Causal Representation Learning via Counterfactual Intervention

Xiutian Li[1*], **Siqi Sun**[1*], **Rui Feng**[1,2,3†]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433
[2]Fudan Zhangjiang Institute, Shanghai, 200120
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
lixt22@m.fudan.edu.cn, {siqi_sun, fengrui}@fudan.edu.cn

Abstract

Existing causal representation learning methods are based on the causal graph they build. However, due to the omission of bias within the causal graph, they essentially encourage models to learn biased causal effects in latent space. In this paper, we propose a novel causally disentangling framework that aims to learn unbiased causal effects. We first introduce inductive and dataset biases into traditional causal graph for the physical concepts of interest. Then, we eliminate the negative effects from these two biases by counterfactual intervention with reweighted loss function for learning unbiased causal effects. Finally, we employ the causal effects into the VAE to endow the latent representations with causality. In particular, we highlight that removing biases in this paper is regarded as a part of learning process for unbiased causal effects, which is crucial for causal disentanglement performance improvement. Through extensive experiments on real-world and synthetic datasets, we show that our method outperforms different baselines and obtains the state-of-the-art results for achieving causal representation learning.

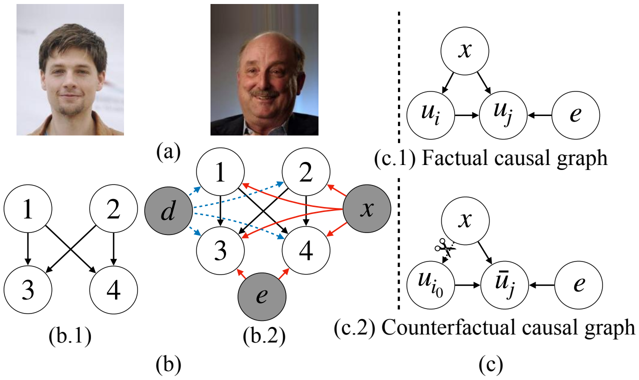
Introduction

Disentangled representation learning (DRL) (Bengio, Courville, and Vincent 2013) aims at identifying and separating underlying independent semantic factors of interest from observed data. Though DRL has achieved many advances (Gilpin et al. 2018; Creager et al. 2019; Zhu et al. 2020; Montero et al. 2020), these methods hold a collective assumption that the underlying semantic factors are mutually independent, which is inadaptable in most real circumstances. This is because the underlying semantic factors of interest are often causally related rather than mutually independent (Bengio et al. 2019). And not capturing causal relationships between underlying factors limits the generalization of disentanglement on causally related factors scenarios (Shen et al. 2022). To this end, the task of causal representation learning (Suter et al. 2019) has been proposed, for learning causal relationships among underlying semantic factors of interest during DRL.

To realize causal representation learning, many methods (Khemakhem et al. 2020; Yang et al. 2021; Brehmer et al. 2022; Lippe et al. 2022; Reddy, Balasubramanian et al. 2022; Shen et al. 2022) design complex mechanisms to endow underlying semantic factors of interest with causality learned from observed data. The important part of these methods is learning causal effects of factor pairs. However, these methods ignore the impact of inductive and dataset biases in the observed data on learning causal effects. Taking CelebA dataset (Liu et al. 2015) in Figure 1 (a) as an example, it includes semantic factors *age*, *gender*, *bald* and *beard*. Figure 1 (b.1) shows the causal graph constructed by current methods. The causal graph illustrates the causal relationships among *age*, *gender*, *bald* and *beard*. When models learn causal effect from *age* to *bald*, there are inductive biases, factors affect *bald* besides *age*, that influence learning causal effect. For instance, the image influences the values of both *age* and *bald*. If models observe that *bald* is always with sunglasses and not *bald* always occurs without sunglasses in images, there will be negative effects from the image to *bald* through sunglasses. Besides, some factors not in the image such as genes, habits and jobs also influence the value of *bald*. On this occasion, the causal effect from *age* to *bald* learned by models is mixed by direct causal effect from *age* to *bald* and negative effects from inductive biases. Moreover, dataset bias affects learning causal effect from *age* to *bald* as well. Concretely, when not *bald* sample size far exceeds *bald* sample size, the learned causal effect from *age* to *bald* will also be weakened, since *bald* sample is not easily observed. With inductive and dataset biases, causal effects they learned are biased that deviate from the actual values, and thus the performance of causal representation learning would be harmed.

To address the problem, in this paper, we propose CounterFactual Intervention Variational AutoEncoder (CFI-VAE) to remove the influences of inductive and dataset biases in causal representation learning. Given images and weak supervision annotations, CFI-VAE first learns an unbiased causal effect matrix by counterfactual intervention with reweighted loss function. Then, the images are transferred to independent exogenous variables through an encoder. The exogenous variables consist of independent representations corresponding to the physical concepts of interest and other information in images. With causal effect matrix, the independent representations are endowed with causality as causal representations through a nonlinear Structural Causal

[*]These authors contributed equally.
[†]Corresponding author.


