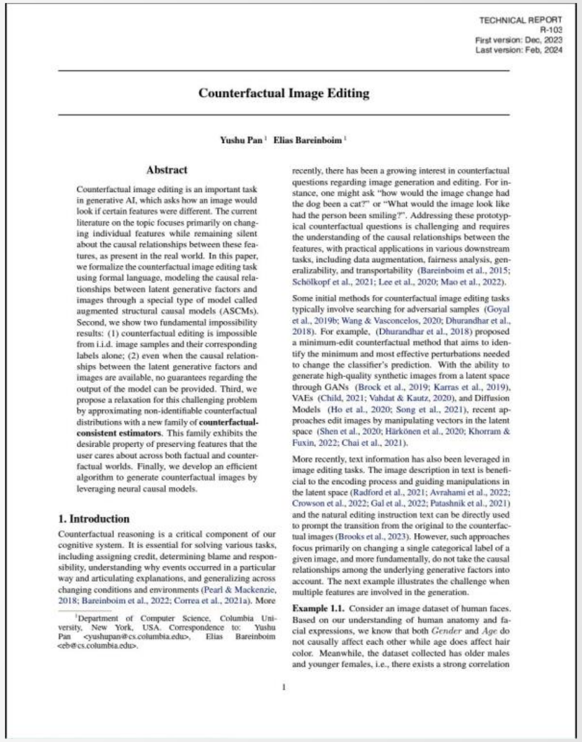
Counterfactual Image Editing

Yushu Pan[1], Elias Bareinboim[1]

Presented by Erik Urzua

[1]Department of Computer Science, Columbia University, New York, USA



TECHNICAL REPORT
R-103
First version: Dec, 2023
Last version: Feb, 2024

Counterfactual Image Editing

Yushu Pan[1] Elias Bareinboim[1]

Abstract

Counterfactual image editing is an important task in generative AI, which asks how an image would look if certain features were different. The current literature on the topic focuses primarily on changing individual features while remaining silent about the causal relationships between these features, as present in the real world. In this paper, we formalize the counterfactual image editing task using formal language, modeling the causal relationships between latent generative factors and images through a special type of 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) even when the causal relationships between the latent generative factors and images are available, no guarantees regarding the output of the model can be provided. Third, we propose a relaxation for this challenging problem by approximating non-identifiable counterfactual distributions with a new family of **counterfactual-consistent estimators**. This family exhibits the desirable property of preserving features that the user cares about across both factual and counterfactual worlds. Finally, we develop an efficient algorithm to generate counterfactual images by 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) and 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. The next example illustrates the challenge when multiple features are involved in the generation.

Example 1.1. Consider an image 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

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

Counterfactual Image Editing?

- Generative AI task asking how an image would have looked had different features been different
- "How would the image change had the dog been a cat?" or " What would the image look like had the person been smiling"

