Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms

Aneesh Komanduri[1], **Yongkai Wu**[2], **Feng Chen**[3] and **Xintao Wu**[1]
[1]University of Arkansas
[2]Clemson University
[3]University of Texas at Dallas
{akomandu, xintaowu}@uark.edu, yongkaw@clemson.edu, feng.chen@utdallas.edu

Abstract

Learning disentangled causal representations is a challenging problem that has gained significant attention recently due to its implications for extracting meaningful information for downstream tasks. In this work, we define a new notion of causal disentanglement from the perspective of independent causal mechanisms. We propose ICM-VAE, a framework for learning causally disentangled representations supervised by causally related observed labels. We model causal mechanisms using nonlinear learnable flow-based diffeomorphic functions to map noise variables to latent causal variables. Further, to promote the disentanglement of causal factors, we propose a causal disentanglement prior learned from auxiliary labels and the latent causal structure. We theoretically show the identifiability of causal factors and mechanisms up to permutation and elementwise reparameterization. We empirically demonstrate that our framework induces highly disentangled causal factors, improves interventional robustness, and is compatible with counterfactual generation.

1 Introduction

Disentangled representation learning aims to learn meaningful and compact representations that capture semantic aspects of data by structurally disentangling the factors of variation [Bengio *et al.*, 2013]. Such representations have been shown to offer useful properties such as better interpretability, robustness to distribution shifts, efficient out-of-distribution sampling, and fairness [Locatello *et al.*, 2019]. However, disentangled representation learning typically assumes that the underlying factors are independent, which is unrealistic in practice. The factors generating the data can contain correlations or even causal relationships that are disregarded when factors are assumed to be independent. Further, a generative model learning from an independent prior assumes that all combinations of the latent factors are equally likely to appear in the training data. Thus, disentangling the factors would yield a sub-optimal likelihood since the assumed support could be well outside the support of the training data.

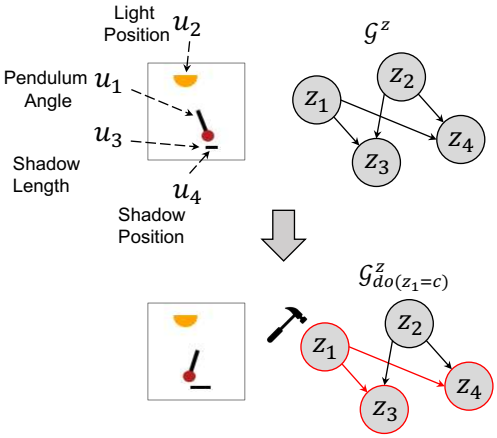

Figure 1: We learn causal models representing images as latent causal variables z. The bottom shows the effect of intervening on the latent code corresponding to the pendulum's angle, propagating effects, and generating a counterfactual image.

Recently, there has been a growing interest in bridging causality [Pearl, 2009] and representation learning [Bengio *et al.*, 2013]. The goal of causal representation learning (CRL) is to map unstructured low-level data to high-level abstract causal variables of interest [Scholkopf *et al.*, 2021]. The key assumption is that high-dimensional observations are generated from a set of underlying low-dimensional *causally related* factors of variation. Causal representations have been shown to be useful for tasks involving reasoning and planning. Causal representations also adhere to the principle of independent causal mechanisms (ICM) [Parascandolo *et al.*, 2018], which states that the mechanisms that generate each causal variable are independent such that a change in one mechanism does not affect another [Schölkopf and von Kügelgen, 2022; Peters *et al.*, 2017]. Learning a generative model that captures the causal structure among latent factors can be crucial for reasoning about the world under manipulation. For example, a pendulum, light source, and shadow, as seen in Figure 1, may be causally related but are separate entities in the world that can be independently manipulated. Particularly, manipulating the pendulum's angle will affect the shadow's position and length. These hypothetical scenarios could be counterfactually generated from a causal model.

Ensuring that causal representations are *disentangled* is useful for the causal controllable generation to generate counterfactual instances unseen during training. Such instances could be utilized as augmented data for robust learning in downstream tasks. The notion of disentanglement may be trivial when the factors are independent but becomes difficult to achieve when there are correlations or causal relationships among factors in the observed data. For highly correlated factors, it can be difficult to separate the factors of variation from their latent codes [Träuble *et al.*, 2021]. Recently, it was shown that it is impossible to learn a disentangled representation in an unsupervised manner without some form of inductive bias [Locatello *et al.*, 2019]. Recent work proved that models with an independent prior are unidentifiable [Shen *et al.*, 2022]. Further, most existing disentanglement methods fail to disentangle factors when correlations exist in the data [Träuble *et al.*, 2021]. However, results from large-scale empirical studies [Locatello *et al.*, 2020] have indicated that supervision in the form of auxiliary labels or contrastive data can effectively disentangle correlated or causal factors.

Related Work. Our work builds upon the ideas presented in iVAE [Khemakhem *et al.*, 2020] and causal variants [Yang *et al.*, 2021; Komanduri *et al.*, 2022] and extends them to consider a principled view of causal disentanglement in the label supervised setting. DIVA [Ilse *et al.*, 2020] and CC-VAE [Joy *et al.*, 2021] are special case implementations of the iVAE framework. Yang et al. [2021] proposed CausalVAE, which uses a causal masking layer and is limited to linear SCMs. Komanduri et al. [2022] extended this to a nonlinear setting and proposed a causal prior. Both works proposed simplistic models to learn causal mechanisms under the strictly additive noise assumption and do not, from an empirical or theoretical perspective, focus on disentanglement. Shen et al. [2022] proposed learning causal representations supervised by a GAN loss. There has also been work focusing on learning causal representations from paired counterfactual data [Brehmer *et al.*, 2022], temporal data [Lippe *et al.*, 2023], in self-supervised learning [von Kügelgen *et al.*, 2021], and when interventional data is available [Ahuja *et al.*, 2023]. Unlike many previous works in label-supervised VAE-based CRL, we consider general nonlinear SCMs instead of restricting to additive noise models, propose a causal prior to causally factorize the latent space, and achieve disentanglement of causal mechanisms, as summarized in Table 1.