(a)

(b.1) (b.2)

(b)

(c.1) Factual causal graph

(c.2) Counterfactual causal graph

(c)

Figure 1: The motivation and key idea of our approach. In part (a), we visualize the CelebA (Liu et al. 2015) with concepts of (1)*age*, (2)*gender*, (3)*bald* and (4)*beard*. In part (b), we show the comparisons between the traditional causal graph (b.1) and our constructed causal graph with inductive and dataset biases (b.2). In our causal graph, the red links denote the influences from inductive biases and the blue links denote the influence from dataset bias. Note that the blue links are dotted lines, because dataset bias has no causal effect on concept nodes while it can interfere with learning causal effect through concept nodes. In part (c), to remove inductive biases, we construct a causal graph with the nodes x (image), u_i (causal variable), u_j (result variable) and e (variables affect result variable while not in image, such as gene, habit and job), where x and e are inductive biases, and get the causal effect. Then, we adopt counterfactual intervention on causal variable u_i to cut off the impact from inductive biases on it and get the counterfactual causal effect. The inductive biases are removed by calculating the subtraction between the original and counterfactual causal effects.

Model (SCM) (Pearl 2009a). Finally, the model reconstructs the images by a decoder with causal representations and other information as input. In the process of CFI-VAE, the core is removing inductive and dataset biases for learning unbiased causal effects. For this, we introduce inductive and dataset biases into traditional causal graph as Figure 1 (b.2). Inspired by the causal inference methods (Pearl, Glymour, and Jewell 2016), we adopt counterfactual intervention to remove the negative effects from inductive biases. Specifically, as shown in Figure 1 (c), we construct a factual causal graph, which includes the inductive biases, the causal variable and the result variable. Then, we adopt counterfactual intervention on causal variable to cut off the impact from inductive biases on it. During this process, inductive biases are unchanged between the factual and counterfactual. Therefore, we can calculate the difference between the factual and counterfactual to remove the negative effects from inductive biases and obtain causal effect on result variable only from causal variable. For removing dataset bias, we improve the loss function that optimizes counterfactual intervention module by adjusting the weight of each class samples contributing to loss function. Through removing inductive and dataset biases, our model learns unbiased causal effects and improve the performance of causal representation learning.

In the experiments, we evaluate our method on real-world and synthetic datasets. The significant performance gained over baselines shows the effectiveness of our causal representation learning method. The contributions of our proposed method are summarized as follows:

- We show that inductive and dataset biases cause learning biased causal effects and harm the performance of causal representation learning.
- We propose a framework called CounterFactual Intervention Variational AutoEncoder (CFI-VAE) to remove inductive and dataset biases for learning unbiased causal effects and improve the performance of causal representation learning.
- We demonstrate that the proposed CFI-VAE is more effective than the state-of-the-art methods for achieving causally disentangling representation.

Related Work

Causal representation learning approaches. Current existing causal representation learning methods mostly fall broadly into two categories. The first category methods (Ahuja, Hartford, and Bengio 2022; Brehmer et al. 2022; Gresele et al. 2021; Lachapelle et al. 2022; Yao et al. 2021; Lippe et al. 2022) realize causal representation learning under supervision of ground truth counterfactual images generated according to causal graph. However, the ground truth counterfactual images are often not accessible in real-world cases, such as CelebA (Liu et al. 2015). The second category methods (Kocaoglu et al. 2017; Yang et al. 2021; Shen et al. 2022; Reddy, Balasubramanian et al. 2022) realize causal representation learning under supervision of annotations and causal graph. Particularly, (Kocaoglu et al. 2017) employed a generator neural network to fit the causal effects of causal graph in annotations, for implicit representation learning. (Yang et al. 2021) utilized VAE (Kingma and Welling 2013) network with an SCM (Pearl 2009a) to endow explicit representations with causal effects of causal graph in annotations, while the causal graph possessed spurious links. (Shen et al. 2022) adopted GAN (Goodfellow et al. 2014) architecture to learn causal effects of causal graph in annotations and transmit them in explicit representations through an SCM. However, all these methods ignore the impact of biases which affects the performance of learning causal effects.

Causal inference. Causal inference (Pearl et al. 2000) has garnered increasing attention in computer vision, including visual recognition (Tang et al. 2020), vision dialog (Liu et al. 2022), semantic segmentation (Zhang et al. 2020) and image reconstruction (Sauer and Geiger 2021), as it helps to discern causal effect from conventional correlation effect (Pearl 2009a). Concretely, causal inference utilizes the additional notion of intervention to remove bias effect for obtaining causal effect from conventional correlation effect (Pearl et al. 2000).

Removing biases. There are inductive bias and dataset bias influencing causally disentangling performance. For inductive bias, causal inference could remove it through causal intervention and counterfactual. (Yue et al. 2020; Wang et al.