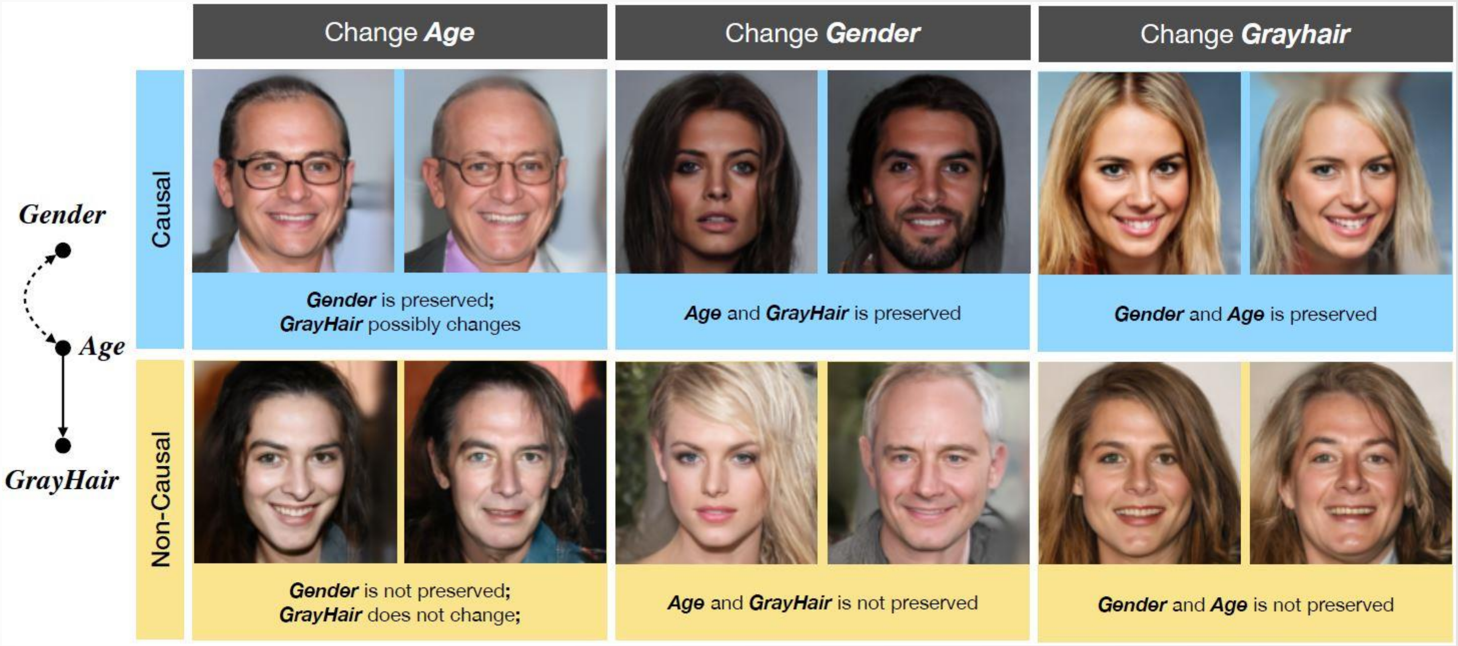


Figure1: (Left) A causal graph depicting the causal relationship among features. Right)Image editing results are displayed, with the first row showing edits incorporating causal relations, and the second row without them

Paper Goals and Results

- Authors formalize the counterfactual image editing task
  - Causal relationship between latent generative factors and images through a special type of SCM
- Show two fundamentally important results regarding the possibility of counterfactual image editing for
- Give a relaxation for identified impossible problems by approximating non-identifiable counterfactual distributions with a new family of counterfactual-consistent estimators
  - Exhibit property of preserving features that are cared about across factual and counterfactual worlds
- Develop an efficient algorithm to generate counterfactual images through neural causal models.

Overview/Outline

- Counterfactual Image Editing Background
- Definitions: Old and New
- Augmented SCM and Image Counterfactual Distributions
- Non-Identifiability of Image Counterfactual Distributions
- Counterfactually Consistent Estimation of Image Counterfactual Distributions
- Experimental Results

Counterfactual Image Editing Background

- Some initial counterfactual image editing tasks involved searching adversarial samples.
- Recently because of the ability to generate high-quality images from a latent space through GANs[4], VAEs[5] and Diffusion Models[6], some have developed approaches to manipulate vectors in the latent space.
- Most recently some have incorporated text information in the image editing task.
  - Editing text instructions can be used to prompt the transition from the original to the counterfactual since image descriptions in text is beneficial in the encoding process and to guide manipulations in the latent space.

Definitions: Old and New

- The following presentation follows that given in the original paper.
- X := random variable where x indicates its corresponding value
- $\mathbf{X}$:= set of random variables with $\boldsymbol{x}$ corresponding values
- χ_X denotes the domain of X and $\boldsymbol{\chi_X} = \chi_{X_1} \times \cdots \times \chi_{X_d}$ for $\mathbf{X} = \{X_1, \dots, X_d\}$.
- A Structural Causal Model (SCM) is a 4-tuples $< \boldsymbol{U}, \boldsymbol{V}, F, P(\boldsymbol{U}) >$, where (1) $\boldsymbol{U}$ is a set of background variables, also called the exogenous variables, that are determined by factors outside model; (2) $\boldsymbol{V} = \{\mathrm{V}_1, \mathrm{V}_2, \dots, V_d\}$ is the set of endogenous variables that are determined by other variables in the model; (3) F is the set of functions $\{f_{V_1}, \dots, f_{V_d}\}$ mapping $\mathbf{U}_{V_j} \cup \boldsymbol{Pa}_{V_j}$ to V_j, where $\boldsymbol{U}_{V_j} \subseteq \boldsymbol{U}$ and $\boldsymbol{Pa}_{V_j} \subseteq \boldsymbol{V} \setminus V_j$;(4)$\boldsymbol{P}(\boldsymbol{U})$ is a probability function over the domain of $\boldsymbol{U}$.
- For any SCM M, let $M_{\boldsymbol{x}}$ be the submodel of M induced by do($\boldsymbol{x}$). For any subset $\boldsymbol{Y} \subseteq \boldsymbol{V}$, the potential outcome $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$ is defined as the solution of $\boldsymbol{Y}$ after feeding $\boldsymbol{U} = \boldsymbol{u}$ into the submodel $M_{\boldsymbol{x}}$. Then $\boldsymbol{Y}_{\boldsymbol{x}}$ is called the counterfactual variable induce by M.
- The counterfactual quantities induced by the model M are defined as:

$$P^M(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}) = \int_{\chi_{\boldsymbol{U}}} \mathbf{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{u})=\boldsymbol{z}} dP(u),$$

where $\boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$. Specifically, $P(\boldsymbol{Y}_{\boldsymbol{x}})$ reduces to an observational distribution $P(\boldsymbol{Y})$ taking $\mathbf{X}$ as an empty set.

Definitions: Old and New (cont.)

