisher divergence measuring the difference between the two priors. As the prior change is the only source of domain change, this term also measures the "OODness" in terms of the effect on prediction. The bound also shows that when the causal mechanism $p(x|s,v)$ is strong (small σ_μ), it dominates prediction over the prior change, as the generalization error becomes small. Compared with other methods, using a CSG enforces causal invariance, so the boundedness of OOD generalization error becomes more plausible in practice.

Remark. The bound also shows the advantage of CSG-ind (Sec. 4.1). The Fisher divergence is revealed [28] to have a similar behavior as the forward KL divergence $p_{s,v} \mapsto \mathrm{KL}(\tilde{p}_{s,v} \| p_{s,v})$ that it is very sensitive to the insufficient coverage of $p_{s,v}$ on the support of $\tilde{p}_{s,v}$ [46, 109], since $\log(\tilde{p}_{s,v}/p_{s,v})$ is infinitely large on the uncovered region. As the independent prior $p^{\perp\!\!\!\perp}_{s,v}$ has a larger support than $p_{s,v}$, it is less likely to miss the support of $\tilde{p}_{s,v}$, so it induces a generally smaller Fisher divergence. CSG-ind thus generally has a smaller OOD generalization error bound than CSG.

5.3 Domain Adaptation Theory

CSG-DA (Sec. 4.2) learns a new prior $\tilde{p}_{s,v}$ by fitting unsupervised test-domain data, with causal mechanisms shared. If the mechanisms are semantic-identified, the ground-truth test-domain prior $\tilde{p}^*_{s,v}$ can also be identified under the learned parameterization, and prediction is made precise.

Theorem 7 (domain adaptation error)**.** *With conditions of Thm. 5, for a semantic-identified CSG p on the training domain with semantic-preserving reparameterization Φ, if its new prior $\tilde{p}_{s,v}$ is well-learned such that $\tilde{p}(x) = \tilde{p}^*(x)$, then $\tilde{p}_{s,v} = \Phi_\#[\tilde{p}^*_{s,v}]$, and $\tilde{\mathbb{E}}[y|x] = \tilde{\mathbb{E}}^*[y|x]$ for any $x \in \mathrm{supp}(\tilde{p}^*_x)$.*

Different from existing domain adaptation bounds (Appx. E), Theorems 6,7 allow different inference models in the two domains, thus go beyond inference invariance.

6 Experiments

For OOD generalization baselines, there is not much choice beyond the standard CE loss optimization, as domain adaptation methods require test-domain data and most domain generalization methods degenerate to CE with one training domain. The exception within our scope is a causal discriminative method CNBB [41]. For domain adaptation, we consider well-acknowledged methods DANN [33], DAN [73], CDAN [74] and recent compelling methods MDD [124] and BNM [25] (shown in Appx. Tables 2,3). Appx. G shows more details, results, and discussions. [10]

Shifted-MNIST. We first consider an OOD prediction task on MNIST to classify digits "0"s and "1"s. To make a spurious correlation, in the training data, we horizontally shift each "0" at random by $\delta_0 \sim \mathcal{N}(-5, 1^2)$ pixels, while each "1" by $\delta_1 \sim \mathcal{N}(5, 1^2)$ pixels. We consider two test domains with different digit-position distributions: each digit is not moved $\delta_0 = \delta_1 = 0$ in the first, and is shifted at random by $\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$ pixels in the second. We implement all methods using a multilayer perceptron which is not naturally shift invariant. We use a larger architecture for non-generative methods to compensate the additional generative component of generative methods.

The performance is shown in Table 1(top 2 rows). For OOD generalization, CE is misled by the more noticeable position factor due to the spurious correlation to digits, and resorts to random guess (even

[7] For categorical y, the expectation of y is taken under the one-hot representation.

[8] See Appx. Thm. 6' for the formal version.

[9] As the induced operator norm for matrices (not the Frobenius norm).

[10] Codes are available at `https://github.com/changliu00/causal-semantic-generative-model`.

Table 1: Test accuracy (%) by various methods (ours in bold) for OOD generalization (left 4 cols) and domain adaptation (right 5 cols) on Shifted-MNIST (top 2 rows), ImageCLEF-DA (middle 4 rows) and PACS (bottom 4 rows) datasets. Averaged over 10 runs. Appx. Tables 2,3 show more results.