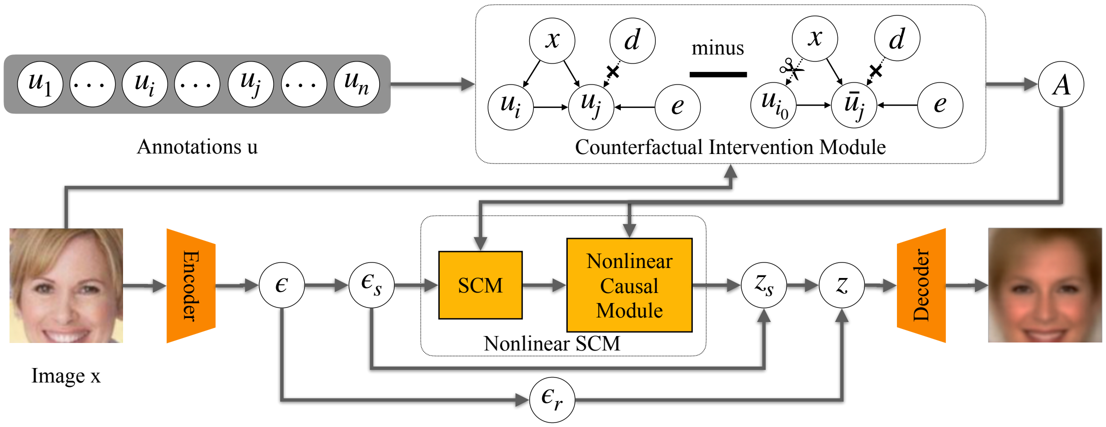


Figure 2: Our causally disentangling framework CFI-VAE. Given image x and annotations u, the model learns an unbiased causal effect matrix A. During this process, counterfactual intervention is adopted to remove inductive biases x and e, and the loss function ($\mathcal{L}_{causal}$) optimizing the counterfactual intervention module is reweighted to re-balance class distribution for removing dataset bias d. Then, the encoder transfers x to exogenous variables ϵ, which consist of independent representations ϵ_s corresponding to the physical concepts of interest and other information in the image ϵ_r. After that, ϵ_s are endowed with A as causal representations z_s through the nonlinear SCM. Next, causal representations z_s and other information ϵ_r are merged into latent representations z. Finally, the decoder reconstruct the image x with z as input.

2021; Zhao et al. 2022) utilized causal intervention to remove inductive bias effect for obtaining causal effect. (Tang et al. 2020; Bhat et al. 2022; Niu et al. 2021) propose that comparing between the factual and counterfactual will naturally remove the effect of bias, since the bias is the only thing unchanged between the two alternatives. Dataset bias is largely attributed to unbalance class distribution of datasets. The most widely-used solution is to re-balance the contribution of each class by re-sampling, re-weighting and logit adjustment (Zhang et al. 2023). Re-sampling methods (Hu et al. 2020; Mahajan et al. 2018) adjust class sample sizes for re-balancing. Re-weighting methods (Cao et al. 2019; Cui et al. 2019) attempt to endow different classes with different weights in loss function. Logit adjustment methods (Menon et al. 2020; Wu et al. 2021) adjust the prediction logits to alleviate the class imbalance.

Methodology

Problem Definition

Consider a set of images X with labels U, where each image $x \in X$ has annotations $u \in U$. The task of causally disentangling is to fit causal effects of causal graph in annotations u and inject them into latent representations of image x. The most important thing in the task is learning appropriate causal effects. Many methods design complex models to learn causal effects. However, they don't take inductive and dataset biases into account, which have negative effects on learning causal effects. Taking learning causal effect from annotation *age* to annotation *bald* as an example, for inductive bias, image x affects the values of two annotations simultaneously and some factors not in image such as genes, habits and jobs also affect *bald*. When these factors mix together, it's difficult to observe causal effect on *bald* only from *age*. For dataset bias, unbalanced class distribution between *bald* and not *bald* also harms learning causal effect from *age* to *bald*. Therefore, our goal is to remove inductive and dataset biases to learn unbiased causal effects for improving causally disentangling performance.

Overall Model

The architecture of CFI-VAE is depicted in Figure 2. Given image x and annotations u, we first obtain the unbiased causal effect matrix A through counterfactual intervention with reweighted loss function. Then, we transfer image x to independent Gaussian exogenous factors ϵ through an encoder. Exogenous factors ϵ consist of independent representations ϵ_s corresponding to the physical concepts of interest and other information in the image ϵ_r. With independent representations ϵ_s and matrix A, familiar with (Yang et al. 2021), we convert ϵ_s to causal representations z_s through a nonlinear SCM (Pearl 2009a). Note that causal representations z_s possess linear causality through structural equations, and then are extended to nonlinear cases through neural networks. Finally, we merge causal representations z_s and other information ϵ_r into latent representations z and z are transformed to image x through a Decoder. Overall, CFI-VAE is a three-step process that firstly extracts latent representations from the image, then injects causal effects of annotations into latent representations through SCM and finally reconstructs the image with latent representations. Therefore, to make sure the learned latent representations conform to actual representations of image, the objective of CFI-VAE is given by:

$$\mathcal{L}_{gen} = D_{KL}(q(z|x,u), p(z|x,u)) \qquad (1)$$

where D_{KL} represents the Kullback-Leibler (KL) divergence, x is the image, z are the latent representations, u are the annotations, q is learned representations distribution and p is actual representations distribution. However, the actual distribution $p(z|x, u)$ cannot be calculated directly (Doersch 2016), thus we convert the objective as Equation 2:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{rec} + \mathcal{L}_{kl} \\ &= -\mathbb{E}_{z\sim q_{E,F}(z|x,u)} log p_D(x|z, u) \\ &+ D_{KL}(q_{E,F}(z|x, u), p_\lambda(z|u))\end{aligned} \tag{2}$$

where E is the Encoder, D is the Decoder, F is the causal functions in nonlinear SCM and λ represents functions of z conditioning on u. Details of proof are given in Appendices. The first term $\mathcal{L}_{rec}$ can be calculated as Binary Cross-Entropy loss between input and reconstructed images. The second term $\mathcal{L}_{kl}$ represents KL divergence between $q_{E,F}(z|x, u)$ and $p_\lambda(z|u)$. The former is a variational posterior distribution and the latter is a factorized Gaussian conditional prior. In order to get $\mathcal{L}_{kl}$, we first obtain $q_{E,F}(z|x, u)$. In the nonlinear SCM, like (Yang et al. 2021), independent representations ϵ_s are transformed to linear causal representations z_l through $z_l = A^T z_l + \epsilon_s \Rightarrow z_l = (I - A^T)^{-1}\epsilon_s$, and to nonlinear causal representations z_s through neural networks f. Thus, $q_{E,F}(z|x, u)$ can be obtained by:

$$q_{E,F}(z|x, u) = (f((I - A^T)^{-1}\epsilon_s), \epsilon_r), \epsilon \sim N(0, I) \tag{3}$$

where I is an identity matrix, ϵ_r is other information in the image and A is the causal effect matrix we need to learn through $\mathcal{L}_{causal}$. Then, $p_\lambda(z|u)$ represents the Gaussian distribution of causal representations z_s conditioning on annotations u, given by (Yang et al. 2021):

$$\begin{aligned}p_\lambda(z|u) = (\Pi_i^n p_\lambda(z_{s_i}|u_i), \epsilon_r), \\ p_\lambda(z_{s_i}|u_i) = \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i))\end{aligned} \tag{4}$$

where λ_1 and λ_2 are functions of z_s conditioning on u, $\lambda_1(u) = u$ and $\lambda_2(u) = 1$ in this paper, and $p_\lambda(z|u)$ is denoted by sufficient statistics $T(z) = (\mu(z), \sigma(z))$.

Note that linear causal representations z_l are transformed to nonlinear causal representations z_s through neural networks f. Since neural networks are black-box and indeterminate, we need to make causal effects in z_s respect to counterparts in z_l. Like (Yang et al. 2021), we optimize f with loss function:

$$\mathcal{L}_f = \|z_s - f(A^T z_s; \eta) - \epsilon_s\|^2 \leq \kappa \tag{5}$$

where η are parameters of $f(\cdot)$ and κ is the small positive constant value.

Therefore, the overall loss of CFI-VAE is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{kl} + a\mathcal{L}_{causal} + b\mathcal{L}_f \tag{6}$$

where a and b denote regularization hyper-parameters and $(a, b) = (1, 1)$ in this paper.

Causal Graph

To systematically study how inductive bias influences learning causal effect, we construct a causal graph $\mathcal{G}$ (Pearl 2022) as Figure 1 (c.1). $\mathcal{G}$ consists of four variables: image x, causal annotation variable u_i, result annotation variable u_j and other variables e (affect result variable while not in image), where x and e are inductive biases. And the causal links in $\mathcal{G}$ show how inductive biases, causal and result variables interacting with each other. In a link $x \rightarrow u_i \rightarrow u_j$, x is the parent variable of u_i, and u_j is the child variable of u_i. This link represents there is a causal effect from parent variable to child variable. $x \rightarrow (u_i, u_j)$ indicates annotations are influenced by the image. In causal inference theory (Pearl 2009b), variable x is a confounder which influences estimating causal effect of variable u_i on u_j, because x simultaneously affects u_i and u_j (Chen et al. 2021). The link $e \rightarrow u_j$ indicates variable u_j is influenced by e besides x and u_i. For example, when estimating causal effect of *age* on *bald*, the causal effect observed by models deviated from the actual value, because it is mixed by effects from image, *age* and other factors (gene, habits and jobs) together.

Learning Unbiased Causal Effects

Since causal effect observed by models is direct causal effect from causal variable to result variable, mixing with effect from inductive biases. Inspired by causal inference methods (Pearl 2009a, 2022), we employ counterfactual intervention to separate direct causal effect from effect of inductive biases. Concretely, counterfactual intervention wipes out causal variable through operation $do(\cdot)$, while inductive biases are not affected and maintain the original values. In this case, causal effect observed by models is just effect given by inductive biases. Through calculating the subtraction between two causal effects before and after counterfactual intervention, we elegantly separate direct causal effect from effect of inductive biases, as shown in Figure 1 (c).

Current methods estimate causal effects through a classifier (Chen et al. 2021; Tang et al. 2020; Reddy, Balasubramanian et al. 2022). Concretely, to learn causal effect from annotation u_i to annotation u_j, given image x and annotations including u_i except u_j as inputs, we train a classifier to predict u_j. In this case, prediction results of classifier should respect to the class of u_j. Then, we wipe out u_i from inputs by setting it to zero, prediction results should deviate from the class of u_j, since u_i is a cause of u_j. Therefore, we optimize the classifier by loss function:

$$\mathcal{L}_{causal} = CE[C(u_i, \Omega), u_j] - \gamma CE[C(u_{i_0}, \Omega), u_j], \tag{7}$$

where CE is Cross-Entropy loss, C is classifier, u_i is the original value, u_{i_0} denotes that wiping out u_i by setting it to zero, Ω are other inputs of classifier and γ denotes a dynamic parameter.

