Counterfactual Image Editing

Yushu Pan [1] **Elias Bareinboim** [1]

Abstract

Counterfactual image editing is a challenging task within generative AI. The current literature on the topic focuses primarily on changing individual features while being silent about the causal relationships between features, which are present in the real world. In this paper, we first formalize this task through causal language, modeling the causal relationships between latent generative factors and images through a special type of causal model called *augmented structural causal models (ASCMs)*. Second, we show two fundamental impossibility results: (1) counterfactual editing is impossible from i.i.d. image samples and their corresponding labels alone; (2) also, even when the causal relationships between latent generative factors and images are available, no guarantees regarding the output of the generative model can be provided. Third, we propose a relaxation over this hard problem aiming to approximate the non-identifiable target counterfactual distributions while still preserving features the users care about and that are causally consistent with the true generative model, which we call **ctf-consistent estimators**. Finally, we develop an efficient algorithm to generate counterfactual image samples leveraging neural causal models.

1. Introduction

Counterfactual reasoning is a critical component of our cognitive system. It is essential for solving various tasks, including assigning credit, determining blame and responsibility, understanding why events occurred in a particular way and articulating explanations, and generalizing across changing conditions and environments (Pearl & Mackenzie, 2018; Bareinboim et al., 2022; Correa et al., 2021a). More recently, there has been a growing interest in counterfactual questions regarding image generation and editing. For instance, one might ask "how would the image change had the dog been a cat?" or "What would the image look like had the person been smiling?". Addressing these prototypical counterfactual questions is challenging and requires the understanding of the causal relationships between the features, with practical applications in various downstream tasks, including data augmentation, fairness analysis, generalizability, and transportability (Bareinboim et al., 2015; Schölkopf et al., 2021; Lee et al., 2020; Mao et al., 2022).

Some initial methods for counterfactual image editing tasks typically involve searching for adversarial samples (Goyal et al., 2019b; Wang & Vasconcelos, 2020; Dhurandhar et al., 2018). For example, (Dhurandhar et al., 2018) proposed a minimum-edit counterfactual method that aims to identify the minimum and most effective perturbations needed to change the classifier's prediction. With the ability to generate high-quality synthetic images from a latent space through GANs (Brock et al., 2019; Karras et al., 2019), VAEs (Child, 2021; Vahdat & Kautz, 2020), and Diffusion Models (Ho et al., 2020; Song et al., 2021), recent approaches edit images by manipulating vectors in the latent space (Shen et al., 2020; Härkönen et al., 2020; Khorram & Fuxin, 2022; Chai et al., 2021).

More recently, text information has also been leveraged in image editing tasks. The image description in text is beneficial to the encoding process and guiding manipulations in the latent space (Radford et al., 2021; Avrahami et al., 2022; Crowson et al., 2022; Gal et al., 2022; Patashnik et al., 2021); also the natural editing instruction text can be directly used to prompt the transition from the original to the counterfactual images (Brooks et al., 2023). However, such approaches focus primarily on changing a single categorical label of a given image, and more fundamentally, do not take the causal relationships among the underlying generative factors into account. We illustrate the challenge when multiple features are involved in the generation task next.

Example 1.1. Consider a dataset of human faces. Based on our understanding of human anatomy and facial expressions, we know that both $Gender$ and Age do not causally affect each other while Age does affect $Hair\ color$. Meanwhile, the dataset collected has older males and younger females, i.e., there exists a strong correlation between Age and $Gender$. Formally, the causal relationships between

[1]Department of Computer Science, Columbia University, New York, USA. Correspondence to: Yushu Pan <yushupan@cs.columbia.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

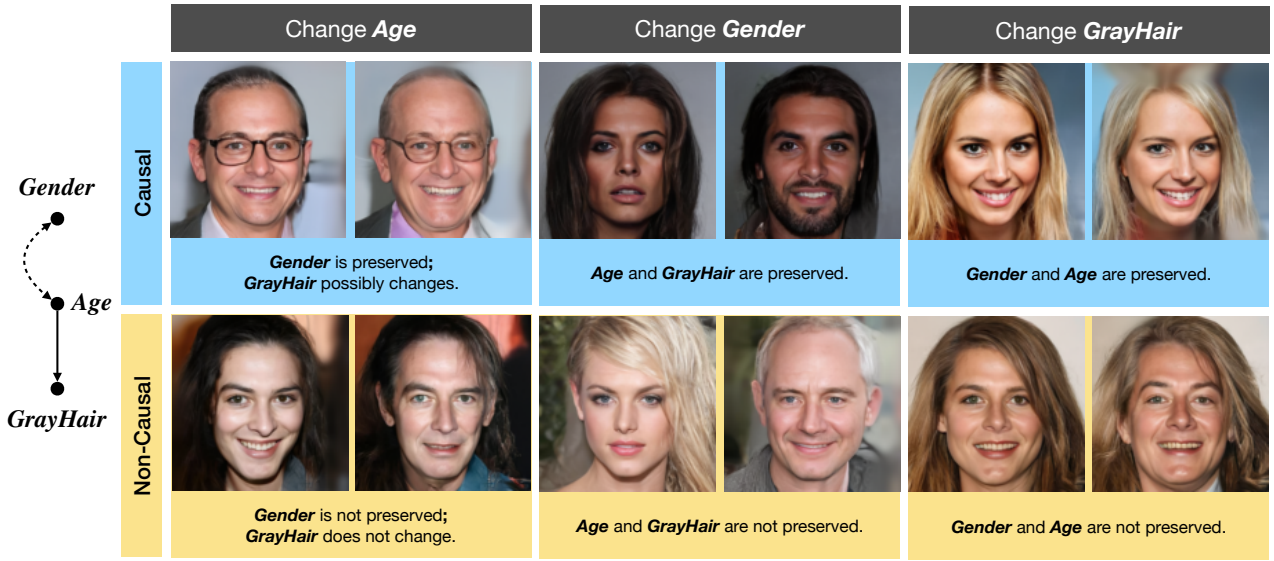

Figure 1: (Left) A causal graph depicting the causal relationships among features. (Right) Image editing results are displayed, with the first row showing edits incorporating causal relations, and the second row without them. Each column represents a unique counterfactual query, altering the age, gender, and gray hair of the individuals.

the three generative factors are shown in Fig. 1.

Existing methods focus on editing a single concept while the effects of the intervened concepts on others are not considered. Suppose we are evaluating the counterfactual query: "Given a certain image, what would the face look like had the person been older?". If the age of the person is changed naively, gender and hair color may also change due to the correlation between these features found in the data. For example, when making an image of a woman older, the AI may inadvertently change her gender to male; see the yellow row in Figure 1. However, it would be expected that changes in age should not affect gender when performing causal editing, as shown in the figure's first row (in blue).

More importantly, existing methods are unable to answer to **what extent hair color should change after an intervention on age.** Even though some recent methods may be able to enforce consistency in terms of gender, the causal effect from the age to the hair color may not be reflected in the counterfactual images. For instance, gray hair may never appear after editing by non-causal approaches. In contrast, causal image editing ensures the effects of target interventions on other features are carried over properly from the factual to the proper counterfactual world. To illustrate, edits in Fig. 1 (blue) are more closely aligned with the reality in which these causal invariances are presented. ■

To capture the causal relationships among generative factors, we build on a class of generative models known as Structural Causal Models (SCMs) (Pearl, 2000). A fully instantiated SCM induces what is known as the Pearl Causal Hierarchy (PCH; also called *ladder of causation*) (Pearl & Mackenzie, 2018; Bareinboim et al., 2022). The PCH consists of families of distributions in increasing levels of refinement: layer 1 ($\mathcal{L}_1$) corresponds to passive observations and typical correlations, layer 2 ($\mathcal{L}_1$) to interventions (e.g., changing a variable to see the effect), and layer 3 ($\mathcal{L}_3$) to counterfactuals (e.g., considering what would happen under hypothetical scenarios). A result known as the causal hierarchy theorem states that higher-layer distributions cannot be answered only from the lower-layer ones (Bareinboim et al., 2022).