Contributions. (1) We propose ICM-VAE, a framework for causal representation learning under label supervision, where causal variables are derived from nonlinear flow-based diffeomorphic causal mechanisms. (2) Based on the ICM principle, we propose the notion of causal disentanglement for causal models from the perspective of mechanisms and design a causal disentanglement prior to causally factorize the learned distribution over causal variables. (3) Using the structure from our causal disentanglement prior, we theoretically show identifiability of the learned causal factors up to permutation and elementwise reparameterization. (4) We experimentally validate our method and show that our model can almost perfectly disentangle the causal factors, improve interventional robustness, and generate consistent counterfactuals.

Framework	General SCM Compatible	Causal Prior	Causal Mechanism Disentanglement
iVAE	✗	✗	✗
CausalVAE	✗	✗	✗
SCM-VAE	✗	✓	✗
ICM-VAE (Ours)	✓	✓	✓

Table 1: Causal and acausal framework compatibilities in label-supervised setting

2 Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the support of the observed data x generated from a set of causally related ground-truth factors. The observations are assumed to be explained by some latent causal factors of variation z with domain $\mathcal{Z} \subset \mathbb{R}^n$, where $n \ll d$. z represents a set of causal factors while z_i represents a single causal factor. We assume x can be decomposed as $x = g(z) + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2 I)$ are mutually independent Gaussian noise terms for reconstruction. Let $g : \mathcal{Z} \to \mathcal{X}$ be the decoder (or mixing function). Each factor z_i contains semantically meaningful information about the observation. In the traditional VAE [Kingma and Welling, 2013], we assume the observed data is generated by the latent generative model with the structure $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ where θ are the true but unknown parameters. We aim to learn the joint distribution $p(x, z)$ to estimate the marginal density and a posterior $p(z|x)$ to describe the underlying factors of variation given a prior $p(z)$ over the latent variables.

2.1 Identifiability

The goal of learning a useful representation that recovers the true underlying data-generating factors is closely tied to the problem of blind source separation (BSS) and independent component analysis (ICA) [Hyvarinen *et al.*, 2018; Comon, 1994; Hyvärinen and Pajunen, 1999]. Provably showing that a learning algorithm achieves this goal up to tolerable ambiguities under certain conditions is formalized as the *identifiability* of a model. In this section, we use the notion of $\sim$-equivalence [Khemakhem *et al.*, 2020] to formulate the notion of identifiability.

Definition 1 ($\sim$-identifiability)**.** *Let $\sim$ be an equivalence relation on θ. The generative model is $\sim$-identifiable if*

$$p_\theta(x) = p_{\hat{\theta}}(x) \implies \theta \sim \hat{\theta} \tag{1}$$

If two different choices of model parameter θ and $\hat{\theta}$ lead to the same marginal density $p_\theta(x)$, then they must be equal and this implies that $p_\theta(x, z) = p_{\hat{\theta}}(x, z)$, $p_\theta(z) = p_{\hat{\theta}}(z)$, and $p_\theta(z|x) = p_{\hat{\theta}}(z|x)$. However, recent work showed that it is impossible to achieve marginal density equivalence $p_\theta(x) = p_{\hat{\theta}}(x)$ with an unconditional prior $p_\theta(z)$ [Khemakhem *et al.*, 2020]. Since the VAE is unidentifiable without some form of additional restriction on the function class of the mixing function or auxiliary information, the identifiable VAE (iVAE) was proposed to utilize auxiliary information in the form of a conditionally factorial prior for identifiability guarantees [Khemakhem *et al.*, 2020]. In iVAE, each factor z_i is assumed to have a univariate exponential family distribution (due to their universal approximation capabilities) given the

conditioning variable u, where a function $\boldsymbol{\lambda}$ determines the natural parameters of the distribution. The general PDF of the conditional distribution is defined as follows:

$$p_{T,\lambda}(z|u) = \prod_i h_i(z_i) \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(u) - \psi_i(u)\right] \quad (2)$$

where $h_i(z_i)$ is the base measure, $\mathbf{T}_i : \mathcal{Z} \to \mathbb{R}^k$ and $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,k})$ are the sufficient statistics, $\boldsymbol{\lambda}_i(u) = (\lambda_{i,1}(u), \ldots, \lambda_{i,k}(u))$ are the corresponding natural parameters, k is the dimension of each sufficient statistic, and the remaining term $\psi_i(u)$ acts as a normalizing constant. A prior conditioned on auxiliary information u can guarantee that the joint densities $p_\theta(x, z) = p_{\hat{\theta}}(x, z)$ are equivalent up to some equivalence class. The following two definitions describe the conditions necessary to achieve the identifiability of a learned model up to linear transformation and block permutation indeterminacies, respectively.

Definition 2 ([Khemakhem *et al.*, 2020])**.** *Let $\sim$ be an equivalence relation on θ, $\mathcal{X} = g(\mathcal{Z})$, and $\hat{\mathcal{X}} = \hat{g}(\mathcal{Z})$. We say that θ and $\hat{\theta}$ are **linearly-equivalent** if and only if there exists an invertible matrix $A \in \mathbb{R}^{nk \times nk}$ and vectors $b, c \in \mathbb{R}^{nk}$ such that $\forall x \in \mathcal{X}$:*

1. $\mathbf{T}(g^{-1}(x)) = A\hat{\mathbf{T}}(\hat{g}^{-1}(x)) + b, \forall x \in \mathcal{X}$
2. $A^T\boldsymbol{\lambda}(u) + c = \hat{\boldsymbol{\lambda}}(u)$

We denote this equivalence as $\theta \sim_A \hat{\theta}$.