task	CE	CNBB	**CSG**	**CSG-ind**	DANN	DAN	CDAN	MDD	**CSG-DA**
$\delta_0 = \delta_1 = 0$	42.9±3.1	54.7±3.3	81.4±7.4	**82.6±4.0**	40.9±3.0	40.4±2.0	41.0±0.5	41.9±0.8	**97.6±4.0**
$\delta_0,\delta_1 \sim \mathcal{N}(0,2^2)$	47.8±1.5	59.2±2.4	61.7±3.6	**62.3±2.2**	46.2±0.7	45.6±0.7	46.3±0.6	45.8±0.3	**72.0±9.2**
C→P	65.5±0.3	72.7±1.1	73.6±0.6	**74.0±1.3**	74.3±0.5	69.2±0.4	74.5±0.3	74.1±0.7	**75.1±0.5**
P→C	91.2±0.3	91.7±0.2	92.3±0.4	**92.7±0.2**	91.5±0.6	89.8±0.4	**93.5±0.4**	92.1±0.6	**93.4±0.3**
I→P	74.8±0.3	75.4±0.6	76.9±0.3	**77.2±0.2**	75.0±0.6	74.5±0.4	76.7±0.3	76.8±0.4	**77.4±0.3**
P→I	83.9±0.1	88.7±0.5	90.4±0.3	**90.9±0.2**	86.0±0.3	82.2±0.2	90.6±0.3	90.2±1.1	**91.1±0.5**
others→**P**	**97.8±0.0**	96.9±0.2	97.7±0.2	**97.8±0.2**	97.6±0.2	97.6±0.4	97.0±0.4	97.6±0.3	**97.9±0.2**
others→**A**	88.1±0.1	73.1±0.3	**88.5±0.6**	**88.6±0.6**	85.9±0.5	84.5±1.2	84.0±0.9	88.1±0.8	**88.8±0.7**
others→**C**	77.9±1.3	50.2±1.2	84.4±0.9	**84.6±0.8**	79.9±1.4	81.9±1.9	78.5±1.5	83.2±1.1	**84.7±0.8**
others→**S**	79.1±0.9	43.3±1.2	80.7±1.0	**81.1±1.2**	75.2±2.8	77.4±3.1	71.8±3.9	80.2±2.2	**81.4±0.8**

worse) when position is not informative for prediction. CNBB ameliorates the position confusion, but not as thoroughly without modeling causal mechanisms. In contrast, our CSG gives more genuine predictions in unseen domains, thanks to the identification of the semantic factor. CSG-ind performs even better, justifying the merit of using an independent prior for prediction. For domain adaptation, CSG-DA achieves the best results. Existing adaptation methods even worsen the result (negative transfer), as the misleading position representation gets strengthened on the unsupervised test data. CSG is benefited from adaptation in a proper way that identifies the semantic factor.

ImageCLEF-DA is a standard benchmark for domain adaptation [1]. It has 12 classes and three domains of real-world images: **C**altech-256, **I**mageNet, **P**ascal VOC 2012. We select four OOD prediction tasks **C**↔**P**, **I**↔**P** that have not seen good enough results. We adopt the same setup as [74]. As shown in Table 1(middle 4 rows), CSG-ind again achieves the best OOD generalization results, and even outperforms some domain adaptation methods. Our CSG also outperforms the baselines mostly. For domain adaptation, CSG-DA is the best in most cases and on par with the best in others.

PACS is a more recent benchmark dataset [69]. It has 7 classes and is named after its four domains: **P**hoto, **A**rt, **C**artoon, **S**ketch; each contains images of a certain style. We follow the same setup as [39]; particularly, we pool together all domains but the test one as the single training domain. Results in Table 1(bottom 4 rows) show the same trend. CSG-DA even outperforms most domain generalization methods reported in [39], which are fed with more information. Appx. Tables 2,3 also show the results on an even larger dataset **VLCS** [30], which present a similar observation.

Visualization. Appx. Fig. 5 visualizes the learned models using LIME [91]. The results show our methods focus more on the semantic regions and shapes, indicating a causal representation is learned.

Dataset analysis. The results indicate our methods are more powerful on shifted-MNIST and PACS (and VLCS) than ImageCLEF-DA. This meets the intuition of identifiability (Thm. 5 Remark **(1)**): the random position or pooled training domain shows a diverse v for each s (while with a misleading spurious correlation), so identification is better guaranteed to overcome the spurious correlation.

Ablation study. To show the benefit of modeling s and v separately, we compare with a counterpart of CSG that treats s and v as a whole (equivalently, $v \to y$ is kept; see Appx. F.1.4 for method details). Appx. Tables 2,3 show that our methods outperform this baseline in all cases. This shows the separate modeling makes CSG consciously drive semantic representation into the dedicated variable s.

7 Conclusion and Discussion

We propose a Causal Semantic Generative model for single-domain OOD prediction tasks, which builds upon a causal reasoning, and models the semantic (cause of prediction) and variation factors separately. By the causal invariance principle, we develop novel and efficient learning and prediction methods, and prove the semantic-identifiability and the subsequent bounded generalization error and the success of adaptation. Experiments show the improved performance over prevailing baselines.

Notably, we answered the questions in the recent farseeing paper [98] on causal representation learning: we found an appropriate condition under which "causal variables can be recovered", and provided "compelling evidence on the advantages (of causal modeling) in terms of generalization". Also, separating semantics from variation extends to broader examples. Neural nets are found to

change their prediction under a different texture [34, 15]. Adversarial vulnerability [107, 38, 67] extends variation factors to human-imperceptible features, *i.e.* adversarial noise, which is found to have a strong correlation to the semantics [50]. The separation also matters for fairness when a sensitive variation factor may affect prediction. This work also inspires the dual connection between causal representation learning ("fill in the blanks" given a graph) and causal discovery ("link the nodes" given observed variables). Our theory shows the identifiability condition for causal discovery (the additive noise assumption) also makes causal representation identifiable. Studying the general connection between the two tasks is an interesting future work.