Recently, researchers have connected SCMs with deep generative models by implicitly finding surrogate models of the true generative model relating images and its generative factors. Despite the progress made so far, many of these works have limitations in different dimensions that are important to our context. First, they assume *Markovianity*, which implies the absence of unobserved confounding among generative factors. While this assumption may hold in specific settings, it is certainly strong and does not hold in many others, such as $Gender$ and Age described in Example 1.1 (Kocaoglu et al., 2018; Pawlowski et al., 2020; De Sousa Ribeiro et al., 2023; Sanchez & Tsaftaris, 2022; Sauer & Geiger, 2021; De Sousa Ribeiro et al., 2023; Dash et al., 2022).

Second, many of these works estimate counterfactual queries for images and generate samples without considering whether the target query is *identifiable*. In particular, samples are generated even though the query is non-identifiable, which implies that no guarantee can be provided in terms of the quality and causal consistency of the image. In particular, it is unclear whether the causal invari-

ances present in the real systems are preserved across the original and generated images.

Third, other works focus on parametric SCMs over generative factors, such as linear mechanisms, while we study a more general class of non-parametric models (Yang et al., 2021; Shen et al., 2022). Recently, a new class of generative models has been developed, the Neural Causal Model (NCM), which encodes causal constraints into deep generative models (Xia et al., 2021; 2022). These models are capable of both identifying and then estimating counterfactual quantities in non-parametric settings. Despite the soundness of this approach to handling general, non-parametric variables in theory, it remains challenging to estimate counterfactual images, as the structure between generative factors and images is not taken into account. In practice, it's hard to scale these models to higher dimensions. Further discussion on related works is provided in Appendix C.

In this paper, we study the principles underpinning counterfactual image editing tasks and develop a causally-grounded, practical framework for these critical generative capabilities in high-dimensional settings. To achieve this, we formalize counterfactual image tasks according to augmented SCMs (ASCMs), a special class of SCMs taking the image generation step into account. This formulation allows for the encoding of causal relationships between generative factors and the low-level representation, an image in this case. It also enables the modeling of image editing tasks as querying counterfactual distributions induced by true yet unknown ASCMs. More specifically, our contributions are as follows:

1. We formally show that image counterfactual distributions are almost never identifiable from only observational i.i.d image samples. Further, even when the causal relationships between generative factors and images are given, the target counterfactual distribution is still non-identifiable (Sec. 3).

2. We relax these settings and develop a new family of **counterfactual (ctf-) consistent estimators** to approximate non-identifiable distributions. This provides the first procedure with formal guarantees of causal consistency w.r.t. the true generative model. With a sufficient condition to obtain ctf-consistent estimators, we then develop an efficient algorithm (ANCMs) to sample counterfactual images in practice (Sec. 4). Extensive experiments are conducted to demonstrate the effectiveness of ANCMs (Sec. 5).

All supplementary material (including proofs) is provided in the full technical report (Pan & Bareinboim, 2023).

Preliminary. Here we provide the necessary background for this work. An uppercase letter X indicates a random variable and a lowercase letter x indicates its corresponding value; bold uppercase $\mathbf{X}$ denotes a set of random variables, and $\mathbf{x}$ is its corresponding values. The domain of X is denoted as $\mathcal{X}_X$. $P(\mathbf{X})$ is a probability distribution over $\mathbf{X}$.

Our work relies on the Structural Causal Models (SCMs) (Pearl, 2000, Ch. 7); we follow the presentation in (Bareinboim et al., 2022). An SCM is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of exogenous variables, that are determined by factors outside the model; $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ is the set of endogenous variables that are determined by other variables in the model; $\mathcal{F}$ is the set of functions $\{f_{V_1}, f_{V_2} \ldots, f_{V_d}\}$ mapping $\mathbf{U}_{V_j} \cup \mathbf{Pa}_{V_j}$ to V_j, where $\mathbf{U}_{V_j} \subseteq \mathbf{U}$ and $\mathbf{Pa}_{V_j} \subseteq \mathbf{V} \backslash V_j$;$P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$. Each SCM $\mathcal{M}$ induces a causal diagram $\mathcal{G}$, which is a directed acyclic graph where every V_j is a vertex. There is a directed arrow from V_j to V_k if $V_j \in \mathbf{Pa}_{V_k}$. And there is a bidirected arrow between V_j and V_k if $\mathbf{U}_{V_j}$ and $\mathbf{U}_{V_k}$ are not independent with each other (Bareinboim et al., 2022, Def. 11).

An intervention on a subset of $\mathbf{X} \subseteq \mathbf{V}$, denoted by $do(\mathbf{x})$, is an operation where $\mathbf{X}$ takes value $\mathbf{x}$, regardless how $\mathbf{X}$ are originally defined. For an SCM $\mathcal{M}$, let $\mathcal{M}_{\mathbf{x}}$ be the submodel of $\mathcal{M}$ induced by $do(\mathbf{x})$. For any subset $\mathbf{Y} \subseteq \mathbf{V}$, the potential outcome $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution of $\mathbf{Y}$ after feeding $\mathbf{U} = \mathbf{u}$ into the submodel $\mathcal{M}_{\mathbf{x}}$. Specifically, the event $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$ represent "$\mathbf{Y}$ would be $\mathbf{y}$ had $\mathbf{X}$ been $\mathbf{x}$". The counterfactual quantities induced by an SCM $\mathcal{M}$ are defined as (Bareinboim et al., 2022, Def. 7):

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) = \int_{\mathcal{X}_{\mathbf{U}}} \mathbf{1}_{\mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}, \ldots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})=\mathbf{z}} dP(\mathbf{u}), \quad (1)$$

where $\mathbf{Y}, \ldots, \mathbf{Z}, \mathbf{X}, \ldots, \mathbf{W} \subseteq \mathbf{V}$. $P(\mathbf{Y}_{\mathbf{x}})$ reduces to an observational distribution $P(\mathbf{Y})$ when $\mathbf{X}$ is an empty set. The counterfactual optimal bounds are closed intervals based on the following optimization problem (Zhang et al., 2022).

Definition 1.2 (Optimal Counterfactual Bounds)**.** For a causal diagram $\mathcal{G}$ and observed distributions $P(\mathbf{V})$, the *optimal bound* $[l, r]$ over a counterfactual probability $P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}})$ is defined as, respectively, the minimum and maximum of the following optimization problem:

$$\max_{\mathcal{M} \in \Omega(\mathcal{G})} / \min_{\mathcal{M} \in \Omega(\mathcal{G})} \; P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) \text{ s.t.} P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V}) \quad (2)$$

where $\Omega(\mathcal{G})$ is the space of all SCMs that agree with the diagram $\mathcal{G}$, i.e., $\Omega(\mathcal{G}) = \{\forall \mathcal{M} | \mathcal{G}_{\mathcal{M}} = \mathcal{G}\}$. ■

We use neural causal models (NCMs) for estimating counterfactual distributions, which are defined as follows (Xia et al., 2021):

Definition 1.3 ($\mathcal{G}$-Constrained Neural Causal Model ($\mathcal{G}$-NCM))**.** Given a causal diagram $\mathcal{G}$, a $\mathcal{G}$-constrained Neural Causal Model (for short, $\mathcal{G}$-NCM) $\widehat{\mathcal{M}}(\boldsymbol{\theta})$ over variables $\mathbf{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$ such that $\widehat{\mathbf{U}} = \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \subseteq \mathbf{V}\}$, where (1) each $\widehat{U}$ is associated with some subset of variables $\mathbf{C} \subseteq \mathbf{V}$, and $\mathcal{X}_{\widehat{U}} = [0, 1]$ for all $\widehat{U} \in \widehat{\mathbf{U}}$; (2) $\widehat{\mathcal{F}} = \{\hat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each $\hat{f}_{V_i}$ is a feed forward neural network parame-

terized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to values of V_i for $\mathbf{U}_{V_i} = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}} \text{ s.t. } V_i \in \mathbf{C}\}$ and $\mathbf{Pa}_{V_i} = Pa_{\mathcal{G}}(V_i)$; (3) $\widehat{P}(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U} \sim \text{Unif}(0,1)$ for each $\widehat{U} \in \widehat{\mathbf{U}}$. ■

2. Augmented SCMs and Image Counterfactual Distributions

In this section, we model the image counterfactual editing problems in causal language. We first define a special type of SCMs to model the generation process of an image variable $\mathbf{I}$, which is called Augmented SCMs.