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

$$\max_{M \in \Omega(G)}/\min P^M(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}})$$

  such that $P^M(\boldsymbol{V}) = P(\boldsymbol{V})$ where $\Omega(G)$ is the space of all SCMs that agree with the diagram

- G-Constrained Neural Causal Model (G-NCM) – Given a causal diagram G, a G-constrained Neural Causal Model $\widehat{M}(\boldsymbol{\theta})$ over the variables $\boldsymbol{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \boldsymbol{V}\}$ is a SCM $\langle \widehat{\boldsymbol{U}}, \boldsymbol{V}, \hat{F}, \hat{P}(\widehat{\boldsymbol{U}}) \rangle$ such that $\widehat{\boldsymbol{U}} = \{\widehat{\boldsymbol{U}}_C : \boldsymbol{C} \subseteq \boldsymbol{V}\}$, where (1) each $\widehat{U}$ is associated with some subset of variables $\mathbf{C} \subseteq \boldsymbol{V}$, and $D_{\widehat{U}} = [0,1]$ for all $\widehat{U} \in \widehat{\boldsymbol{U}}$; (2) $\hat{F} = \{\hat{f}_{V_i} \in \boldsymbol{V}\}$ where each $\hat{f}_{V_i}$ is a feed forward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\boldsymbol{U}_{V_i} \cup \boldsymbol{Pa}_{V_i}$ to values of V_i for $\boldsymbol{U}_{V_i} = \{\widehat{U}_{\boldsymbol{C}} : \widehat{U}_{\boldsymbol{C}} \in \widehat{\boldsymbol{U}}\ s.t\ V_i \in \boldsymbol{C}\}$ and $\boldsymbol{Pa}_{V_i} = Pa_G(V_i)$; (4) $\hat{P}(\widehat{\boldsymbol{U}})$ is defined s.t. $\widehat{U} \sim Unif(0,1)$ for each $\widehat{U} \in \widehat{\boldsymbol{U}}$.

Augmented SCM and Image Counterfactual Distributions

- Augmented Structural Causal Model – An Augmented Structural Causal Model (ASCM) over a generative level SCM $M_0 = \langle \boldsymbol{U}_0, \boldsymbol{V}_0, F_0, P^0(\boldsymbol{U}_0) \rangle$ is a tuple $M = \langle \boldsymbol{U}, \{\boldsymbol{V}, \boldsymbol{I}\}, F, P(\boldsymbol{U}) \rangle$ such that (1) exogenous variables $\boldsymbol{U} = \{\boldsymbol{U}_0, \boldsymbol{U}_I\}$; (2) $\boldsymbol{V} = \boldsymbol{V}_0$ are labeled observed endogenous variables and $\boldsymbol{I}$ is an m dimensional image variable; (3) $F = \{F_0, f_I\}$ where f_I maps from (the respective domains of) $\boldsymbol{V} \cup \boldsymbol{U}_I$ to $\boldsymbol{I}$, which is an invertible function regarding $\boldsymbol{V}$. Namely, there exists a function h such that $\boldsymbol{V} = h(\boldsymbol{I})$. (4) $P(\boldsymbol{U}_0) = P^0(\boldsymbol{U}_0)$.
- ASCM are “larger” SCM describing a two-stage generative process: In the first stage, the $\boldsymbol{U}_I$ interact with the $\boldsymbol{V}$ to produce other unlabeled features $\widetilde{\boldsymbol{U}}$ through part of f_I. In the second stage the other parts of f_I mix with the observed $\boldsymbol{V}$ and unobserved generative factor $\widetilde{\boldsymbol{U}}$ to create the images set of pixels.
- The inverse of f_I, h, is important as it represents a labeling process that assigns the correct labels of $\boldsymbol{V}$ to $\boldsymbol{i}$. So for any $\boldsymbol{W} \subseteq \boldsymbol{V}$,

$$P(\boldsymbol{w}|\boldsymbol{i}) = \begin{cases} 1 & \boldsymbol{w} = h_{\boldsymbol{W}}(\boldsymbol{i}) \\ 0 & otherwise \end{cases}$$

  where $h_{\boldsymbol{W}}(\cdot)$ is the subfunction of h mapping $\boldsymbol{I}$ to $\boldsymbol{W}$.

ASCM Example – Earlier Face Modeling

- $M^* = \langle \boldsymbol{U} = \{U_F, U_Y, U_{H_1}, U_{H_2}, \boldsymbol{U_I}\}, \{\{F, Y, H\}, \boldsymbol{I}\}, F^*, P^*(\boldsymbol{U})\rangle$ where

$$F^* = \begin{cases} F \leftarrow U_F \oplus U_Y \\ Y \leftarrow U_Y \\ H \leftarrow (\neg Y \wedge U_{H_1}) \oplus (Y \wedge U_{H_2}) \\ \boldsymbol{I} \leftarrow f_I^{face}(F, Y, H, \boldsymbol{U_I}) \end{cases}$$

  and variables $U_F, U_Y, U_{H_1}, U_{H_2}$, are independent binary variables and $P(U_F = 1) = 0.4$, $P(U_Y = 1) = 0.4$, $P(U_{H_1} = 1) = 0.4$, $P(U_{H_2} = 1) = 0.2$.

- Prior to an image being taken, $\boldsymbol{U_I}$ and $\{F, Y, H\}$ are produce other unobserved generative factors $\tilde{\boldsymbol{U}}$. Amongst them, some factors can be produced by $\boldsymbol{V}$ and $\boldsymbol{U_I}$ while some can only be produced by $\boldsymbol{U_I}$. All generative factors are then mapped by f_I to image pixels $\boldsymbol{I}$ at the second stage.