Definition 3 ([Khemakhem *et al.*, 2020])**.** *We say θ and $\hat{\theta}$ are **permutation-equivalent**, denoted $\theta \sim_P \hat{\theta}$, if and only if P is permutation matrix that has block-permutation structure respecting $\mathbf{T}$. That is, there exist n invertible $k \times k$ matrices $A_1, \ldots, A_n$ and an n-permutation π such that for all $z \in \mathbb{R}^{nk}$, $P\hat{z} = [z_{\pi(1)}A_1^T, z_{\pi(2)}A_2^T, \ldots, z_{\pi(n)}A_n^T]^T$.*

Linear equivalence indicates the true representation is a linear transformation of the learned representation and only guarantees the learned representation captures the true representation. In general, linear-equivalent identifiability does not guarantee that the factors of variation are disentangled since the linear transformation can mix up the variables (i.e. one component of g^{-1} corresponds to multiple components of $\hat{g}^{-1}$). Permutation equivalence implies that the i-th factor z_i of one representation corresponds to a unique factor in another representation, given the permutation π. To truly disentangle factors of variation, we must ensure that each coordinate of the learned representation is equal to the scaled and shifted coordinate of the ground truth up to some permutation. To this end, we define the notion of disentanglement as permutation equivalence [Lachapelle *et al.*, 2022] as follows.

Definition 4 (Permutation Disentanglement)**.** *Given some ground-truth model, a learned model $\hat{\theta}$ is said to be **disentangled** if θ and $\hat{\theta}$ are permutation-equivalent.*

2.2 Structural Causal Model

Henceforth, we assume z is described by a structural causal model (SCM) [Pearl, 2009], which is formally defined by a tuple $\mathcal{M} = \langle \mathcal{Z}, \mathcal{E}, F \rangle$, where $\mathcal{Z}$ is the domain of the set of n endogenous causal variables $z = \{z_1, \ldots, z_n\}$, $\mathcal{E}$ is the domain of the set of n exogenous noise variables $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$, which is learned as an intermediate latent variable, and $F = \{f_1, \ldots, f_n\}$ is a collection of n independent causal mechanisms of the form

$$z_i = f_i(\epsilon_i, z_{\mathbf{pa}_i}) \quad (3)$$

where $\forall i$, $f_i : \mathcal{E}_i \times \prod_{j \in \mathbf{pa}_i} \mathcal{Z}_j \to \mathcal{Z}_i$ are **causal mechanisms** that determine each causal variable as a function of the parents and noise, $z_{\mathbf{pa}_i}$ are the parents of causal variable z_i; and a probability measure $p_\mathcal{E}(\epsilon) = p_{\mathcal{E}_1}(\epsilon_1)p_{\mathcal{E}_2}(\epsilon_2)\ldots p_{\mathcal{E}_n}(\epsilon_n)$, which admits a product distribution. For the purposes of this work, we assume a causally sufficient setting (no hidden confounding), no SCM misspecification, and faithfulness is satisfied. We depict the causal structure of z by a directed acyclic graph (DAG) $\mathcal{G}^z$ with adjacency matrix $G^z \in \{0, 1\}^{n \times n}$.

3 Causal Mechanism Equivalence

Although the existing notions of disentanglement may be suitable for independent factors of variation [Khemakhem *et al.*, 2020], they fail to capture important information in a causal model where the factors are causally related. As formulated in Def. 2 and Def. 3, linear equivalence or permutation equivalence cannot capture the causal mechanisms accurately or distinguish the mechanisms afflicted to factors. For a counterexample to the definitions, refer to the appendix in [Komanduri *et al.*, 2024]. The framework of iVAE captures identifiability in the sense that the joint distributions of the latent variables of two different models are equivalent. However, for a causally factorized model, we have that $p_\theta(z) = p_{\hat{\theta}}(z)$ does not imply $p_\theta(z_i|z_{\mathbf{pa}_i}) = p_{\hat{\theta}}(z_i|z_{\mathbf{pa}_i})$. That is, the ground-truth causal factors and the learned causal factors should entail the same causal conditional mechanisms, where the minimal conditioning set is the set of causal parents.

Based on the intuition that causal models are described by mechanisms, we define a new notion of disentanglement that takes into account conditional distributions of causal variables under the Markov factorization. The new causal conditional equivalence preserves information about the independent causal mechanisms (ICM), which is a unique formulation for a causal model and important for performing correct interventions. The following two definitions describe the conditions necessary to satisfy causal mechanism equivalence.

Definition 5 (Causal Mechanism Permutation Equivalence)**.** *Let $\sim$ be an equivalence relation between $\hat{\theta}$ and θ, $\mathcal{X} = g(\mathcal{Z})$, and $\hat{\mathcal{X}} = \hat{g}(\mathcal{Z})$. If the factors z are causally related, we say that θ is causal mechanism permutation equivalent to $\hat{\theta}$ iff:*

1. *There exists a permutation matrix P such that $I = P \cdot J$ where I and J are indices of z and $\hat{z}$, respectively.*
2. *Given an equivalence pair (i, j), i.e., $P_{ij} \neq 0$, from this permutation matrix, one has $\mathbf{T}_i(z_i|z_{\boldsymbol{pa}_i}) = D_{ij}\hat{\mathbf{T}}_j(z_j|z_{\boldsymbol{pa}_j}), \forall z_i \in \mathcal{Z}_i, \forall \hat{z}_j \in \mathcal{Z}_j$, where D_{ij} is a scaling coefficient.*
3. *For all $i, j \in \{1, \ldots, n\}$, we have the mechanism equivalence $\boldsymbol{\lambda}_i(z_{\boldsymbol{pa}_i}, u) = D_{ij}\hat{\boldsymbol{\lambda}}_j(z_{\boldsymbol{pa}_j}, u)$, where D is a diagonal scaling matrix.*

Definition 6 (Causal Disentanglement)**.** *Given some ground-truth model θ, a learned model $\hat{\theta}$ is said to be* ***causally disentangled*** *if θ and $\hat{\theta}$ are causal mechanism permutation-equivalent.*

4 Proposed Framework

We design a framework to achieve *causal* disentanglement. We propose ICM-VAE, a VAE-based framework based on the independent causal mechanisms (ICM) principle that achieves disentanglement of causal mechanisms. Figure 2 shows the overall architecture of our proposed framework.