In this way, we define causal effect observed by models as prediction logits Y for class of u_j (Tang, Huang, and Zhang 2020). Then, in the field of causal inference, the direct causal effect from u_i to u_j is calculated as Total Direct Effect (TDE) (VanderWeele 2013; Pearl 2022):

$$TDE(u_i \rightarrow u_j) = Y_{u_i}(\Omega) - Y_{u_{i_0}}(\Omega) \tag{8}$$

Now we remove the influence of inductive bias on learning causal effect. However, dataset bias also harms learning causal effect. Concretely, the momentum in Adam optimizer (Kingma and Ba 2014) or SGD optimizer (Qian 1999), is influenced by unbalanced class distribution of datasets (Tang, Huang, and Zhang 2020). Before analyzing how dataset bias influences learning causal effects with classifiers, we take a brief review on the Adam optimizer with momentum (Kingma and Ba 2014):

$$\begin{aligned} & m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t, \\ & \hat{m}_t = m_t/(1-\beta_1^t), \\ & \theta_t = \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) \end{aligned} \tag{9}$$

where in the t-th iteration: model parameters θ_t, gradient g_t, momentum m_t, momentum decay ratio β_1, stepsize α and second raw moment estimate $\hat{v}_t$.

Theorem 1 *Let m is momentum in Adam optimizer, and it depends on the gradient over all classes. For a balanced dataset, m is equally contributed by every class. When class distribution becomes unbalanced, m will be dominated by head classes. In this way, classifier trained under unbalanced m would generate inappropriate prediction logits (causal effects).*

Based on Theorem 1, we adopt a reweighted loss function (Cui et al. 2019) to remove influence from dataset bias. Concretely, we convert $\mathcal{L}_{causal}$ as:

$$\mathcal{L}_{causal} = \tfrac{1-r}{1-r^k}(CE[C(u_i,\Omega),u_j] - \gamma CE[C(u_{i_0},\Omega),u_j]), \tag{10}$$

where r is a hyper-parameter and k is sample number of the class of u_j and $r = 0.5$ in this paper.

In this way, we remove influences of inductive and dataset biases and obtain unbiased causal effect matrix A to guide causal representation learning.

Identifiability Analysis

In this section, we show that our model is identifiable. The core of the problem is that the true parameters θ and the learned parameters $\tilde{\theta}$ by hypothetical functions should be equivalent (Yang et al. 2021). Therefore, we define our causally disentangling model (CDM) firstly:

Definition 1 *(Causally disentangling model (CDM)) A causally disentangling model $\mathcal{M}_\theta = \langle E, S, D, T, \lambda\rangle$ consists of*

- *an encoder E: image $\mathcal{X} \to$ exogenous variables ϵ,*
- *a nonlinear SCM S, which faithfully respects to its causal graph,*
- *a decoder D: latent representations $\mathcal{Z} \to$ image $\mathcal{X}$,*
- *sufficient statistics T, and*
- *conditioning functions λ.*

With Definition 1, we define the true causally disentangling model as $\mathcal{M}_\theta = \langle E, S, D, T, \lambda\rangle$ and the learned causally disentangling model as $\tilde{\mathcal{M}}_{\tilde{\theta}} = \langle \tilde{E}, \tilde{S}, \tilde{D}, \tilde{T}, \tilde{\lambda}\rangle$. Then, the problem of equivalence between θ and $\tilde{\theta}$ is transformed to equivalence between CDMs $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_{\tilde{\theta}}$. Similar to (Brehmer et al. 2022), we define two CDMs are equivalent as:

Definition 2 *(CDM equivalence) Let $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_{\tilde{\theta}}$ be two CDMs with identical observation space. A mapping between them is defined as ψ for exogenous variables and latent representations that tells us how to reparameterize them, such that encoder, nonlinear SCM, decoder, sufficient statistics and conditioning function of $\tilde{\mathcal{M}}_{\tilde{\theta}}$ are compatible with corresponding elements of $\mathcal{M}_\theta$ reparameterized through the mapping ψ. $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_{\tilde{\theta}}$ are equivalent, $\mathcal{M}_\theta \sim \tilde{\mathcal{M}}_{\tilde{\theta}}$, if and only if there is a CDM mapping ψ between them.*

Reviewing the causally disentangling process, we can summarize the true causally disentangling process and the learned causally disentangling process as follows:

$$\begin{aligned} & x \sim \mathcal{X}, \epsilon = E(x), z = (S(\epsilon_s), \epsilon_r), x = D(z), \\ & x \sim \mathcal{X}, \tilde{\epsilon} = \tilde{E}(x), \tilde{z} = (\tilde{S}(\tilde{\epsilon}_s), \tilde{\epsilon}_r), x = \tilde{D}(\tilde{z}). \end{aligned} \tag{11}$$

Based on Equation 11, then we could define a mapping:

$$\psi : \begin{cases} T(E(x)) = L_1\tilde{T}(\tilde{E}(x)) + l_1 : \epsilon \to \tilde{\epsilon} \\ T(D^{-1}(x)) = L_2\tilde{T}(\tilde{D}^{-1}(x)) + l_2 : \mathcal{Z} \to \tilde{\mathcal{Z}} \end{cases} \tag{12}$$

where decoder function D is differentiable and the Jacobian matrix of D is of full rank, the sufficient statistics $T \neq 0$ almost everywhere for z, L_1 is an invertible matrix, L_2 is an invertible diagonal matrix with diagonal elements associated to u and l_1, l_2 are vectors.