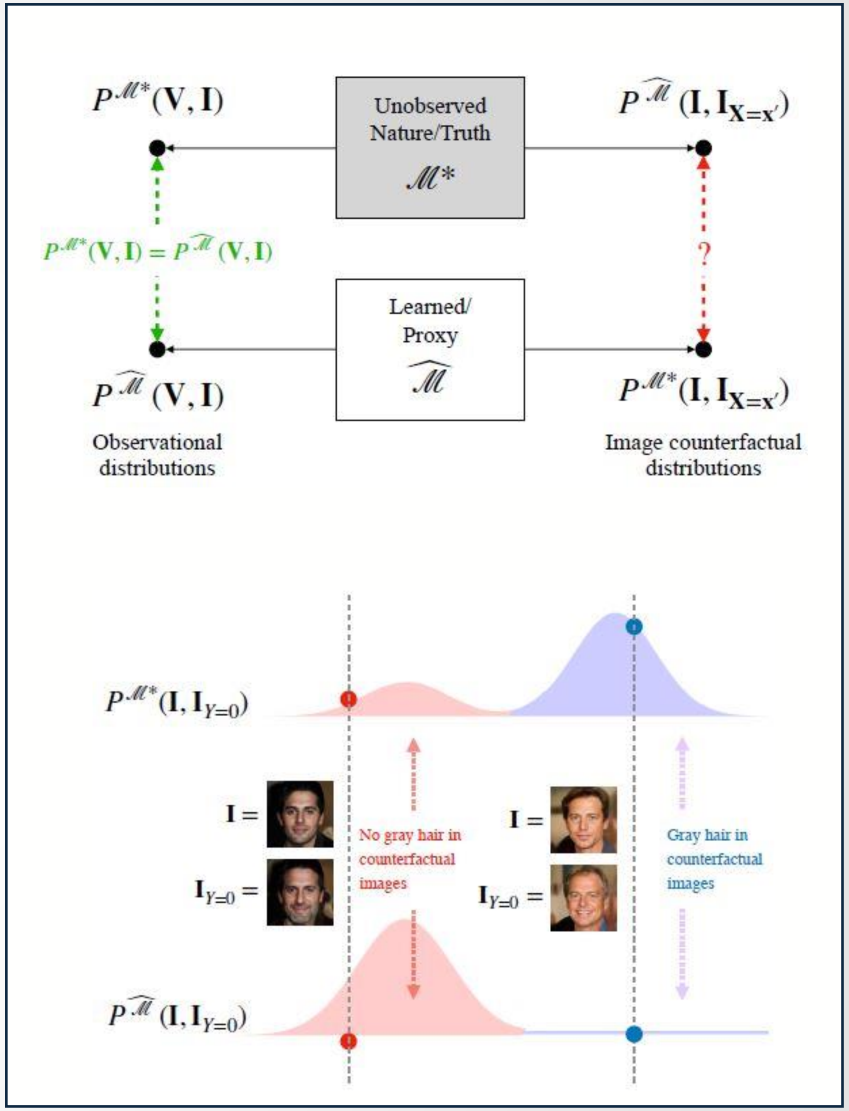
F	Y	H	$P(F, Y, H)$
0	0	0	0.216
0	0	1	0.144
0	1	0	0.128
0	1	1	0.032
1	0	0	0.144
1	0	1	0.096
1	1	0	0.192
1	1	1	0.048

Formalizing Counterfactual Image Generation Tasks.

- Suppose that the true unobserved underlying ASCM is M^*. Goal is to query a specific type of counterfactual distribution induce by M^* given the input distribution $P(\boldsymbol{V}, \boldsymbol{I})$. That is $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{X}'})$ where $\boldsymbol{X} \subseteq \boldsymbol{V}$.
- Factorizing the joint distribution, we get that $P^{M^*}(\boldsymbol{I} = \boldsymbol{i}, \boldsymbol{I}_{\boldsymbol{X}'} = \boldsymbol{i}') = P^{M^*}(\boldsymbol{I} = \boldsymbol{i}) P^{M^*}(\boldsymbol{I}_{\boldsymbol{X}'} = \boldsymbol{i}' | \boldsymbol{I} = \boldsymbol{i})$ which is explained as, the initial image $\boldsymbol{i}$ is sampled from $P^{M^*}(\boldsymbol{I})$ and the goal is to edit $\boldsymbol{i}$ to a counterfactual version $\boldsymbol{i}'$ with modified features $\boldsymbol{X} = \boldsymbol{x}'$ where $\boldsymbol{i}'$ is sampled from $P^{M^*}(\boldsymbol{I}_{\boldsymbol{X}'} | \boldsymbol{I} = \boldsymbol{i})$
- In the case of our example $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{Y=0})$ can answer the query to generate an image for a persons face and edit the face to make them look older.
- These types of distribution are called Image Counterfactual Distributions. A particular instantiation of the image variable is called an Image Counterfactual Query.

Non-Identifiability of Image Counterfactual Distributions

- Classical counterfactual image generation tasks generally train a generator $\widehat{M}$ to match the distribution $P(\boldsymbol{V}, \boldsymbol{I})$ and then the initial image and counterfactual image pair can be sampled from $P^{\widehat{M}}(\boldsymbol{I}, \boldsymbol{I}_{x'})$ from the generator. However counterfactual distribution cannot be computed merely from correlation.
- **Image Causal Hierarchy Theorem** – Any image counterfactual distribution is almost never uniquely computable from the observational distribution (or its samples).
- That is, $P^{\widehat{M}}(\boldsymbol{I}, \boldsymbol{I}_{x'})$ induced by the proxy generator may not be consistent with the true $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{x'})$ even when the proxy generator fits the observational distribution perfectly.
- Broadly, there is nothing in the observational distribution indicating how an image would change under hypothetical interventional world.