4.1 Structural Causal Flow

Rather than assuming the limiting linear causal graphical model (CGM), as done in CausalVAE [Yang *et al.*, 2021], we consider causal mechanisms to be complex nonlinear functions. Diverging from the strictly additive noise model assumption, we propose to parameterize causal mechanisms with a more general *diffeomorphic*[1] function. Flow-based models [Papamakarios *et al.*, 2021] are often quite expressive in low-dimensional settings, which makes them desirable for learning complex distributions due to efficient and exact evaluation of densities. We parameterize the causal mechanisms with a conditional flow, which we refer to as the latent structural causal flow (SCF), that learns to map the independent noise distribution to a distribution over causal variables. This module is inspired by the causal autoregressive flow [Khemakhem *et al.*, 2021]. This type of model is more realistic and general to better capture the complex distribution over the latent causal variables compared to simple linear mappings and leads to counterfactual identifiability [Nasr-Esfahany *et al.*, 2023]. The SCF, denoted as f^{RF}, is the reduced form (RF) of a nonlinear SCM function that conceptually maps noise variables ϵ to causal variables z as follows

$$z = f^{\text{RF}}(\epsilon) \tag{4}$$

where $f^{\text{RF}} : \mathcal{E} \to \mathcal{Z}$ is derived from the recursive substitution of causal mechanisms f_i in topological order of the causal graph as follows

$$z_i = f_i(\epsilon_i; z_{\mathbf{pa}_i}), \qquad \forall i \in \{1, \dots, n\} \tag{5}$$

realized as a function of the noise term and parent variables. The noise encoding ϵ_i is exactly the SCM noise variable corresponding to the causal variable z_i.

Similar to several prior works [Shen *et al.*, 2022; Liang *et al.*, 2023], we assume that the latent causal graph is known in the form of a binary adjacency matrix to focus on formulating the problem of causal disentanglement. To implement a diffeomorphic function f^{RF}, we use flow-based models to parameterize the causal mechanisms. Specifically, this flow is implemented as an affine-form autoregressive flow, where we derive each causal variable one at a time in topological order such that each variable is dependent only on a subset of previously derived variables (i.e. parents). Thus, the change of variables can be computed quite easily for exact and efficient likelihood estimation. In general, one can parameterize causal mechanisms using any nonlinear diffeomorphic function, as long as it takes into account the topological ordering. Let's take the pendulum example in Figure 1 to illustrate. The causal structure is $z_1 \to z_3, z_4$ and $z_2 \to z_3, z_4$. Then, the SCF would be $f^{\text{RF}} : (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \mapsto (z_1 = f_1(\epsilon_1), z_2 = f_2(\epsilon_2), z_3 = f_3(\epsilon_3, z_1, z_2), z_4 = f_4(\epsilon_4, z_1, z_2))$, where $z_i = f_i^{\text{RF}}(\epsilon_i; \epsilon_{\mathbf{pa}_i}) = f_i(\epsilon_i; z_{\mathbf{pa}_i})$ and each f_i is a diffeomorphic transformation of the form

$$z_i = f_i(\epsilon_i; z_{\mathbf{pa}_i}) = \exp(a_i) \cdot \epsilon_i + b_i \tag{6}$$

where $a_i = r_1(z_{\mathbf{pa}_i})$ and $b_i = r_2(z_{\mathbf{pa}_i})$ are the slope and offset parameters of the affine transformation, respectively, learned via neural networks r_1 and r_2 that capture information about the causal parents. Since the Jacobian of the function will be triangular by construction and the slope parameter is learned for each variable, the slope is equivalent to the diagonal elements of the Jacobian matrix as follows

$$\log \prod_i \left| \frac{\partial \epsilon_i}{\partial z_i} \right| = \sum_i \log \left| \frac{\partial f_i^{\text{RF}}(\epsilon_i; \epsilon_{\mathbf{pa}_i})}{\partial \epsilon_i} \right|^{-1} = \sum_i a_i \tag{7}$$

where $\epsilon_{\mathbf{pa}_i}$ denotes the noise terms associated with the parents of causal variable z_i. The structural causal flow can easily be generalized to multivariate scenarios by masking groups of latent codes corresponding to each causal variable.

4.2 Generative Model

To achieve an identifiable model, we leverage auxiliary information as a weak supervision signal [Khemakhem *et al.*, 2020]. Let $u \in \mathbb{R}^n$ be the auxiliary observed labels corresponding to the causally related ground-truth factors with support $\mathcal{U} \subset \mathbb{R}^n$. We assume that the decoder g is diffeomorphic onto its image. Several prior works [Locatello *et al.*, 2020; Khemakhem *et al.*, 2020; Lachapelle *et al.*, 2022] assume that the nonlinear mixing function mapping $\mathcal{Z}$ to $\mathcal{X}$ is a diffeomorphism. Consider the pendulum system from Figure 1 consisting of a light source, a pendulum, and a shadow. Given only the image, it is completely certain that we can identify where each object appears in the image. So, we find it reasonable to assume a diffeomorphic mixing function g for our exploration. Let $\theta = (g, \mathbf{T}, \boldsymbol{\lambda}, G^z)$ be the parameters of the conditional generative model defined as follows