Definition 2.1 (Augmented Structure Causal Model)**.** An Augmented Structure Causal Model (for short, ASCM) over a generative level SCM $\mathcal{M}_0 = \langle \mathbf{U}_0, \mathbf{V}_0, \mathcal{F}_0, P^0(\mathbf{U}_0)\rangle$ is a tuple $\mathcal{M} = \langle \mathbf{U}, \{\mathbf{V}, \mathbf{I}\}, \mathcal{F}, P(\mathbf{U})\rangle$ such that
(1) exogenous variables $\mathbf{U} = \{\mathbf{U}_0, \mathbf{U}_\mathbf{I}\}$;
(2) $\mathbf{V} = \mathbf{V}_0$ are labeled observed endogenous variables; $\mathbf{I}$ is an m dimensional image variable;
(3) $\mathcal{F} = \{\mathcal{F}_0, f_\mathbf{I}\}$, where $f_\mathbf{I}$ maps from (the respective domains of) $\mathbf{V} \cup \mathbf{U}_\mathbf{I}$ to $\mathbf{I}$, which is an invertible function regarding $\mathbf{V}$. Namely, there exists a function h such that $\mathbf{V} = h(\mathbf{I})$;
(4) $P(\mathbf{U}_0) = P^0(\mathbf{U}_0)$. ■

The ASCM $\mathcal{M}$ is a "larger" SCM describing a two-stage generative process. $\mathbf{U}_\mathbf{I}$ interact with labeled $\mathbf{V}$ to produce other unlabeled features $\tilde{\mathbf{U}}$ through part of $f_\mathbf{I}$ in the first stage. In the second stage, the remaining part of $f_\mathbf{I}$ mixes the observed $\mathbf{V}$ and unobserved generative factors $\tilde{\mathbf{U}}$ to create the image's set of pixels. Throughout this paper, we assume that domains of observed generative factors $\mathbf{V}$ are discrete and finite. An important aspect of $f_\mathbf{I}$ is that it is invertible regarding $\mathbf{V}$ since generative factors $\mathbf{V}$ are present directly in a given image $\mathbf{i}$. The inverse h represents a labeling process that assigns the correct labels of $\mathbf{V}$ to $\mathbf{i}$. Then, for any $\mathbf{W} \subseteq \mathbf{V}$:

$$P(\mathbf{w} \mid \mathbf{i}) = \mathbf{1}\left[\mathbf{w} = h_\mathbf{W}(\mathbf{i})\right] \tag{3}$$

where $h_\mathbf{W}(\cdot)$ is the subfunction of h mapping from $\mathbf{I}$ to $\mathbf{W}$. The next example illustrates the modeling of face images.

Example 2.2. (Example 1.1 continued). Now we consider the augmented generative process, ASCM $\mathcal{M}^*$: $\langle \mathbf{U} = \{U_F, U_Y, U_{H_1}, U_{H_2}, \mathbf{U}_\mathbf{I}\}, \{\{F, H, Y\}, \mathbf{I}\}, \mathcal{F}^*, P^*(\mathbf{U})\rangle$, where the mechanisms

$$\mathcal{F}^* = \begin{cases} F \leftarrow U_F \oplus U_Y \\ Y \leftarrow U_Y \\ H \leftarrow (\neg Y \wedge U_{H_1}) \oplus (Y \wedge U_{H_2}) \\ \mathbf{I} \leftarrow f^{\text{face}}_\mathbf{I}(F, Y, H, \mathbf{U}_\mathbf{I}) \end{cases} \tag{4}$$

and the exogenous variables $U_F, U_Y, U_{H_1}, U_{H_2}$ are independent binary variables, and $P(U_F = 1) = 0.4, P(U_Y = 1) = 0.4, P(U_{H_1} = 1) = 0.4, P(U_{H_2} = 1) = 0.2$. $\mathbf{U}_\mathbf{I}$ can be correlated with $U_F, U_Y, U_{H_1}, U_{H_2}$. The variable F represents gender (male $F = 0$; female $F = 1$), Y represents age (young $Y = 0$; old $Y = 1$), and H represents whether the person has gray hair (gray $H = 1$; non-gray $H = 0$). We can verify the observational distribution and find that $Y = 1$ and $H = 1$ are positively correlated and older people are more likely to have gray hair.

Before the image is taken, $\mathbf{U}_\mathbf{I}$ and $\{F, Y, H\}$ produce other unobserved generative factors $\tilde{\mathbf{U}}$, such as wrinkles and smiling at the generative level. Among them, some factors (such as wrinkles) can be produced by both $\mathbf{V}$ and $\mathbf{U}_\mathbf{I}$, and other factors (such as smiling) can be produced only by $\mathbf{U}_\mathbf{I}$. Then, $f_\mathbf{I}$ maps all generative factors (including unobserved and observed ones) to image pixels $\mathbf{I}$ at the second stage. Looking at the image, $\{F, Y, H\}$ are deterministic and one can label them through function h, the inverse of $f^{\text{face}}_\mathbf{I}$ $\{F, Y, H\}$. ■

Equipped with ASCMs, we now formalize the counterfactual image generation tasks through causal semantics. Suppose the true underlying ASCM is $\mathcal{M}^*$, which is unobserved. The goal is to query a specific type of counterfactual distribution, i.e., $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$, induced by $\mathcal{M}^*$ given the observed distribution $P(\mathbf{V}, \mathbf{I})$, where $\mathbf{X} \subseteq \mathbf{V}$. Factorizing this joint probability distribution, we have $P^{\mathcal{M}^*}(\mathbf{I} = \mathbf{i}, \mathbf{I}_{\mathbf{x}'} = \mathbf{i}') = P^{\mathcal{M}^*}(\mathbf{I} = \mathbf{i}) P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'} = \mathbf{i}' \mid \mathbf{I} = \mathbf{i})$. This $\mathcal{L}_3$-quantity can be explained as follows. The initial image $\mathbf{i}$ is sampled from $P^{\mathcal{M}^*}(\mathbf{I})$ and the goal is to edit $\mathbf{i}$ to a counterfactual version $\mathbf{i}'$ with modified features $\mathbf{X} = \mathbf{x}'$, where $\mathbf{i}'$ is sampled from $P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'} \mid \mathbf{I} = \mathbf{i})$ [1]. For example, the distribution $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{Y=0})$ (induced by the ASCM introduced in Example 2.2) answers the query "generate an image of a person's face and edit the face to make the person look older".

Throughout this paper, we call this type of $\mathcal{L}_3$-distributions as *Image Counterfactual (I-ctf) Distributions*. A particular instantiation of the image variable, such as $P(\mathbf{I} = \mathbf{i}, \mathbf{I}_{\mathbf{x}'} = \mathbf{i}')$, is called on *Image Counterfactual (I-ctf) Query*. The explanation of I-ctf distributions at the generative level is that given all generative factors in the initial images, what would they be had $\mathbf{X}$ taken value $\mathbf{x}'$. For instance, $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{Y=0})$ is asking what would observed factors (gender, hair color) and unobserved factors (wrinkles, smiling, narrow eyes, ...) be had the person been older.