Causal Diagrams to the Rescue?

- **Goal**: Infer target image counterfactual query $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ given a causal diagram G over $\{\boldsymbol{V}, \boldsymbol{I}\}$ and observational distributions $P(\boldsymbol{V}, \boldsymbol{L}\ )$ after qualitative knowledge about the generative process has been placed into the causal model.
- _Identifiability_ – *Consider the true underlying ASCM M^* defined over $\{\boldsymbol{V}, \boldsymbol{I}\}$ and the corresponding causal diagram G and observational distribution $P(\boldsymbol{V}, \boldsymbol{I})$. An image counterfactual query $P(\boldsymbol{i}, \boldsymbol{i}'_{\boldsymbol{x}})$ is said to be identifiable from the input $\langle P(\boldsymbol{V}, \boldsymbol{I}), G\rangle$ if $P^{M^{(1)}}(\boldsymbol{i}, \boldsymbol{i}'_{\boldsymbol{x}}) = P^{M^{(2)}}(\boldsymbol{i}, \boldsymbol{i}'_{\boldsymbol{x}})$ for every pair of ASCM $M^1, M^2 \in \Omega_{\boldsymbol{I}}(G)$ such that $P^{M^{(1)}}(\boldsymbol{V}, \boldsymbol{I}) = P^{M^{(2)}}(\boldsymbol{V}, \boldsymbol{I})$, where $\Omega_{\boldsymbol{I}}$ is the space of ASCMs. The distribution $P(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ is said to be identifiable if $P(\boldsymbol{i}\ \boldsymbol{i}'_{\boldsymbol{x}'})$ for every $\boldsymbol{i}\ \boldsymbol{i}' \in \chi_{\boldsymbol{I}}$.*
- *Identifiability of $P(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ is equivalent to saying that $P(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ is uniquely computable given the observational distribution and the graphical constrains in G.*

Causal Diagrams to the Rescue? - NO!

- **Theorem** (ID) – The image counterfactual distribution $P(\boldsymbol{I}, \boldsymbol{I}_{x'})$ is not identifiable from any combination of $\langle P(\boldsymbol{V}, \boldsymbol{I}), G \rangle$
- Challenges in the non-identifiability come from two areas/perspectives:
  1. Unknow how U_I interacts with V to produce the unobserved factors $\widetilde{U}$
  2. Given observed values of a generative factor X and its child Y, $P(\mathrm{y}'_{\mathrm{x}'} | y, x)$ is never point identifiable from the observational distribution.
- What can we do then?

Counterfactually Consistent Estimation of Image Counterfactual Distributions

- Propose two directions to relax the exact estimation
  1. Case Set W: In practical situation we may only really care about how some specific features behave after intervention but not the whole image. E.g Gender and age but not hair colour, smiling, background etc
  2. Optimal Bounds: When a query is not point identifiable, still possible to compute information bounds over the target distribution from observational data and causal diagram.

Care Set $\boldsymbol{W}$

- Feature Counterfactual Query – Denote $\boldsymbol{W}$ as a set of feature one cares about ϕ as a function mapping from $\boldsymbol{I}$ to $\boldsymbol{W}$. The feature counterfactual query regarding to $P(\boldsymbol{i}\,\boldsymbol{i}_{x'})$ is defined as

$$\int_{\boldsymbol{i}^{(1)},\boldsymbol{i}^{(2)}\chi_I} 1\left[\phi\left(\boldsymbol{i}^{(1)}\right) = \boldsymbol{w}, \phi\left(\boldsymbol{i}^{(2)}\right) = \boldsymbol{w}'\right] dP\left(\boldsymbol{i}^{1}, \boldsymbol{i}^{2}_{x'}\right)$$

  where $\boldsymbol{w} = \phi(\boldsymbol{i})$ and $\mathbf{w}' = \phi(\boldsymbol{i}')$. We denote the feature counterfactual query as $\phi\left(P\left(\boldsymbol{i}\,\boldsymbol{i}'_{x'}\right)\right)$.

- That is, the feature counterfactual query is a "push-forward" measure of $P\left(\boldsymbol{i}\,\boldsymbol{i}'_{x'}\right)$ through ϕ

- Example: Consider counterfactual image query $P(\boldsymbol{i}, \boldsymbol{i}_{Y=\mathbf{0}})$, where $\boldsymbol{i}$ is a smiling young man without gray hair and $\boldsymbol{i}'$ is a smiling old man with gray hair. If the care set $\boldsymbol{W}$ contains features gender and age, the feature counterfactual $\phi\left(P\left(\boldsymbol{i}\,\boldsymbol{i}'_{x'}\right)\right)$ calculate the probability that the original image describes a young male and the counterfactual image describes an old male with grey hair.