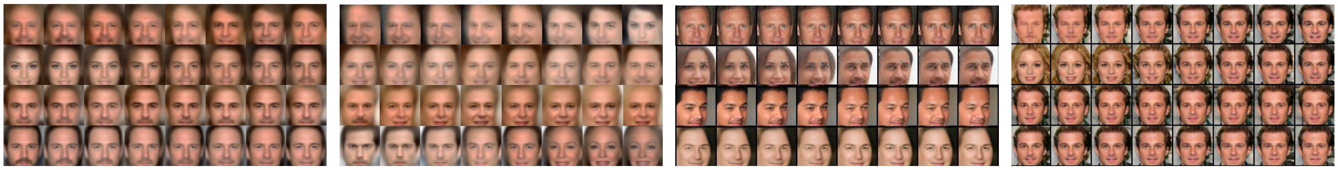
Through this, we can show that ψ is a mapping that proves CDM equivalence $\mathcal{M}_\theta \sim \tilde{\mathcal{M}}_{\tilde{\theta}}$ and makes learned parameters θ are compatible with the true ones. The detailed proof is available in Appendices.

Experiments

Experimental Setups

Datasets. we conduct experiments on real-world CelebA dataset (Liu et al. 2015) and synthetic datasets. ***1)*** CelebA dataset contains face images with 40 attributes annotations. Following previous causally disentangling methods (Yang et al. 2021; Shen et al. 2022), we select images with different attributes subsets as **Age** dataset and **Smile** dataset respectively. For causal graph, *age* and *gender* influence *bald* and *beard* in Age dataset, and *gender*, *smile* and *mouth* influence *eye* in Smile dataset. In CelebA dataset, image, gene, habit and job are inductive biases and unbalanced class distribution is dataset bias. For both Age and Smile datasets, we randomly select 30,000 images for training and 7,500 images for test. ***2)*** For synthetic datasets, we build Pendulum (**Pend**) dataset and **Flow** dataset like (Yang et al. 2021). For causal graph, *pendulum angle* and *light position* influence *shadow length* and *shadow position* in Pend dataset. And *ball size* influences *water height*, and *water height* and *hole position* influence *water flow* in Flow dataset. In synthetic datasets, the original states of physical concepts are inductive biases and no dataset bias because of balanced class distribution. For both Pend and Flow, we build 5,000 images for training and 1,000 images for test. Due to pages limitations, we visualize causal graphs of all datasets in Appendices.

Baselines. We compare our model with three state-of-the-art methods on causal representation learning under supervision of annotations: CausalVAE (Yang et al. 2021), CausalGAN (Kocaoglu et al. 2017), and DEAR (Shen et al. 2022). Among these models, causal graphs are given correctly in CausalGAN and DEAR, while with spurious relationships in CausalVAE. To compare fairly, we remove spurious relationships in CausalVAE as CausalVAE-D.



(a) CFI-VAE (b) CausalVAE-D (c) CausalGAN (d) DEAR

Figure 3: Intervention results of CFI-VAE, CausalVAE-D, CausalGAN and DEAR on Age dataset, where the intervention factors are *age*, *gender*, *bald* and *beard* respectively from up to bottom.

	MIC↑				**TIC↑**				**ISR(%)↑**				**CERD(%)↑**				**SCR(%)↓**			
Method	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**
CausalVAE	0.609	0.800	0.937	0.721	6.238	8.873	10.553	8.192	**100**	**100**	**100**	**100**	50	**100**	**100**	67	63	25	38	55
CausalVAE-D	0.608	0.802	0.951	0.773	6.427	8.973	10.845	7.452	**100**	**100**	**100**	75	75	75	**100**	67	38	25	38	11
CausalGAN	0.244	0.308	-	-	1.466	2.018	-	-	75	**100**	-	-	50	50	-	-	25	13	-	-
DEAR	0.554	0.660	0.373	0.406	4.884	5.921	1.868	1.968	75	75	75	75	25	75	**100**	67	13	25	**0**	100
CFI-VAE	**0.626**	**0.873**	**0.997**	**0.992**	**7.495**	**10.429**	**11.534**	**11.291**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**0**	**0**	**0**	**0**

Table 1: Comparison with the state-of-the-arts on Age, Smile, Pend and Flow datasets. MIC is a normalized metric and TIC is based on 64 batch size. Best performances are bold. Note that CausalGAN model setup is based on binary classification datasets and cannot be used for multiclass classification datasets (Pend and Flow).

Evaluation Metrics. Similar to (Yang et al. 2021), we adopt Maximal Information Coefficient (MIC) and Total Information Coefficient (TIC) (Kinney and Atwal 2014), which estimates disentanglement ability that how much physical meaning that latent representations possess. However, MIC and TIC cannot reflect whether model learned causality among representations. Therefore, we propose Causal Effect Realization Degree (CERD) and Spurious Correlation Rate (SCR) for evaluating causality. CERD measures the ratio of the number of learned causal relationships to the number of causal relationships set in advance, and SCR measures the ratio of the number of learning spurious relationships to the number of all spurious relationships. Besides, we propose Intervention Success Rate (ISR), which measures the ratio of intervening representations successfully, as an auxiliary verification for MIC and TIC. In the experiments, CERD, SCR and ISR are evaluated through randomly sampling a set of images where representations are intervened respectively.

Implementation Details. We train and evaluate our model on both real-world and synthetic datasets. Similar to CausalVAE (Yang et al. 2021), the real-world CelebA dataset are resized at 128 × 128 resolution while the synthetic dataset are built at 96 × 96 resolution. During training, batch size is set as 64. For optimizing, we utilize Adam (Kingma and Ba 2014) with Cosine Annealing (Loshchilov and Hutter 2016). More details are given in Appendices.