$$p_\theta(x, \epsilon, z|u) = p_\theta(x|\epsilon, z) p_\theta(\epsilon, z|u) \tag{8}$$

where

$$p_\theta(x|\epsilon, z) = p_\theta(x|z) = p_\xi(x - g(z)) \tag{9}$$

If we assume that the distribution over the noise ξ is Gaussian with infinitesimal variance, we can model non-noisy observations as a special case of Eq. (9). The prior distribution in the generative model is given by

$$p_\theta(\epsilon, z|u) = p(\epsilon) p_\theta(z|u) \tag{10}$$

where we choose $p(\epsilon)$ as a standard Gaussian base distribution and $p(z|u)$ is assumed to be conditionally factorial. However, the conditional prior in Eq. (2) cannot properly capture causal mechanisms for causally related factors. We next define a causally factorized prior suitable to achieve causal disentanglement.

[1] A diffeomorphism is a differentiable bijection with a differentiable inverse.

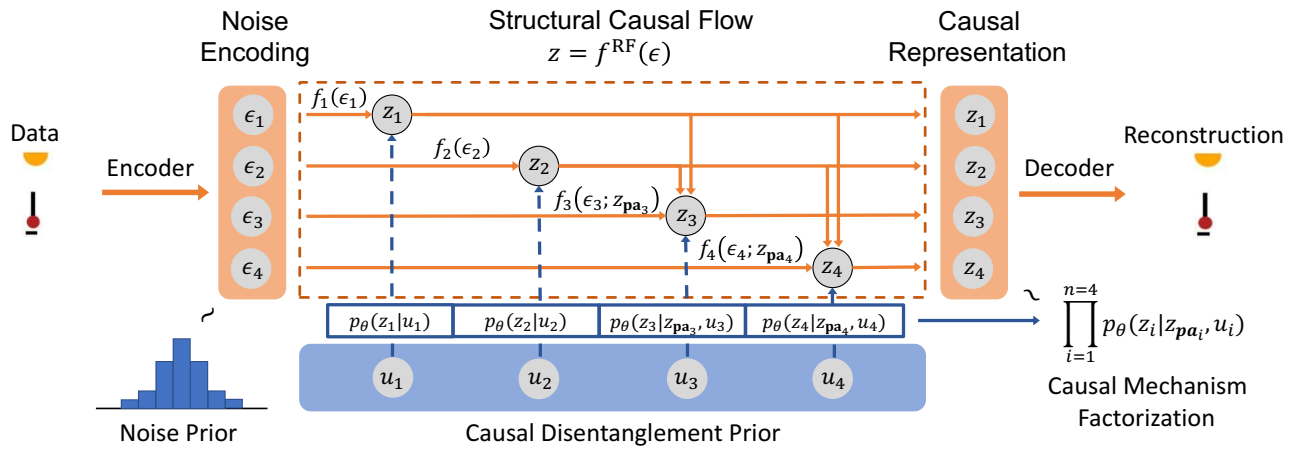

Figure 2: Architecture of ICM-VAE Framework, which contains two main components: (i) Structural Causal Flow (SCF), and (ii) Causal Disentanglement Prior. The blue color represents prior components and the orange represents the learning process.

4.3 Causal Disentanglement Prior

We aim to use a structured prior and perform conditioning in the latent space, similar to previous work on nonlinear ICA [Khemakhem *et al.*, 2020], to enforce z to be a disentangled causal representation. However, for a model incorporating causal structure, the form of the conditional prior in Eq. (2) needs to be modified and generalized to *causally* factorized distributions. To enforce the disentanglement of z, we parameterize the prior distribution to learn a mapping from u to z. That is, since the goal of causal disentanglement is to map each latent/mechanism to exactly one corresponding ground-truth factor/mechanism, we can explicitly incorporate this into the prior. Using u as our observational labels, we parameterize the factorized causal conditionals with a conditional flow between u and z to establish a bijective relationship. The goal is for the distribution over the causal variables to tend towards the learned prior. The prior over z is defined as follows

$$p_\theta(z|u) = \prod_{i=1}^{n} p_\theta(z_i|z_{\mathbf{pa}_i}, u_i) = \prod_{i=1}^{n} p(u_i) \left| \frac{\partial \boldsymbol{\lambda}_i(u_i; z_{\mathbf{pa}_i})}{\partial u_i} \right|^{-1} \tag{11}$$

$$p_\theta(z_i|z_{\mathbf{pa}_i}, u_i) = h_i(z_i) \exp(\mathbf{T}_i(z_i|z_{\mathbf{pa}_i}) \boldsymbol{\lambda}_i(G_i^z \odot z, u_i) - \psi_i(z, u)) \tag{12}$$

where $\boldsymbol{\lambda}_i(G_i^z \odot z, u_i)$ is the estimated parameter vector of the prior obtained via mechanism $\boldsymbol{\lambda}_i$, G_i^z is the ith column of the adjacency matrix of the causal graph of z, $h_i(z)$ is the base measure, and $\mathbf{T}_i(z) = (z, z^2)$ is the sufficient statistic. The prior induces a causal factorization of z with causal conditionals $p_\theta(z_i|z_{\mathbf{pa}_i}, u_i)$, where u_i is introduced as a weak supervision signal for identifiability. Eq. (11) is reminiscent of temporal priors that define a distribution over a latent variable conditioned on the variable at a previous time step [Lippe *et al.*, 2023]. In our case, we view the causal factors as derived autoregressively. With a slight abuse of notation, we define $\boldsymbol{\lambda}(z, u)$ to be the concatenation of all $\boldsymbol{\lambda}_i(G_i^z \odot z, u_i)$. The function $\boldsymbol{\lambda}(z, u)$ outputs the natural parameter vector for the causally factorized distribution. We further require $\boldsymbol{\lambda} : \mathcal{Z} \times \mathcal{U} \to \mathcal{Z}$ to be a bijective map between u and learned representation z to encourage disentanglement of the causal mechanisms. In practice, we choose $p(u)$ from a location-scale family such as Gaussian. The learned mechanism $\boldsymbol{\lambda}_i$ is defined as the following diffeomorphic map:

$$\boldsymbol{\lambda}_i(u_i; z_{\mathbf{pa}_i}) = \exp(c_i) \cdot u_i + d_i \tag{13}$$

where $c_i = s_1(z_{\mathbf{pa}_i})$ and $d_i = s_2(z_{\mathbf{pa}_i})$ are the slope and offset parameters of the flow, respectively, learned via neural networks. To obtain a causally factorized conditional prior over z, we map the base distribution $p(u)$, which is known beforehand, to a distribution over z.