- **Lemma:** Consider the true underlying ASCM M^* over $\{\boldsymbol{V}, \boldsymbol{I}\}$, and a feature set with mapping function $\boldsymbol{\phi} = h^*_{\boldsymbol{W}}$, where $h^*_{\boldsymbol{W}}$ is the inverse function $f^*_{\boldsymbol{I}}$ with respect to $\boldsymbol{W}$ and a proxy ASCM $\widehat{M}$ over $\{\boldsymbol{V}, \boldsymbol{I}\}$. If $P^{\widehat{M}}(\boldsymbol{V}, \boldsymbol{I}) = P^{M^*}(\boldsymbol{V}, \boldsymbol{I})$,

$$h^*_{\boldsymbol{W}}\left(P^{\widehat{M}}(\boldsymbol{i}, \boldsymbol{i}'_{x'})\right) = P^{\widehat{M}}\left(\boldsymbol{w}, \boldsymbol{w}'_{x'}\right),$$

  where $\boldsymbol{w} = h_{\boldsymbol{W}}(\boldsymbol{i})$, and $\mathbf{w}' = h_{\boldsymbol{W}}(\boldsymbol{i}')$.

- If $\widehat{M}$ agrees on the observational distribution of $M^*$and the care set $\boldsymbol{W}$ is a subset of observed generative factors, the feature counterfactual query is equivalent to a counter factual query $P^{\widehat{M}}(\boldsymbol{w}, \boldsymbol{w}'_{x'})$ over $\boldsymbol{W}$ induced by $\widehat{M}_0$ at the generative level.

Ctf-Consistent Estimator

- Consider a features set $\boldsymbol{W} \subseteq \boldsymbol{V}$ and its mapping function $\phi = h^*_{\boldsymbol{W}}$ where $h^*_{\boldsymbol{W}}$ is the inverse function of $f^*_{\boldsymbol{I}}$ regarding $\boldsymbol{W}$. $P^{\widehat{M}}(\boldsymbol{i}, \boldsymbol{i}'_{\boldsymbol{x}'})$ is said to be a Ctf-consistent estimator of $P^{M^*}(\boldsymbol{i}, \boldsymbol{i}'_{\boldsymbol{x}'})$ with respect to $\boldsymbol{W}$ if (1) the observational distribution induced by $\widehat{M}$ and M^* are the same, namely $P^{\widehat{M}}(\boldsymbol{V}, \boldsymbol{I}) = P^{M^*}(\boldsymbol{V}, \boldsymbol{I})$ and (2) the feature counterfactual query $\phi\left(P^{\widehat{M}}(\boldsymbol{w}, \boldsymbol{w}'_{\boldsymbol{x}'})\right)$ is within the optimal bound of $P(\boldsymbol{w}, \boldsymbol{w}'_{\boldsymbol{x}'})$ derived by $P(\boldsymbol{V})$ and G, where $\boldsymbol{w} = h^*_{\boldsymbol{W}}(\boldsymbol{i})$ and $\boldsymbol{w}' = h_{\boldsymbol{W}}(\boldsymbol{i}')$; The proxy quantity $P^{\widehat{M}}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ is said to be a Ctf-Consistent estimator of the true $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ with respect to $\boldsymbol{W}$ if $P^{\widehat{M}}(\boldsymbol{i}\ \boldsymbol{i}_{\boldsymbol{x}'})$ is Ctf-Consistent for every $\boldsymbol{i}, \boldsymbol{i}' \in \chi_{\boldsymbol{I}}$.
- In other words, the observational distribution induced by the proxy model is the same as the true model and the feature counterfactual query induced by the proxy model is within the optimal bound of $P(\boldsymbol{w}, \boldsymbol{w}'_{\boldsymbol{w}})$, then the corresponding image counterfactual query can be regarded as a Ctf-consistent estimation of the true image counterfactual image querry .
- **Theorem** (Counterfactually Consistent Estimation): $P^{\widehat{M}}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ is a Ctf-consistent Estimator with respect to $\boldsymbol{W} \subseteq \boldsymbol{V}$ of $P^{M^*}(\boldsymbol{I}, \boldsymbol{I}_{\boldsymbol{x}'})$ if $\widehat{M} \in \Omega_{\boldsymbol{I}}(G)$ and $P^{\widehat{M}}(\boldsymbol{V}, \boldsymbol{I}) = P(\boldsymbol{V}, \boldsymbol{I})$.