3. Non-identifiability of I-ctf Distributions

In classic counterfactual image editing tasks, a generator $\widehat{\mathcal{M}}$ is trained to match the distribution $P(\mathbf{V}, \mathbf{I})$, and then the pair of an initial image and its counterfactual can be sampled from $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ induced by the generator. How-

[1] $P^{\mathcal{M}^*}(\mathbf{I}_{\mathbf{x}'} \mid \mathbf{I} = \mathbf{i})$ serves for editing real images when the initial image $\mathbf{i}$ is a real one given by an user.

ever, as alluded to earlier, the Causal Hierarchy Theorem (Bareinboim et al., 2022, Thm. 1) states that counterfactual distributions cannot be computed merely from correlations. In particular, we show next the non-identifiability of any I-ctf query from pure observational data:

Corollary 3.1 (Image Causal Hierarchy Theorem)**.** *Any I-ctf distribution is almost never uniquely computable from the observational distribution.* ■

In other words, Corol. 3.1 states that $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ induced by the proxy generator may not be consistent with the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ even when $\widehat{\mathcal{M}}$ fits the observed distributions perfectly (i.e., $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$) as illustrated in Fig. 2. This inconsistency implies the effect of intervention $\mathbf{X} = \mathbf{x}'$ on other generative factors (features) may differ from the true model and the proxy generator (see Fig. 17 and Example D.1 in Appendix. D).

One of the realizations from the broader causal inference literature is that further assumptions are needed in order to perform counterfactual reasoning. Then we will leverage the causal diagram of the true underlying ASCM to discuss whether an I-ctf distribution is uniquely computable.

A causal diagram encodes constraints over counterfactual distributions compatible with the true and unobserved ASCM, narrowing down the hypothesis space of the proxy generator (Bareinboim et al., 2022, Sec. 1.4). It can be obtained from prior information about concepts in images. For instance, the qualitative understanding that getting older likely leads to gray hair suggests that there should be a direct edge from Y to H in Example 2.2. The causal diagram can be regarded as a causal inductive bias based on human knowledge. The complete causal diagram induced by $\mathcal{M}^*$ is shown in Figure 3; the diagram induced by $\mathcal{M}_0^*$, at the generative level, is in the dashed box.

Once qualitative knowledge about the generative process is encoded in the causal model, our new goal is to infer a target query $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ given a causal diagram $\mathcal{G}$ over $\{\mathbf{V}, \mathbf{I}\}$ and observational distributions $P(\mathbf{V}, \mathbf{L})$. We next define the notation of identifiability in the context of ASCMs.

Definition 3.2 (Identifiability)**.** Consider the true underlying ASCM $\mathcal{M}^*$ defined over $\{\mathbf{V}, \mathbf{I}\}$ and the corresponding causal diagram $\mathcal{G}$ and observational distribution $P(\mathbf{V}, \mathbf{I})$. $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is said to be identifiable from the input $\langle P(\mathbf{V}, \mathbf{I}), \mathcal{G}\rangle$ if $P^{\mathcal{M}^{(1)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}) = P^{\mathcal{M}^{(2)}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ for every pair of ASCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)} \in \Omega_{\mathbf{I}}(\mathcal{G})$ s.t. $P^{\mathcal{M}^{(1)}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^{(2)}}(\mathbf{V}, \mathbf{I})$, where $\Omega_{\mathbf{I}}(\mathcal{G})$ is the space of ASCMs that induces $\mathcal{G}$. The distribution $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is said to be identifiable if $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is identifiable for every $\mathbf{i}, \mathbf{i}' \in \mathcal{X}_{\mathbf{I}}$. ■

Compared to the previous definition of identifiability in literature (e.g., (Pearl, 2009, Ch. 3)), Def. 3.2 restricts the space of SCMs to ASCMs and only considers I-ctfquery

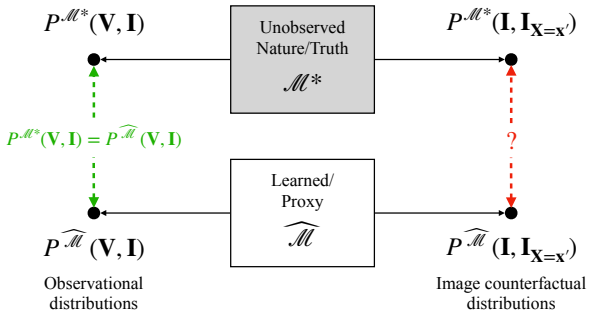

Figure 2: The proxy generator $\widehat{\mathcal{M}}$ is compatible with the same observational distributions with the unobserved true model but is not guaranteed to induce the same target query.

$P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$. The identifiability of $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is equivalent to saying that $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is uniquely computable given the observational distribution and the graphical constraints encoded in $\mathcal{G}$. However, the following proposition implies that even with prior causal information about $\mathbf{V}$ as encoded in $\mathcal{G}$, $P(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is still not identifiable.

Theorem 3.3. *The I-ctf distribution $P(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is not identifiable from any combination of $\langle P(\mathbf{V}, \mathbf{I}), \mathcal{G}\rangle$.* ■

This non-identifiability challenge comes from two perspectives. First, it is unknown how $U_{\mathbf{I}}$ interacts with $\mathbf{V}$ to produce unobserved factors (denoted as $\tilde{\mathbf{U}}$) while these interactions have implications for determining how the counterfactual image should look like. Second, another perspective follows that given the observed values of a generative factor X and its child Y, $P(y'_{x'} \mid y, x)$ is never point identifiable from the observational distribution (see also Fig.17 and Examples D.2 and D.3 in Appendix D).

4. Counterfactually consistent estimation of I-ctf Distributions

So far, we have seen that no I-ctf distribution is identifiable given the causal diagram and the observational distribution. A question naturally arises considering this situation: can these non-identifiable distributions be estimated in any reasonable way? In other words, when the proxy generator ($\widehat{\mathcal{M}}$) does not induce the exact same I-ctf distributions with the true model, what tolerance could be acceptable between $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ and the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$? In addition, we need an estimator to guarantee the approximation of $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ be within the tolerance no matter what observed distribution and causal diagram are given as input. To achieve this, we propose the following two directions to relax the exact estimation of query $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ while retaining causal principles and reasonable results.

(1) **Care set W.** As illustrated in Sec. 2, $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ takes into account how all generative factors ($\{\mathbf{V}, \tilde{\mathbf{U}}\}$) in an im-

age would change after the intervention $do(\mathbf{X} = \mathbf{x})$ takes place. Still, in practical situations, one may only be concerned about how some specific features behave after the intervention but not the whole image. In Example 2.2, all facial features should change causally after making the person older. To illustrate, the intervention on age should preserve the gender and smiling status, and change the hair color with probability 0.4 since $P^{\mathcal{M}^*}(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0) = 0.4$. However, in practice, one may only care about the gender and age after the intervention, but not whether the hair color, smiling status, and background of the image are presented the same way or not. If so, the counterfactual image can have gray hair features and smiling features with arbitrary probability. We introduce the following definition to describe the counterfactual distributions among the selected set of features.

Definition 4.1 (Feature Counterfactual Query)**.** Denote $\mathbf{W}$ as a set of features one cares about and ϕ as a function mapping from $\mathbf{I}$ to $\mathbf{W}$ ($\mathbf{W} = \phi(\mathbf{I})$). The feature counterfactual query (for short, F-ctf) query regarding to $P(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ is defined as:

$$\int_{\mathbf{i}^{(1)}, \mathbf{i}^{(2)} \in \mathcal{X}_{\mathbf{I}}} \mathbf{1}\Big[\phi(\mathbf{i}^{(1)}) = \mathbf{w}, \phi(\mathbf{i}^{(2)}) = \mathbf{w}'\Big]\, dP(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'}) \quad (5)$$

where $\mathbf{w} = \phi(\mathbf{i})$, and $\mathbf{w}' = \phi(\mathbf{i}')$. We denote the F-ctf query as $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$.