References

[1] The imageclef-da challenge 2014. https://www.imageclef.org/2014, 2014.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] M. T. Bahadori, K. Chalupka, E. Choi, R. Chen, W. F. Stewart, and J. Sun. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017.

[5] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[6] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[8] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[9] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[10] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565. PMLR, 2018.

[11] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *Proceedings of the International Conference on Learning Representations (ICLR 2020)*, 2020.

[12] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[13] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New Jersey, 2012. ISBN 978-1-118-12237-2.

[14] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[15] W. Brendel and M. 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.

[16] P. Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

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

[18] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Conference of IJCAI*, volume 2019, page 2060. NIH Public Access, 2019.

[19] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[20] J. Chen and K. Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

[21] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[23] C.-Y. Chuang, A. Torralba, and S. Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *International Conference on Machine Learning*, pages 1984–1994. PMLR, 2020.

[24] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[25] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020.

[26] B. Dai and D. Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

[27] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[28] C. Durkan and Y. Song. On maximum likelihood training of score-based generative models. *arXiv preprint arXiv:2101.09258*, 2021.

[29] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[30] C. Fang, Y. Xu, and D. N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[31] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[32] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059, 2016.

[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.

[34] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. 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.

[35] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848, 2016.

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

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, Montréal, Canada, 2014. NIPS Foundation.

[38] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, 2015.

[39] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Y. He, Z. Shen, and P. Cui. Towards non-i.i.d. image classification: A dataset and baselines. *arXiv preprint arXiv:1906.02899*, 2019.

[42] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *stat*, 1050:13, 2019.

[43] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

[44] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[45] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[46] F. Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

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

[48] M. Ilse, J. M. Tomczak, and P. Forré. Designing data augmentation for simulating interventions. *arXiv preprint arXiv:2005.01856*, 2020.

[49] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling. DIVA: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[50] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

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

[52] 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)*, pages 383–391. AUAI Press, 2011.

[53] J. Jiang, B. Fu, and M. Long. Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library`, 2020.

[54] F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019.

[55] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[56] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, C. Pal, and Y. Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[57] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, 2020.

[58] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models. *arXiv preprint arXiv:2002.11537*, 2020.

[59] N. Kilbertus, G. Parascandolo, and B. Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

[60] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[61] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

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

[63] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[64] M. Kocaoglu, S. Shakkottai, A. G. Dimakis, C. Caramanis, and S. Vishwanath. Entropic latent variable discovery. *arXiv preprint arXiv:1807.10399*, 2018.

[65] T. C. Koopmans and O. Reiersol. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

[66] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (REx). *arXiv preprint arXiv:2003.00688*, 2020.

[67] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[68] C. M. Lee, C. Hart, J. G. Richens, and S. Johri. Leveraging directed causal discovery to detect latent common causes. *arXiv preprint arXiv:1910.10174*, 2019.

[69] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[70] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[71] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

[72] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[73] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015.

[74] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[75] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[76] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[77] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

[78] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, Vancouver, Canada, 2008. NIPS Foundation.

[79] J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=9p2ekP904Rs`.

[80] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[81] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[82] R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

[83] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[84] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[85] J. Pearl. *Causality*. Cambridge university press, 2009.

[86] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[87] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[88] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[89] F. Qiao, L. Zhao, and X. Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[90] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[91] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[92] T. Richardson, P. Spirtes, et al. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

[93] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[94] J.-W. Romeijn and J. Williamson. Intervention and identifiability in latent variable modelling. *Minds and machines*, 28(2):243–264, 2018.

[95] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[96] B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

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

[98] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[99] E. Sgouritsa, D. Janzing, J. Peters, and B. 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)*, pages 556–575. AUAI Press, 2013.

[100] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR.org, 2017.

[101] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi. Generalizing across domains via cross-gradient training. In *Proceedings of the International Conference on Learning Representations (ICLR 2018)*, 2018.

[102] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen. Causally regularized learning with agnostic data selection bias. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 411–419. ACM, 2018.

[103] I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[104] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

[105] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[106] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[107] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, 2014.

[108] T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9458–9469, 2020.

[109] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–10, 2016.

[110] T. Tieleman and G. 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.

[111] T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

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

[113] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization, 2021. Survey Track.

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

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

[116] Y. Yacoby, W. Pan, and F. Doshi-Velez. Learning deep bayesian latent variable regression models that generalize: When non-identifiability is a problem. *arXiv preprint arXiv:1911.00569*, 2019.

[117] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[118] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[119] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang. Towards a theoretical framework of out-of-distribution generalization. *arXiv preprint arXiv:2106.04496*, 2021.

[120] K. You, X. Wang, M. Long, and M. Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 7124–7133, 2019.

[121] C. Zhang, K. Zhang, and Y. Li. A causal view on robustness of neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[122] 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)*, pages 647–655. AUAI Press, 2009.

[123] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

[124] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.

[125] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.