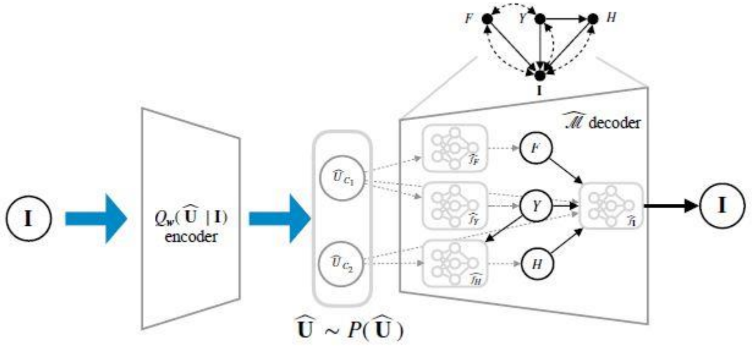
NCM for Estimating and Sampling

- Need method to train G-Constrained causal deep generative models (G-NCM)for to objectives:
  - Fit observation a distribution $P(\boldsymbol{V}, \boldsymbol{I})$
    - Proxy G-NCM $\widehat{M}$ serves as a decoder to approx. $P(\boldsymbol{I}|\widehat{\boldsymbol{U}})$ with the prior $P(\widehat{\boldsymbol{U}})$.
    - Separate deep NN $Q_{\omega}(\widehat{\boldsymbol{U}}|\boldsymbol{I})$ acts as an encoder to approx. $P(\widehat{\boldsymbol{U}}|\boldsymbol{I})$
  - Sample images $\boldsymbol{i}$ and their counterfactual counterparts $\boldsymbol{i}'$ from them.
- Prefer to fit $P(\boldsymbol{I})$ and $P(\boldsymbol{V}|\boldsymbol{I})$ separately
  - $P(\boldsymbol{I})$ learned by min. data negative loglikelihood through VAEs
  - For $P(\boldsymbol{V}|\boldsymbol{I})$ min. the cross entropy of the true labels I the image sampled and predicted labels.
- After training, generate samples of the target $P(\boldsymbol{I}|\boldsymbol{I}_{x'})$ by first sampling $\hat{\boldsymbol{u}}$ from $P(\widehat{\boldsymbol{U}})$ giving image sample $\boldsymbol{i}$ derived from $\boldsymbol{I}^{\widehat{M}}(\hat{\boldsymbol{u}})$
- The counterfactual image $\hat{\boldsymbol{i}}_{x'}$ could be derived through $\boldsymbol{I}^{\widehat{M_{x'}}}(\hat{\boldsymbol{u}})$

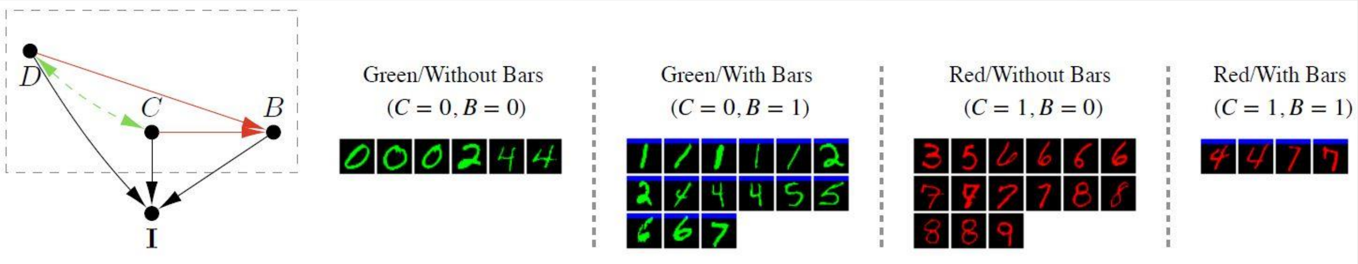
Algorithm 1 ANCM

Input: Data $\{\widehat{P}^{\mathcal{M}^*}(\mathbf{V}, \mathbf{I}) = \{\mathbf{v}_k, \mathbf{I}_k\}_{k=1}^n\}$, causal diagram $\mathcal{G}$, temperature λ, learning rate η, training epochs T.

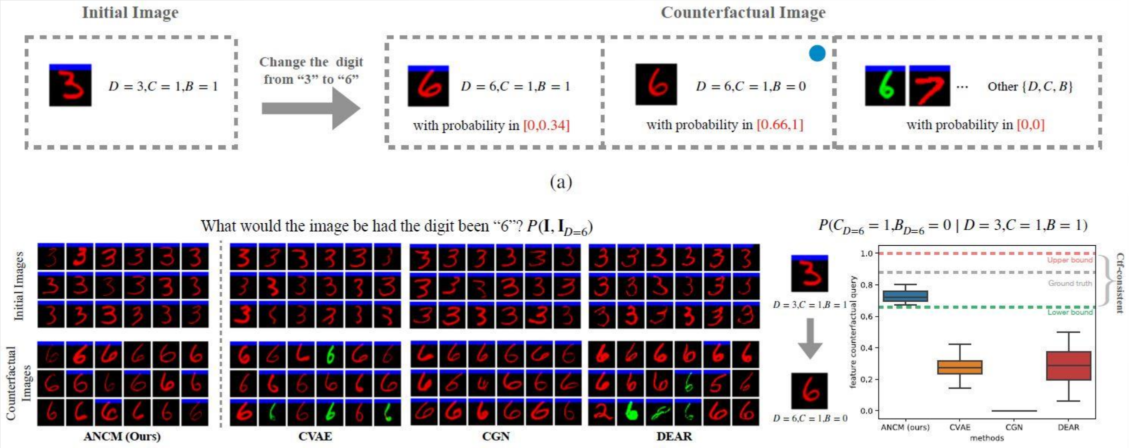
```
M̂ ← NCM(V, G)                                   {from Def. 1.4}
Initialize parameters θ for M̂ and ω for the inference
   network Q_ω(Û | I)
for t ← 1 to T do
   L ← L_1(θ, ω, P̂^{M*}(V, I)) + λ L_2(θ, ω, P̂^{M*}(V, I))
   θ ← θ − η∇L
   ω ← ω − η∇L
end for
```