In other words, the F-ctf query is a push-forward measure from $P(\mathbf{i}, \mathbf{i}'_{x'})$ through ϕ. The quantity in Eq. 5 integrates over all $P(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'})$ such that $\{\mathbf{i}^{(1)}, \mathbf{i}^{(2)}\}$ has the same cared features $\{\mathbf{w}, \mathbf{w}'\}$ with $\{\mathbf{i}, \mathbf{i}'\}$ in the target query. For concreteness, consider the counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a smiling young male without gray hair and $\mathbf{i}'$ is a smiling old male with gray hair. Suppose the care set $\mathbf{W}$ contains the features: gender and age. The F-ctf query $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$ calculates the probability that the original image describes a young male and the counterfactual image describes an old male after editing. Following Equation (5), $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$ sums over $P(\mathbf{i}^{(1)}, \mathbf{i}^{(2)}_{\mathbf{x}'})$, where $\mathbf{i}^{(1)}$ describes a young male, $\mathbf{i}^{(2)}$ describes an old male. In addition, $\mathbf{i}^{(1)}$ and $\mathbf{i}^{(2)}$ can have arbitrary hair and smiling features since those are not part of $\mathbf{W}$. The F-ctf query induced by a proxy ASCM can be simplified using the following result.

Lemma 4.2. *Consider the true underlying ASCM $\mathcal{M}^*$ over $\{\mathbf{V}, \mathbf{I}\}$, a feature set $\mathbf{W} \subseteq \mathbf{V}$ with mapping function $\phi = h^*_{\mathbf{W}}$, where $h^*_{\mathbf{W}}$ is the inverse function of $f^*_{\mathbf{I}}$ w.r.t. $\mathbf{W}$, and a proxy ASCM $\widehat{\mathcal{M}}$ over $\{\mathbf{V}, \mathbf{I}\}$. if $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$, then*

$$h^*_{\mathbf{W}}(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})) = P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'}), \quad (6)$$

where $\mathbf{w} = h^*_{\mathbf{W}}(\mathbf{i})$, and $\mathbf{w}' = h^*_{\mathbf{W}}(\mathbf{i}')$. ■

This result says that if $\widehat{\mathcal{M}}$ agrees on the observational distribution of $\mathcal{M}^*$ and the care set $\mathbf{W}$ is a subset of observed generative factors $\mathbf{V}$, the F-ctf query $\phi(P(\mathbf{i}, \mathbf{i}'_{x'}))$ is equivalent

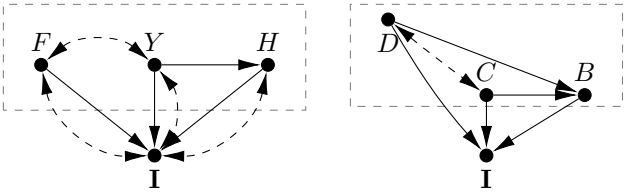

Figure 3: The causal diagram of the in $\mathcal{M}^*$ in Example 2.2 (left) and "Backdoor" setting in Sec. 5.1 (right).

to a counterfactual quantity $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ over $\mathbf{W}$ induced by $\widehat{\mathcal{M}}_0$ at the generative level. We normalize $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ as $P^{\widehat{\mathcal{M}}}(\mathbf{w}'_{\mathbf{x}'} \mid \mathbf{w})$ by dividing $P^{\widehat{\mathcal{M}}}(\mathbf{w})$ and will focus on this *conditional F-ctf query* when the proxy model satisfies $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ as illustrated next.

Example 4.3. Consider the counterfactual image query $P(\mathbf{i}, \mathbf{i}_{Y=0})$, where $\mathbf{i}$ is a young male without gray hair ($F = 0, Y = 1, H = 0$) and $\mathbf{i}'$ describes an old male with gray hair ($F = 0, Y = 0, H = 1$). Suppose the care set $\mathbf{W}$ contains the feature gender (F) and age (Y) as in Example 2.2, i.e., $\mathbf{W} = \{F, Y\}$. Lem. 4.2 suggests the F-ctf query is $P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0, H_{Y=0} = 1, F = 0, Y = 1)$, whenever $\widehat{\mathcal{M}}$ is compatible with $\mathcal{M}^*$ w.r.t. the observational distribution. The normalized conditional F-ctf query is $P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1)$, which illustrates the probability that the gender was still male had a young male gotten older. ■

(2) **Optimal Bounds.** A complementary relaxation arises from the observation that even when a query is not point identifiable, it is still possible to compute informative bounds over the target distribution from a combination of the observational data and the causal diagram (Manski, 1990; Balke & Pearl, 1994; Zhang et al., 2022). These bounds serve as a natural measure of distance, or tolerance, between what is empirically obtainable from the data and the true, yet unobserved, counterfactual distribution. This occurs because numerous ASCMs, compatible with the observed data, can generate counterfactual distributions encompassing the bound. Any value within the optimal bound $[l, r]$ (Def. 1.2) falls within the range of some possible ground truth, contingent on the given assumptions. As assumptions are strengthened, the bounds naturally narrow.

Based on the above discussion, we formally define a class of counterfactual consistent estimators of the target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$.

Definition 4.4 (Ctf-Consistent Estimator w.r.t. Feature Set $\mathbf{W}$)**.** Consider a feature set $\mathbf{W} \subseteq \mathbf{V}$ and its mapping function $\phi = h^*_{\mathbf{W}}$, where $h^*_{\mathbf{W}}$ is the inverse function of $f^*_{\mathbf{I}}$ regarding $\mathbf{W}$. $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ is said to be a *ctf-consistent estimator* of $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ w.r.t. $\mathbf{W}$ if

(1) the observational distributions induced by $\widehat{\mathcal{M}}$ and $\mathcal{M}^*$ are the same, namely, $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ and

(2) the F-ctf query $\phi(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}))$ is within the optimal bound of $P(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ derived by $P(\mathbf{V})$ and $\mathcal{G}$, where $\mathbf{w} = h^*_{\mathbf{W}}(\mathbf{i})$ and $\mathbf{w}' = h^*_{\mathbf{W}}(\mathbf{i}')$;
The proxy quantity $P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ is said to be a ctf-consistent estimator of the true $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ w.r.t. $\mathbf{W}$ if $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}_{\mathbf{x}'})$ is ctf-consistent for every $\mathbf{i}, \mathbf{i}' \in \mathcal{X}_{\mathbf{I}}$. ■

Notice that the F-ctf query $\phi(P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'}))$ is equivalent to $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ according to Lem. 4.2. Def. 4.4 states that if (1) the observational distribution induced by the proxy model agrees with the true model, and (2) the F-ctf query induced by the proxy model is within the optimal bound of $P(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$, then the $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$ can be regarded as a ctf-consistent estimation of the true target query $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}'_{\mathbf{x}'})$. Def. 4.4 does not require that the proxy model $\widehat{\mathcal{M}}$ induces the same $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ but expect $\widehat{\mathcal{M}}$ to be ctf-consistent with $\mathcal{M}^*$ regarding the care set $\mathbf{W}$ while ignoring other observed generative factors $\mathbf{V}\backslash\mathbf{W}$ and $\tilde{\mathbf{U}}$. Specifically, $P^{\widehat{\mathcal{M}}}(\mathbf{w}, \mathbf{w}'_{\mathbf{x}'})$ should be within the optimal bound but no restriction is imposed over the features in $\mathbf{V}\backslash\mathbf{W}$ and $\tilde{\mathbf{U}}$. The next example illustrates this idea.

Example 4.5. (Example 4.3 continued). Def. 4.4 suggests the conditional F-ctf query $Q = P^{\widehat{\mathcal{M}}}(F_{Y=0} = 0 \mid F = 0, Y = 1)$ induced by the proxy model $\widehat{\mathcal{M}}$ should be in the optimal bound $[r, l]$, where

$$r = l = P^{\widehat{\mathcal{M}}}(F = 0 \mid F = 0, Y = 1) = 1 \tag{7}$$

since the intervention $do(Y = 0)$ has no effect on F in the causal diagram (Figure 3).This implies that the gender must remain the same after the editing. In the meantime, it does not matter whether the hair is gray ($\mathbf{V}\backslash\mathbf{W}$) or not and whether the person is smiling ($\tilde{\mathbf{U}}$) since these features are not in the care set.