4.4 Learning Objective

Putting all the components together, ICM-VAE consists of a stochastic encoder $q_\phi(\epsilon, z|x, u)$, a decoder $p_\theta(x|\epsilon, z)$, and diffeomorphic causal transformations $f_i(\cdot; \epsilon)$. All components are learnable and implemented as neural networks. Formally, we aim to optimize the following variational lower bound:

$$\begin{aligned} \log p_\theta(x, u) \geq \mathbb{E}_{\epsilon, z \sim q_\phi(\epsilon, z|x, u)} \Big[& \log p_\theta(x|\epsilon) + \log p_\theta(x|z) \\ & - \beta \{ \log q_\phi(\epsilon|x, u) + \log q_\phi(z|x, u) \\ & - \log p(\epsilon) - \log p_\theta(z|u) \} \Big] \end{aligned} \tag{14}$$

where β is the latent bottleneck parameter. We train the model by minimizing the negative of the ELBO loss and learn to map low-level pixel data to noise variables and map the noise variable distribution to a distribution over the causal variables. For a detailed derivation of the ELBO, refer to the appendix in [Komanduri *et al.*, 2024]. Causal structure learning could be heuristically incorporated jointly with the learning objective by enforcing acyclicity and sparsity of the causal graph. However, most causal discovery methods, such as GraN-DAG [Lachapelle *et al.*, 2020], require restricting parametric assumptions on the causal model (i.e., additive noise), to be practically applied. For an extended discussion on incorporating causal discovery and challenges, refer to the appendix in [Komanduri *et al.*, 2024].

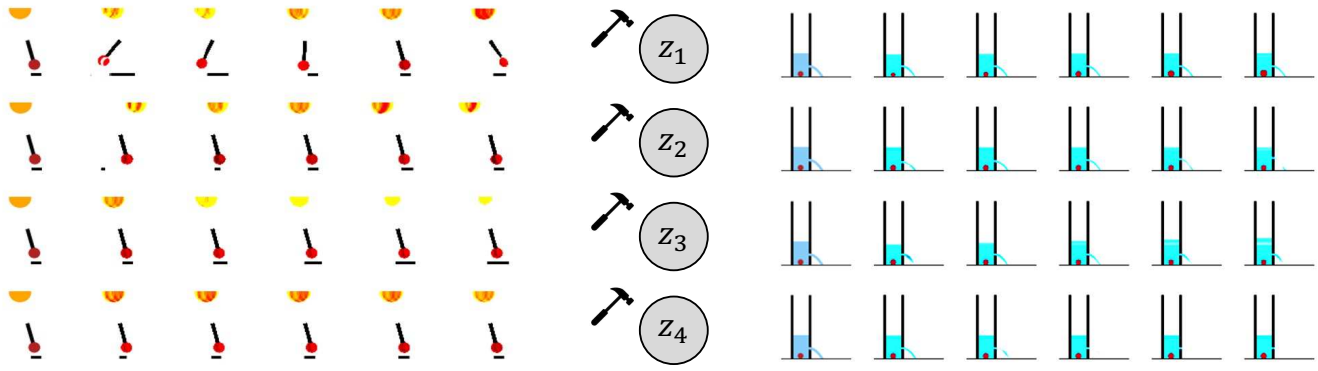

Figure 3: Pendulum (left) and Flow (right) counterfactual images after intervening on causal factors, individually, and propagating effects.

5 Identifiability Analysis

We design our framework to satisfy the conditions necessary to achieve causal mechanism equivalence and causally disentangle the factors of variation. The causally factorized prior in Eq. (11) induces disentanglement of causal mechanisms. Theorem 1 extends the identifiability theorem from iVAE [Khemakhem *et al.*, 2020] to show causal mechanism equivalence identifiability when we have a causal model. We note that causal mechanism disentanglement implies the disentanglement of causal factors.

Theorem 1 (Identifiability of ICM-VAE)**.** *Suppose that we observe data sampled from a generative model defined according to (8)-(12) with two sets of model parameters* $\theta = (g, \mathbf{T}, \boldsymbol{\lambda}, G^z)$ *and* $\hat{\theta} = (\hat{g}, \hat{\mathbf{T}}, \hat{\boldsymbol{\lambda}}, \hat{G}^z)$*. Suppose the following assumptions hold*

1. *The set* $\{x \in \mathcal{X} | \phi_\xi(x) = 0\}$ *has measure zero, where* ϕ_ξ *is the characteristic function of the density* p_ξ *defined in Eq. (9).*
2. *The decoder* g *is diffeomorphic onto its image.*
3. *The sufficient statistics* $\mathbf{T}_i$ *are diffeomorphic.*
4. ***[Sufficient Variability]*** *There exists* $nk+1$ *distinct points* $u_0, \ldots, u_{nk}$ *such that the matrix*

$$L = (\boldsymbol{\lambda}(z_{\boldsymbol{pa}_{(1)}}, u_{(1)}) - \boldsymbol{\lambda}(z_{(0)}, u_{(0)}), \ldots, \boldsymbol{\lambda}(z_{\boldsymbol{pa}_{(nk)}}, u_{(nk)}) - \boldsymbol{\lambda}(z_{(0)}, u_{(0)})) \tag{15}$$

of size $nk \times nk$ *is invertible, the ground-truth function* $\boldsymbol{\lambda}$ *is affected sufficiently strongly by each individual label* u_i *and previously derived variables* $z_{\boldsymbol{pa}_i}$*, and* $\forall i$*,* $\boldsymbol{\lambda}_i(z_{\boldsymbol{pa}_i}, u_i) \neq 0$*.*

Then θ *and* $\hat{\theta}$ *are causal mechanism permutation-equivalent, and the model* $\hat{\theta}$ *is causally disentangled.*

For the proof of Theorem 1, refer to the appendix in [Komanduri *et al.*, 2024].