Quantitative Studies

To verify the effectiveness of our method, we compare the performance of CFI-VAE with baseline methods as in Table 1. Our model can consistently outperform different baselines on both real-world datasets and synthetic datasets, justifying the effectiveness of our framework in causal representation learning. Concretely, for disentanglement, CFI-VAE raises MIC to 0.872 on average and TIC to 10.187 on average. In contrast, the representations learned by those compared methods do not correspond to the causal concepts of interest well. These results illustrates that CFI-VAE possesses better disentanglement ability than baseline methods. For causality, which is evaluated using CERD and SCR, and the results in Table 1 demonstrate that our model captures all causalities according to causal graph set in advance when implementing intervention on latent representations. In contrast, CERD and SCR of baseline methods illustrate that they cannot learn all causalities and capture spurious relationships. These results suggest that our model endows latent representations with better causality than baselines.

Qualitative Studies

To qualitatively analyze how our method realizes better causally disentangling than baselines, we obtain intervened images through performing interventions on latent representations and observe the validity of intervened images corresponding to the causal graph set in advance. Taking **Age** dataset as an example, Figure 3 shows intervention images of CFI-VAE, CausalVAE-D, CausalGAN and DEAR respectively. For CFI-VAE, we can observe that when *age* is intervened, *i.e.* old turns young, *bald* and *beard* change, and that when *gender* is intervened, *i.e.* female turns male, *bald* and *beard* change. And when *bald* is intervened, *i.e.* one turns balder, *age*, *gender* and *beard* are unaffected, and when *beard* is intervened, *i.e.* beard disappears, *age*, *gender* and *bald* are unaffected. These observation results strictly conform to the causal graph of Age dataset set in advance as shown in Figure 1 (b.2). For CausalVAE-D, we can observe that intervention on *age* changes *gender*, that intervention on *bald* changes *gender* and *beard*, and that intervention on *beard* changes *gender*. For CausalGAN, intervention on *age*

	MIC↑				**TIC↑**				**ISR(%)↑**				**CERD(%)↑**				**SCR(%)↓**			
Method	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**
CFI-VAE w/db	0.619	0.860	-	-	7.223	10.240	-	-	50	**100**	-	-	25	25	-	-	**0**	**0**	-	-
CFI-VAE w/ib	0.615	0.851	0.785	0.973	6.466	9.125	8.091	10.751	**100**	50	**100**	**100**	25	75	**100**	33	**0**	**0**	13	11
CFI-VAE	**0.626**	**0.873**	**0.997**	**0.992**	**7.495**	**10.429**	**11.534**	**11.291**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**0**	**0**	**0**	**0**

Table 2: Ablation analysis for removing biases process of our method on Age, Smile, Pend and Flow datasets. *db* refers to dataset bias and *ib* refers to inductive bias. Best performances are bold. Note that Pend and Flow datasets do not have dataset bias because of obeying uniform distribution.

	MIC↑				**TIC↑**				**ISR(%)↑**				**CERD(%)↑**				**SCR(%)↓**			
Method	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**	**Age**	**Smile**	**Pend**	**Flow**
CausalVAE	0.380	0.520	0.801	0.600	3.202	4.529	8.602	4.933	**100**	75	**100**	75	25	50	**100**	67	38	63	50	33
CausalVAE-D	0.397	0.539	0.764	0.643	3.340	4.779	8.024	6.184	**100**	**100**	**100**	75	50	**100**	**100**	67	38	63	50	33
CausalGAN	0.198	0.258	-	-	0.931	1.008	-	-	75	**100**	-	-	25	25	-	-	**13**	**13**	-	-
DEAR	0.248	0.325	0.245	0.241	1.377	1.884	1.624	1.653	50	50	75	50	25	25	**100**	**100**	25	63	75	44
CFI-VAE	**0.461**	**0.664**	**0.921**	**0.953**	**5.391**	**6.319**	**9.807**	**10.260**	**100**	**100**	**100**	**100**	**75**	**100**	**100**	**100**	**13**	**13**	**25**	**11**

Table 3: Comparison with the state-of-the-arts on Age, Smile, Pend and Flow datasets under proportion of labeled samples as 10%. Best performances are bold.

cannot change *beard*, and *beard* is not intervened successfully. For DEAR, intervention on *age* cannot change *bald*, *bald* is not intervened successfully and intervention on *beard* changes *age*. All these spurious results of baselines violate the causal graph set in advance. Due to pages limitations, we present qualitative results on other datasets that **Smile** of CelebA dataset and **synthetic** datasets in Appendices.

Ablation Studies

For all of our ablation studies, we conduct experiments on **Age**, **Smile** of CelebA datasets and **Pend** and **Flow** of synthetic datasets.

Ablation study for removing biases process. To demonstrate the effectiveness of removing inductive bias and dataset bias in learning causal effects to causally disentangling, we maintain two kinds of bias respectively, and other components are kept unchanged. The ablation analysis results are shown in Table 2. With inductive bias, MIC and TIC are 0.806 and 8.608 on average, which are obviously less than the full model. And CERD and SCR results are also weakened. These results illustrate that the model cannot realize both disentangled representation and causality well with inductive bias in learning causal effects. With dataset bias, though MIC, TIC and ISR are comparative to the full model, CERD is weakened. These results illustrate that the model could not realize endowing disentangled representation with causality well with dataset bias in learning causal effects.