Now suppose the user cares about gender, age, and hair color, namely, $\mathbf{W} = \{F, Y, H\}$ (instead of $\{F, Y\}$). Based on Def. 4.4 and Lemma 4.2, the corresponding conditional F-ctf query is

$$Q = P(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0), \tag{8}$$

and Q illustrates the probability that the individual is still a male and has gray hair after getting older. This optimal bound analytically can be derived as (see (Pearl, 2009, Thm. 9.2.12)):

$$\begin{aligned} l &= \max\{0, 1 - \frac{P(H = 0 \mid F = 0, Y = 0)}{P(H = 0 \mid F = 0, Y = 1)}\} = 0.25 \\ r &= \min\{1, \frac{P(H = 1 \mid F = 0, Y = 0)}{P(H = 0 \mid F = 0, Y = 1)}\} = 0.5 \end{aligned} \tag{9}$$

Any $P^{\widehat{\mathcal{M}}}(\mathbf{i}, \mathbf{i}_{Y=0})$ such that $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I})$ and $Q^{\widehat{\mathcal{M}}} \in [0.25, 0.5]$ is a ctf-consistent estimator of $P^{\mathcal{M}^*}(\mathbf{i}, \mathbf{i}_{Y=0})$. Even if $Q^{\widehat{\mathcal{M}}}$ is not equal to the true F-ctf query $Q^{\mathcal{M}^*} = 0.4$, the error is acceptable compared to the non-causal method currently used in practice. For example, one may only make the person older (change Y from 0 to 1), but keep other features as close as possible with the initial image. Using such methods, the counterfactual image will never have gray hair, thus the estimation $Q = P(F_{Y=0} = 0, H_{Y=0} = 1 \mid F = 0, Y = 1, H = 0) = 0$. The causal effect of the intervention $Y = 0$ on H is not reflected. ■

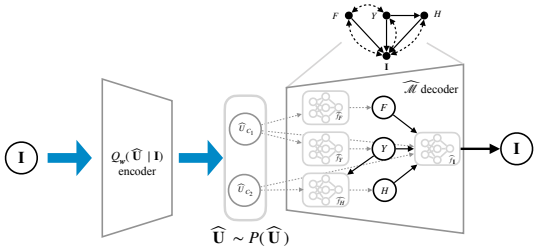

Figure 4: The ANCM structure for Example 2.2.

From now on, our goal is to obtain a ctf-consistent estimator of the non-identifiable target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$ w.r.t. the care set $\mathbf{W}$.

Theorem 4.6. *$P^{\widehat{\mathcal{M}}}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ is a ctf-consistent estimator w.r.t. $\mathbf{W} \subseteq \mathbf{V}$ of $P^{\mathcal{M}^*}(\mathbf{I}, \mathbf{I}'_{\mathbf{x}'})$ if $\widehat{\mathcal{M}} \in \Omega_{\mathbf{I}}(\mathcal{G})$ and $P^{\widehat{\mathcal{M}}}(\mathbf{V}, \mathbf{I}) = P(\mathbf{V}, \mathbf{I})$.* ■

The above result says that any proxy ASCM that is compatible with the diagram $\mathcal{G}$ and $P(\mathbf{V}, \mathbf{I})$ guarantees the estimation of the target distribution being ctf-consistent with the true one. Specifically, in order to construct ctf-consistent estimators, apart from fitting the generator $\widehat{\mathcal{M}}$ with the given observation $P(\mathbf{V}, \mathbf{I})$, it is sufficient to enforce the graphical constraints into $\widehat{\mathcal{M}}$.

4.1. Estimating and Sampling with NCMs

We learned in the previous section, in theory, one could generate ctf-consistent samples by fitting observational distributions to an SCM $\widehat{\mathcal{M}}$ that is compatible with the given diagram. In this section, we develop a practical method for training $\mathcal{G}$-Constrained models, $\mathcal{G}$-NCMs (Def. 1.3, Xia et al. (2021)), with two primary objectives: (a) to fit the observational distribution $P(\mathbf{V}, \mathbf{I})$; (b) to sample images ($\mathbf{i}$) and their counterfactual counterparts ($\mathbf{i}'$).

Towards these goals, we first train $\widehat{\mathcal{M}}$ to match an empirical distribution $\widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = \{\mathbf{v}_k, \mathbf{i}_k\}_{k=1}^n$ derived from finite datasets. Given the substantial difference in the dimensions of variables $\mathbf{V}$ (feature labels) and $\mathbf{I}$ (images), we will fit $P(\mathbf{I})$ and $P(\mathbf{V} \mid \mathbf{I})$ separately. Initially, $P(\mathbf{I})$ will be learned by minimizing the data negative log-likelihood through VAEs (Kingma & Welling, 2013). In this context,

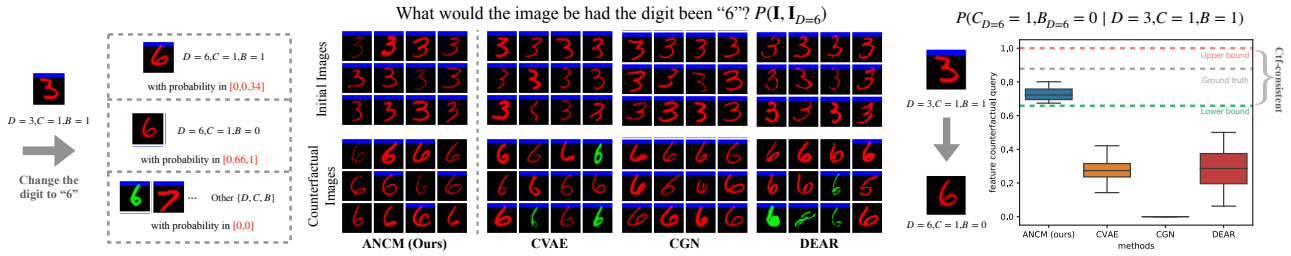

Figure 5: (a) The optimal bound of F-ctf queries when editing a red "3" with a bar to "6". (b) The counterfactual image generation results. (c) The selected F-ctf query estimation.

the proxy $\mathcal{G}$-NCM $\widehat{\mathcal{M}}$ serves as the decoder to approximate $P(\mathbf{I} \mid \widehat{\mathbf{U}})$ with the prior $P(\widehat{\mathbf{U}})$ (details about $\mathcal{G}$-NCM $\widehat{\mathcal{M}}$ can be found in Appendix B.1.3). Furthermore, a separate deep neural network $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$ is utilized to approximate the posterior $P(\widehat{\mathbf{U}} \mid \mathbf{I})$, acting as the encoder, with $\boldsymbol{\omega}$ denoting the network's parameters. The network structure for Example 2.2 is illustrated in Figure 4. To match $P(\mathbf{V} \mid \mathbf{I})$, we minimize the cross-entropy loss L_2 of the true labels of an image sample and its predictions, which can be inferred through $Q_{\boldsymbol{\omega}}(\widehat{\mathbf{U}} \mid \mathbf{I})$ and $\mathcal{M}$ like Locatello et al. (2020b); Shen et al. (2022). We refer to this approach as *ANCM*. More details about the architecture and hyperparameters can be found in Appendix B.1.3.