6 Experimental Evaluation

In this section, we empirically evaluate the effectiveness of ICM-VAE. We consider a setting where the causal variables can be multi-dimensional and fix the dimension of each causal variable, which has also been explored in previous work [Yang *et al.*, 2021; Lippe *et al.*, 2023]. This assumption enables the representation to learn more informative and specific latent codes to describe the factors of variation (i.e. x-position, y-position, etc.). We run experiments on datasets with continuous factors, but discrete flows [Tran *et al.*, 2019] can be used for discrete factors. Our results suggest a component-wise correspondence between the learned and true causal factors.

Datasets. We show the performance of our framework on three datasets, each consisting of four real-valued causal variables. The Pendulum dataset [Yang *et al.*, 2021] consists of causal variables with causal graph (pendulum angle $\rightarrow$ shadow length, shadow position) and (light position $\rightarrow$ shadow length, shadow position). The Flow dataset [Yang *et al.*, 2021] consists of variables with causal graph (ball size $\rightarrow$ water height), (hole position $\rightarrow$ water flow), and (water height $\rightarrow$ water flow). We also show experiments on a more complex 3D dataset of a robot arm interacting with colored buttons called CausalCircuit [Brehmer *et al.*, 2022], which consists of variables with causal graph (robot arm $\rightarrow$ blue light intensity, green light intensity, and red light intensity), (blue light intensity $\rightarrow$ red light intensity), and (green light intensity $\rightarrow$ red light intensity).

Evaluation Metrics. The DCI metric [Eastwood and Williams, 2018] quantifies the degree to which ground-truth factors and learned latents are in one-to-one correspondence. We compute the DCI disentanglement (D) and completeness (C) scores, which are based on a feature importance matrix quantifying the degree to which each latent code is important for predicting each ground truth causal factor. The informativeness (I) score is the prediction error in the latent factors predicting the ground-truth generative factors and is constant ($I = 0$) throughout all datasets and models, so we omit it for brevity. We train models with 3 random seeds and select the median DCI score to report. We note that DCI is highly correlated with other disentanglement metrics, such as MCC, with strong connections to identifiability [Eastwood *et al.*, 2023]. To evaluate how changes in the generative factors affect the latent factors, we compute the interventional robustness score (IRS) [Suter *et al.*, 2019], which is similar to an R^2 value.

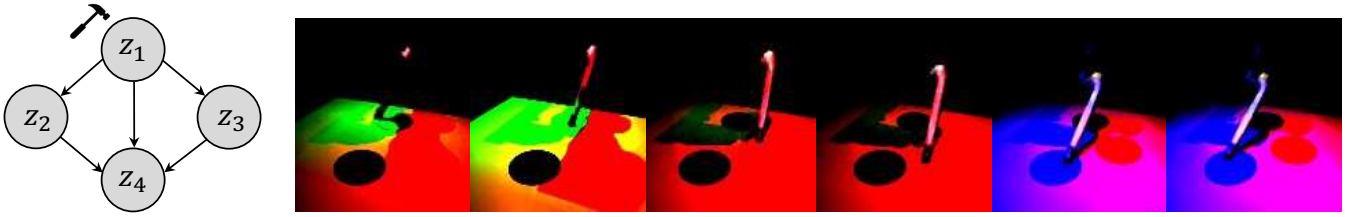

Figure 4: CausalCircuit counterfactual images by intervening on robot arm to turn on a colored light, which can causally affect other lights.

Implementation. For the Pendulum (6K training and 1K testing) and Flow (6K training and 2K testing) datasets, we linearly increase the β parameter throughout training from 0 to 1. We train for $9 \cdot 10^3$ steps using a batch size of 64. We use a Gaussian encoder and decoder with mean and variance computed by fully connected neural networks. For the CausalCircuit dataset (35K training and 10K testing), we linearly increase β from 0 to 0.05. We train for $3.5 \cdot 10^4$ steps using a batch size of 100. We use a convolutional neural network architecture with 6 layers and ReLU activation followed by a fully connected layer to estimate the mean and variance. The noise level for the variance of the Gaussian distribution of the conditional prior is controlled by $\sigma^2 \in \{0.01, 0.00001\}$. The structural causal flow and $\boldsymbol{\lambda}$ are implemented as affine form autoregressive flows with the slope and offset computed by fully connected 3-layer neural networks with 100 unit hidden layers and ReLU activation. We set the learning rate to 0.001 for all experiments. We set the dimension of each causal variable to 4 for all datasets. Our code is available at `https://github.com/Akomand/ICM-VAE`.

Baselines. We compare the performance of our approach, in terms of disentanglement and interventional robustness, with four baseline models: β-VAE [Higgins *et al.*, 2017] (unsupervised and acausal), iVAE [Khemakhem *et al.*, 2020] (acausal), CausalVAE [Yang *et al.*, 2021] (causal), and SCM-VAE [Komanduri *et al.*, 2022] (causal). The causal baselines propose relatively simplistic models that do not necessarily guarantee the disentanglement of causal factors.