Ablation study for proportion of supervision annotations. Since DEAR (Shen et al. 2022) expands causally disentangling task under semi-supervision besides full-supervision. To demonstrate full proportion supervision are needed for causally disentangling, we reduce the proportion of supervision annotations according to DEAR's setting that supervision proportion of 10%, 1% and 0.1%, and other components are kept unchanged. Due to pages limitations, we only present the results under supervision proportion of 10% in Table 3, and results under other supervision proportions are in Appendices. MIC and TIC are 0.749 and 7.944 on average for supervision proportion of 10% for CFI-VAE. For all methods, the decline of MIC and TIC illustrates that the performance of disentangling representation is greatly weakened as supervision proportion reduces. And this directly leads that model cannot intervene latent representations successfully in intervention experiments, which is shown as the decline of ISR. Besides, the drop of supervision proportion also leads that model fails to endow latent representations causality according to causal graph set in advance, which is shown by CERD and SCR. Moreover, we can observe that CFI-VAE outperforms different baselines in causally disentangling on different supervision proportions. In summary, the ablation analysis illustrates that full proportion of supervision is still needed for causally disentangling task strictly conforming to causal graph set in advance.

Conclusion

In this paper, we propose a novel causally disentangling framework called CFI-VAE, aiming to remove inductive and dataset biases for learning unbiased causal effects in the latent space. Counterfactual intervention is applied by replacing the factual annotations of images with the counterfactual ones, and get factual causal effect and counterfactual causal effect through the classifier. By calculating the subtraction of two causal effects, the inductive bias is removed. For removing dataset bias, loss function optimizing counterfactual intervention module is reweighted to re-balance class distribution. Through removing inductive and dataset biases, our method can learn unbiased causal effects for improving performance of causally disentangling. In the experiment, we evaluate our method on real-world and synthetic datasets. The improvement in performance over various baselines demonstrates the effectiveness of our method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62172101), the Science and Technology Major Project of Commission of Science and Technology of Shanghai(No.21XD1402500), Regional social experimentation with the "Clinical Decision Support System for Paediatric Outpatient Clinics" (No. 21002411800); Auxiliary Diagnosis and Rare Disease Screening System for Children's Pneumonia (No. yg2022-7).

References

Ahuja, K.; Hartford, J. S.; and Bengio, Y. 2022. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35: 15516–15528.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Bhat, S.; Jiang, J.; Pooladzandi, O.; and Pottie, G. 2022. De-Biasing Generative Models using Counterfactual Methods. *arXiv preprint arXiv:2207.01575*.

Brehmer, J.; De Haan, P.; Lippe, P.; and Cohen, T. S. 2022. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35: 38319–38331.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Chen, G.; Li, J.; Lu, J.; and Zhou, J. 2021. Human trajectory prediction via counterfactual analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9824–9833.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Doersch, C. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 118.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gresele, L.; Von Kügelgen, J.; Stimper, V.; Schölkopf, B.; and Besserve, M. 2021. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34: 28233–28248.

Hu, X.; Jiang, Y.; Tang, K.; Chen, J.; Miao, C.; and Zhang, H. 2020. Learning to segment the tail. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14045–14054.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR.

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kinney, J. B.; and Atwal, G. S. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9): 3354–3359.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2017. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*.

Lachapelle, S.; Rodriguez, P.; Sharma, Y.; Everett, K. E.; Le Priol, R.; Lacoste, A.; and Lacoste-Julien, S. 2022. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, 428–484. PMLR.

Lippe, P.; Magliacane, S.; Löwe, S.; Asano, Y. M.; Cohen, T.; and Gavves, S. 2022. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, 13557–13603. PMLR.

Liu, B.; Wang, D.; Yang, X.; Zhou, Y.; Yao, R.; Shao, Z.; and Zhao, J. 2022. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18041–18050.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Van Der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 181–196.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Montero, M. L.; Ludwig, C. J.; Costa, R. P.; Malhotra, G.; and Bowers, J. 2020. The role of disentanglement in generalisation. In *International Conference on Learning Representations*.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.

Pearl, J. 2009a. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

Pearl, J. 2009b. *Causality*. Cambridge university press.

Pearl, J. 2022. Direct and indirect effects. In *Probabilistic and causal inference: The works of Judea Pearl*, 373–392.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. Causal inference in statistics: A primer. 2016. *Internet resource*.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Qian, N. 1999. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1): 145–151.

Reddy, A. G.; Balasubramanian, V. N.; et al. 2022. On causally disentangled representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8089–8097.

Sauer, A.; and Geiger, A. 2021. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*.

Shen, X.; Liu, F.; Dong, H.; Lian, Q.; Chen, Z.; and Zhang, T. 2022. Weakly Supervised Disentangled Generative Causal Representation Learning. *Journal of Machine Learning Research*, 23: 1–55.

Suter, R.; Miladinovic, D.; Schölkopf, B.; and Bauer, S. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 6056–6065. PMLR.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33: 1513–1524.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

VanderWeele, T. J. 2013. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 224–232.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3091–3100.

Wu, T.; Liu, Z.; Huang, Q.; Wang, Y.; and Lin, D. 2021. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8659–8668.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021. CausalVAE: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9593–9602.

Yao, W.; Sun, Y.; Ho, A.; Sun, C.; and Zhang, K. 2021. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional few-shot learning. *Advances in neural information processing systems*, 33: 2734–2746.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, H.; Ma, C.; Dong, X.; Luu, A. T.; Deng, Z.-H.; and Zhang, H. 2022. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning*, 26958–26970. PMLR.

Zhu, Y.; Min, M. R.; Kadav, A.; and Graf, H. P. 2020. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6538–6547.