After training the ANCM, we first sample $\hat{\mathbf{u}}$ from $P(\widehat{\mathbf{U}})$ to generate samples of the target $P(\mathbf{I}, \mathbf{I}_{\mathbf{x}'})$. The initial image sample $\hat{\mathbf{i}}$ could be derived from $\mathbf{I}^{\widehat{\mathcal{M}}}(\hat{\mathbf{u}})$, where $\mathbf{I}^{\widehat{\mathcal{M}}}$ is the network mapping from $\hat{\mathbf{u}}$ to $\mathbf{i}$ in the decoder $\widehat{\mathcal{M}}$. To edit the concept $\mathbf{X} = \mathbf{x}'$, the counterfactual image sample $\hat{\mathbf{i}}_{\mathbf{x}'}$ could be derived through $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}(\hat{\mathbf{u}})$, where $\mathbf{I}^{\widehat{\mathcal{M}}_{\mathbf{x}'}}$ is the network but evaluated through submodel $\widehat{\mathcal{M}}_{\mathbf{x}'}$ of the trained NCM.

5. Experiments

In this section, we conduct an empirical evaluation of the method newly proposed, first with a modified Colored MNIST dataset (Section 5.1) and then with the CelebA-HQ dataset (Karras et al., 2018) (which describes peoples' faces) (Section 5.2). More experiments and further details of the model architectures are provided in Appendix B.

5.1. Colored MNIST with Bars

We first conduct experiments on the modified handwritten MNIST dataset (Deng, 2012), featuring colored digits and a horizontal blue bar in images. [2] The observed generative factors include $\{D, C, B\}$, where D denotes the digits from 0 to 9; C indicates the digit color (green for $C = 0$; red for $C = 1$); B determines whether the top of the image features a blue bar ($B = 1$) or not ($B = 0$). We explore 4 tasks: editing digits in "Backdoor" (shown in this section), editing bars in "Backdoor", editing digits in "Frontdoor", and editing color in "Frontdoor"(shown in Appendix B.1).

In the Backdoor setting, the digit (B) and the color (C) are confounded with a positive correlation, but they do not directly affect each other. There are more red/larger (≥ 5) digits and green/smaller (< 5) digits in the dataset. The digit (D) has a negative effect on the existence of the bar (B). Larger digits are less likely to have a bar on the top and The color (C) also has a negative effect on the existence of the bar (B). red digits are less likely to have a bar on top. Fig. 3(right) shows the causal diagram $\mathcal{G}$ induced by the true ASCM in the backdoor setting.

The first task we consider is to edit the digit D counterfactually. We let the cared features be the digit, color, and whether the image has a blue bar, namely, $\mathbf{W} = \{B, C, D\}$. This implies we do not care how other generative factors (i.e., position, thickness) change in the counterfactual world. For counterfactual editing, changing D should not affect C while it possibly changes B, since D is confounded with C but directly affects B. For instance, suppose we are editing a red "3" with a bar (an image with $\{D = 3, C = 1, B = 1\}$) and wonder what would happen had the digit "3" been a "6". In this case, the optimal bounds of conditional F-ctf distribution are shown in Fig. 5a. For example, the probability that the counterfactual image has features $\{D = 6, C = 1, B = 0\}$ is $P(C_{D=6} = 1, B_{D=6} = 0 \mid D = 6, C = 1, B = 1)$, which should be within $[0.66, 1]$. To achieve ctf-consistency, we expect the proxy model to follow these theoretical bounds.

We compare three baseline methods with ANCM. The first one is a naive conditional VAE that learns the correlation between the digit and the image variable $P^{\mathcal{M}_{\mathbf{B}}}(\mathbf{I} \mid D)$ (Sohn et al., 2015). The second one is CGN (Sauer & Geiger, 2021), which approximates SCMs over variables $Shape, Texture$, and $Background$. The third method DEAR (Shen et al., 2022) is designed for Markovian settings among generative factors.

[2] A bar in an image refers to complete rows of blue pixels.

The counterfactual image editing results are shown in Figure 5b. After changing the digit, ANCM preserves the original colors in counterfactual images and is likely to remove the bar, reflecting the bound value discussed above. CVAE is likely to change the color C as it uses the spurious correlation between D and C. Also, the CVAE method fails to capture the causal effect from D to B since the bar hardly disappears after the intervention $do(D = 6)$. CGN preserves the original colors but the bars are never removed, which implies the causal effect from D to B is not reflected. DEAR fails to preserve the color because it is restricted to be used for Markovian models. We re-run each method 4 times and calculate the empirical probability $P(C_{D=6} = 1, B_{D=6} = 0 \mid D = 6, C = 1, B = 1)$. The result is shown in Figure 5c. We can see that queries generated by all baseline methods are not within the optimal bound. In contrast, ANCMs provide in-bound estimation. Both the visualization, numerical results, and theoretical results state the ANCMs capture the causal effects among $\{D, C, B\}$ and offer ctf-consistent estimators while baselines do not. Further tasks are provided in Appendix B.1.

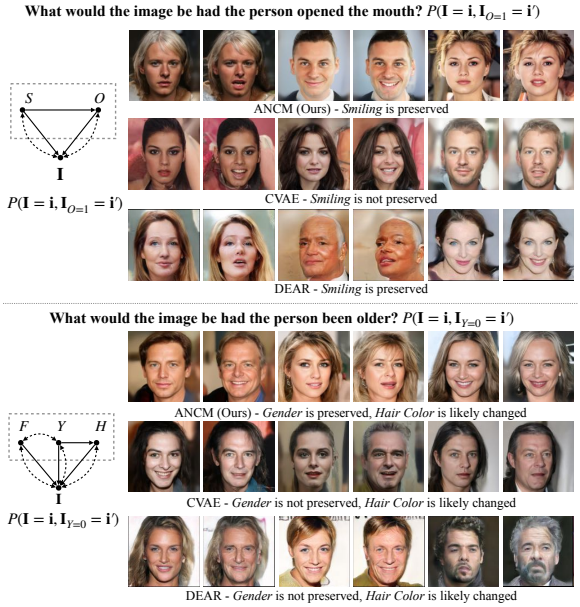

Figure 6: Editing results of the CelebaHQ Experiment.

5.2. Celeba-HQ

In CelebA-HQ experiment, we consider two causal diagrams as shown in Fig. 6. In the first experiment, we consider generative factors $Smile$ (S) and $Open\ Mouth$ (O), and in the second experiment, we consider $Female$ (F), $Young$ (Y) and $Grayhair$ (H). The first target counterfactual queries are "What would the image be had the person opened the mouth?", and the second is "What would the image be had the person been older?". The feature sets are $\mathbf{W} = \{S, O\}$ and $\mathbf{W} = \{F, Y, H\}$ in these two settings, respectively. We also compare ANCM (ours) against the CVAE and DEAR baselines. CGN is not compared here since the variables of CGN are restricted to $Shape, Texture$, and $Background$. Meanwhile, DiffuseVAE (Pandey et al., 2022) is leveraged for ANCM and CVAE here to refine samples to high quality since VAEs often produce blurry images that lack high-frequency information (Dosovitskiy & Brox, 2016).

The empirical results are shown in Fig. 6. In the first setting, the feature set $\mathbf{W} = \{S, O\}$ implies the counterfactual query is $P(S, O, S_{O=1})$, namely, "Would the person smile (or not) had the person opened the mouth?". The constraints induced by the ground truth model imply that changing the mouth should not affect smiling since O is the direct child of S and not the other way around. As shown Figure 6, the smiling features are preserved after the editing by ANCM and DEAR. However, CVAE only captures the correlation between these factors, thus the non-smiling person changes to smiling after editing of mouth. On the other hand, ANCM produces higher-quality images compared to DEAR.

The second causal diagram indicates the correlations between gender and age in Example 1.1. The dataset has more face images of young females and old males. More specifically, 71% of the young people are female, and 66% of the old people are male. The features set $\mathbf{W} = \{F, Y, H\}$ implies the counterfactual distribution is "What would the gender of the person and the hair color of the person be had the person been older?". The causal constraints suggest that the gender of the person should be preserved and the likelihood of gray hair should increase. Our methods match these causal relationships while baseline methods may change the original gender as shown in Figure 6, which is of course undesirable. Further details are provided in Appendix B.2.