Causal Disentanglement. Our experiments show that learning diffeomorphic causal mechanisms, rather than linear SCM, and incorporating the causal structure to learn a bijective $\boldsymbol{\lambda}$ to estimate the parameters of the causally factorized distribution significantly improves the disentanglement and interventional robustness of learned causal factors compared with baselines, as shown in Table 2. Consistent with our intuition, iVAE fails to disentangle the causal factors. The results indicate that ICM-VAE disentangles the causal factors and mechanisms almost perfectly. A high DCI disentanglement score indicates a permutation matrix mapping the latent factors to ground-truth generative factors in an ideal one-to-one mapping [Eastwood *et al.*, 2023]. Further, our model improves the interventional robustness of the representation, where interventions on ground-truth factors map to interventions on the corresponding learned factors.

Counterfactual Generation. We show counterfactual generated results of intervening on learned latent factors. Figure 4 shows the CausalCircuit system and the result of intervening on the robot arm factor and propagating causal effects. We observe that the red light also turns on as the robot arm interacts with the blue or green lights. On the other hand, when the arm interacts with the red light, only the red light turns on and the other lights remain off. We observe a similar phenomenon in the Pendulum and Water Flow systems in Figure 3, which shows the result of intervening on causal factors and propagating effects. Intervening on the pendulum angle or light position has causal effects on the shadow. However, interventions on the shadow factors do not change the parent factors. For the counterfactual generation procedure, results from intervening on other causal factors, and iVAE latent traversals, refer to appendix in [Komanduri *et al.*, 2024].

Dataset	Model	D	C	IRS
Pendulum	β-VAE	0.182	0.285	0.449
	iVAE	0.483	0.385	0.670
	CausalVAE	0.885	0.539	0.817
	SCM-VAE	0.764	0.475	0.829
	ICM-VAE (Ours)	**0.997**	**0.882**	**0.869**
Flow	β-VAE	0.308	0.332	0.452
	iVAE	0.730	0.481	0.674
	CausalVAE	0.819	0.522	0.707
	SCM-VAE	0.854	0.483	0.811
	ICM-VAE (Ours)	**0.988**	**0.598**	**0.893**
CausalCircuit	β-VAE	0.692	0.442	0.982
	iVAE	0.745	0.541	0.992
	CausalVAE	0.886	0.625	0.994
	SCM-VAE	0.867	0.652	0.993
	ICM-VAE (Ours)	**0.982**	**0.689**	**0.999**

Table 2: Causal Disentanglement of ICM-VAE and baselines

7 Conclusion

We propose ICM-VAE, a framework for causal representation learning under label supervision. We model causal mechanisms as flow-based transformations from noise to causal variables. We extend the idea of disentanglement to causal models and propose the notion of causal mechanism disentanglement. To this end, we design a causal disentanglement prior to causally factorize the distribution over causal variables. We theoretically show permutation-equivalent identifiability of the learned factors. Experimental results show that ICM-VAE almost perfectly disentangles the causal factors, improves interventional robustness, and generates consistent counterfactuals. Future work will incorporate causal discovery and disentanglement given partially observed labels.

Acknowledgements

This work is supported in part by National Science Foundation under awards 1910284, 1946391 and 2147375, the National Institute of General Medical Sciences of National Institutes of Health under award P20GM139768, and the Arkansas Integrative Metabolic Research Center at University of Arkansas.

References

[Ahuja *et al.*, 2023] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[Brehmer *et al.*, 2022] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

[Comon, 1994] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. Higher Order Statistics.

[Eastwood and Williams, 2018] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[Eastwood *et al.*, 2023] Cian Eastwood, Andrei Liviu Nicolicioiu, Julius Von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, and Bernhard Schölkopf. DCI-ES: An extended disentanglement framework with connections to identifiability. In *The Eleventh International Conference on Learning Representations*, 2023.

[Higgins *et al.*, 2017] 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 *International Conference on Learning Representations*, 2017.

[Hyvarinen *et al.*, 2018] Aapo Hyvarinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

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

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

[Joy *et al.*, 2021] Tom Joy, Sebastian Schmon, Philip Torr, Siddharth N, and Tom Rainforth. Capturing label characteristics in vaes. In *International Conference on Learning Representations*, 2021.

[Khemakhem *et al.*, 2020] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[Khemakhem *et al.*, 2021] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

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

[Komanduri *et al.*, 2022] Aneesh Komanduri, Yongkai Wu, Wen Huang, Feng Chen, and Xintao Wu. Scm-vae: Learning identifiable causal representations via structural knowledge. In *IEEE International Conference on Big Data (Big Data)*, 2022.

[Komanduri *et al.*, 2024] Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. *arXiv preprint arXiv:2306.01213*, 2024.

[Lachapelle *et al.*, 2020] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

[Lachapelle *et al.*, 2022] Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi LE PRIOL, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022.

[Liang *et al.*, 2023] Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, 2023.

[Lippe *et al.*, 2023] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

[Locatello *et al.*, 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 *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[Locatello *et al.*, 2020] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[Nasr-Esfahany *et al.*, 2023] Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[Papamakarios *et al.*, 2021] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(1), 2021.

[Parascandolo *et al.*, 2018] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[Peters *et al.*, 2017] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[Scholkopf *et al.*, 2021] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021.

[Schölkopf and von Kügelgen, 2022] Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning. *arXiv preprint arXiv:2204.00607*, 2022.

[Shen *et al.*, 2022] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

[Suter *et al.*, 2019] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[Tran *et al.*, 2019] Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, and Ben Poole. Discrete flows: Invertible generative models of discrete data. In *Advances in Neural Information Processing Systems*, 2019.

[Träuble *et al.*, 2021] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[von Kügelgen *et al.*, 2021] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

[Yang *et al.*, 2021] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.