6. Conclusions

We study the problem of counterfactual image generation and editing through formal causal language. We first showed that image counterfactual distributions are not identifiable from a combination of observational data and prior causal knowledge about the generating model represented as a causal diagram. Given such impossibility results, we proposed a new family of counterfactual (ctf-) estimator estimators that come accompanied with guarantees that the generated counterfactual images remain causally consistent with the true image counterfactual distribution for any causal relationship between generative factors, which is important towards building more trustworthy AI. We developed an efficient algorithm to train neural causal models and sample counterfactual images. Finally, we demonstrate our methods are able to generate high-quality counterfactual images.

Acknowledgements

This research was supported in part by the NSF, ONR, AFOSR, DARPA, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation. We thank Kevin Xia for the feedback provided in the early versions of this manuscript.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many societal implications of our work and we hope to be beneficial, as elaborated next.

Reflecting on the broader literature, we propose the first method capable of providing formal guarantees over counterfactual image generation and editing. The main advancement of our work lies in its emphasis on preserving the causal relationships among features, enabling sound, robust, and more realistic counterfactual generation. This approach differs significantly from the existing literature, which primarily focuses on reflecting the intervened features in the image. The critical distinction centers on what happens with the other features that were not intervened upon, and determining which features are shared or not between factual and counterfactual worlds. Although almost never formally articulated, there are two prevalent approaches to this problem in the prior literature. Some works remain silent regarding the counterfactual status of the non-intervened features. This means that the neural network might leverage the correlation between features found in the factual world, leading to the various spurious results discussed earlier. For instance, instructing a generative AI to change a specific feature of an individual might result in a completely different person with other features, such as a different gender or race, despite they are not being causally related. This occurs because the neural model tends to leverage the correlation between factors found in the observational data, which is oblivious to their causal relationship. Other works attempt to ensure that the non-intervened features are preserved across factual and counterfactual worlds. However, this approach is also inadequate in settings where some of the features exert causal influence on others, and the generative AI should accordingly ascertain these relations. For instance, making a person older should logically lead to changes in hair color (or its amount) in both factual and counterfactual images.

After all, we believe the results stemming from this work have broad implications for the development of the next generation of generative AI. First, we note that the training datasets used for large generative models are almost never balanced (see, for example, (Buolamwini & Gebru, 2018)), which implies spurious correlations across features and the generated images. In practice, this often leads to more frequent, unexpected inaccuracies and biases in these models (e.g., refer to (Plecko & Bareinboim, 2022).) Understanding and accounting for the causal relationships among generative factors is fundamental for the accuracy and fairness of these models. Second, the lack of proper treatment of the causal invariances required for sound counterfactual reasoning translates into the impossibility of providing any sort of guarantees over what these models generate as output and their plausibility, a certainly undesirable state of affairs.

References

Augustin, M., Boreiko, V., Croce, F., and Hein, M. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Avin, C., Shpitser, I., and Pearl, J. Identifiability of Path-Specific Effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–363. Morgan-Kaufmann Publishers, 2005.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Balke, A. and Pearl, J. Counterfactual Probabilities: Computational Methods, Bounds, and Applications. In de Mantaras, R. L. and D.˜Poole (eds.), *Uncertainty in Artificial Intelligence 10*, pp. 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pp. 1342–1350, 2015.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 2022.

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

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

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*,

2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chai, L., Wulff, J., and Isola, P. Using latent space regression to analyze and leverage compositionality in {gan}s. In *International Conference on Learning Representations*, 2021.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Child, R. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34: 6856–6867, 2021a.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6856–6867. Curran Associates, Inc., 2021b.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., and Raff, E. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pp. 88–105. Springer, 2022.

Dash, S., Balasubramanian, V. N., and Sharma, A. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924, 2022.

De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., and Glocker, B. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 7390–7425, 2023.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

Dosovitskiy, A. and Brox, T. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

Falcon, W. and Cho, K. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*, 2020.

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 5744–5753, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goyal, Y., Feder, A., Shalit, U., and Kim, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019a.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019b.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850, 2020.

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

Heckman, J. J. Randomization and Social Policy Evaluation. In Manski, C. and Garfinkle, I. (eds.), *Evaluations: Welfare and Training Programs*, pp. 201–230. Harvard University Press, Cambridge, MA, 1992.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 978–987. AUAI Press, Aug 2018.

Jaber, A., Zhang, J., and Bareinboim, E. Identification of conditional causal effects under Markov equivalence. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11512–11520. Curran Associates, Inc., 2019.

Jaber, A., Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Causal discovery from soft interventions with unknown targets: Characterization and learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9551–9561. Curran Associates, Inc., 2020.

Jaber, A., Ribeiro, A., Zhang, J., and Bareinboim, E. Causal identification under markov equivalence: Calculus, algorithm, and completeness. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3679–3690. Curran Associates, Inc., 2022.

Jahanian*, A., Chai*, L., and Isola, P. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HylsTT4FvB.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

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

Khorram, S. and Fuxin, L. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10203–10212, 2022.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

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

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

Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems 30*, pp. 7018–7028. Curran Associates, Inc., 2017a.

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

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Kocaoglu, M., Jaber, A., Shanmugam, K., and Bareinboim, E. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems 32*, pp. 14346–14356, Vancouver, Canada, 2019. Curran Associates, Inc.

Kwon, G. and Ye, J. C. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18062–18071, 2022.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *First Conference on Causal Learning and Reasoning*, 2021.

Lee, S., Correa, J., and Bareinboim, E. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Li, A., Jaber, A., and Bareinboim, E. Causal discovery from observational and interventional data across multiple environments. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *ArXiv*, abs/1811.12359, 2019a.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019b.

Locatello, F., Poole, B., Raetsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/locatello20a.html.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=SygagpEKwB.

Manski, C. F. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80: 319–323, 1990.

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7521–7531, 2022.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

Moraffah, R., Karami, M., Guo, R., Raglin, A., and Liu, H. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

Nasr-Esfahany, A., Alizadeh, M., and Shah, D. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning*, pp. 25733–25754. PMLR, 2023.

Pan, Y. and Bareinboim, E. Counterfactual image editing. Technical Report R-103, Causal Artificial Intelligence Lab, Columbia University, December 2023.

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Pawlowski, N., Coelho de Castro, D., and Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

Pearl, J. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

Pearl, J. and Mackenzie, D. *The Book of Why*. Basic Books, New York, 2018.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Plecko, D. and Bareinboim, E. Causal fairness analysis. Technical Report R-90, Causal Artificial Intelligence Lab, Columbia University, Jul 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. Explaingan: Model explanation via decision boundary crossing transformations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–681, 2018.

Sanchez, P. and Tsaftaris, S. A. Diffusion causal models for counterfactual estimation. In Schölkopf, B., Uhler, C., and Zhang, K. (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 11–13 Apr 2022.

Sauer, A. and Geiger, A. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

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

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55, 2022.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Shpitser, I. and Pearl, J. Effects of Treatment on the Treated: Identification and Generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, 2009. AUAI Press.

Shpitser, I. and Sherman, E. Identification of Personalized Effects Associated With Causal Pathways. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 530–539, 2018.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 650–665. Springer, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., and Shah, C. Counterfactual explanations and algorithmic recourses for machine learning: a review. *arXiv preprint arXiv:2010.10596*, 2020.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wang, P. and Vasconcelos, N. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8981–8990, 2020.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability,

and inference. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10823–10836. Curran Associates, Inc., 2021. URL https://causalai.net/r80.pdf.

Xia, K., Pan, Y., and Bareinboim, E. Neural causal models for counterfactual identification and estimation. In *International Conference on Learning Representations*, 2022.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9593–9602, 2021.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

Zhang, J. and Bareinboim, E. Fairness in decision-making–the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 2037–2045, 2018.

Zhang, J., Jin, T., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML-22)*, 2022.

Zhang, J., Squires, C., Greenewald, K., Srivastava, A., Shanmugam, K., and Uhler, C. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.