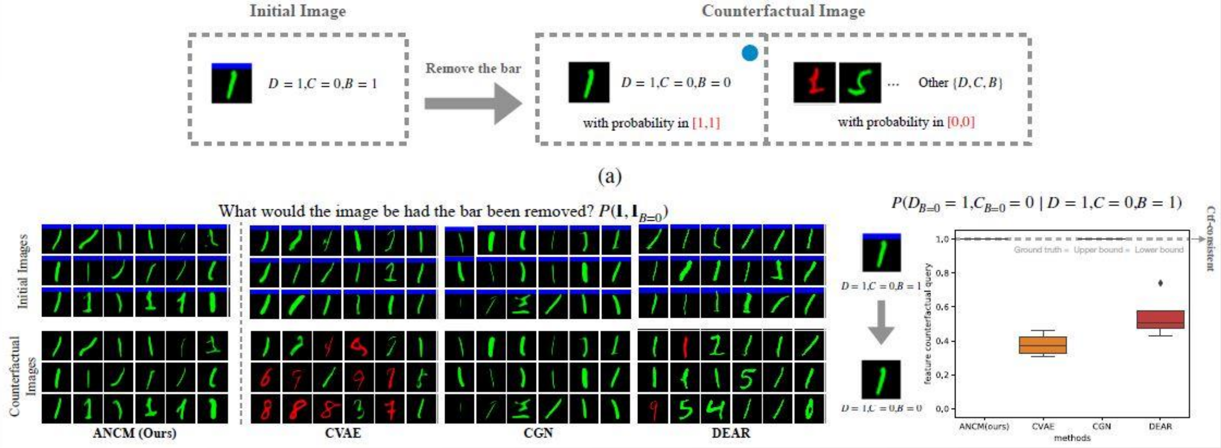
Experimental Results



Causal Diagram and sample for "Backdoor" setting. There are more red larger digits and green smaller digits; larger digits are less likely to have a bar on top; red digits are less likely to have a bar on top.

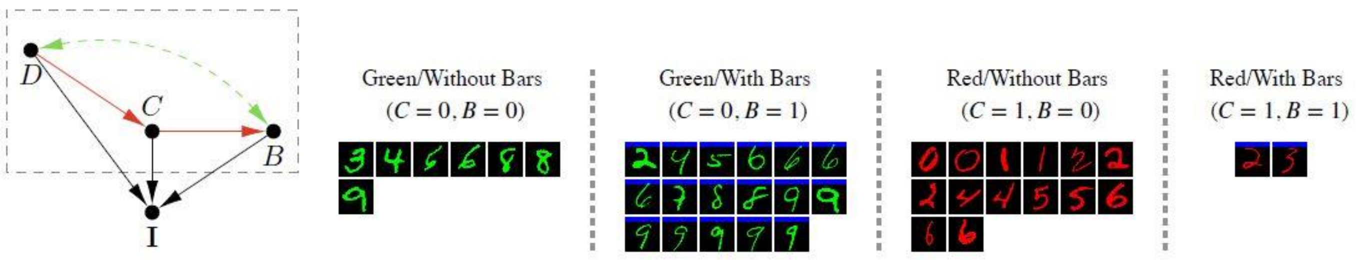


(a) Optimal bound of feature counterfactual queries when editing a red 3 with a bar to a 6.
(b) The counterfactual image generation results when editing a red 3 with a bar to 6.
(c) Selected feature counterfactual query estimates.

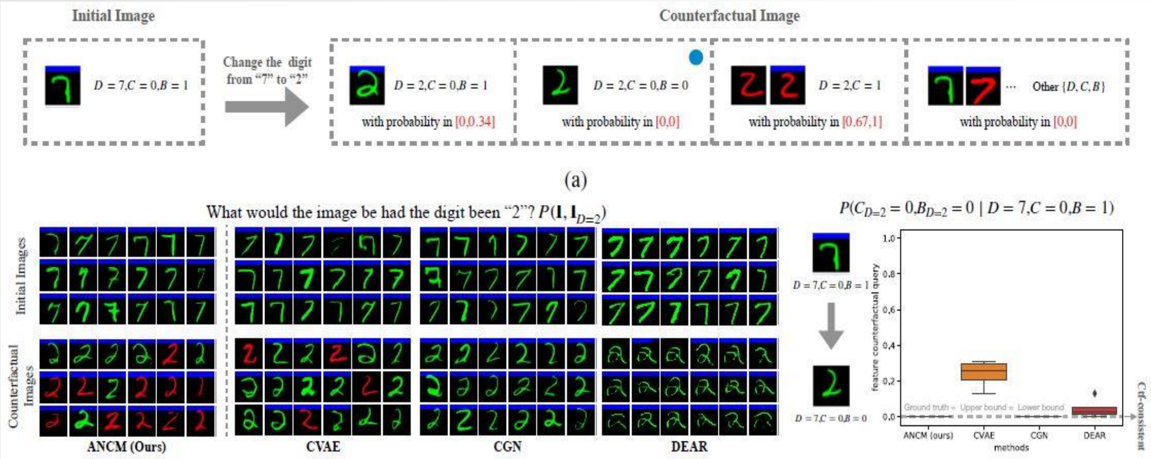


(a) Optimal bound of feature counterfactual queries when editing a red 3 with a bar to a 6.
(b) The counterfactual image generation results when editing a red 3 with a bar to 6.
(c) Selected feature counterfactual query estimates.

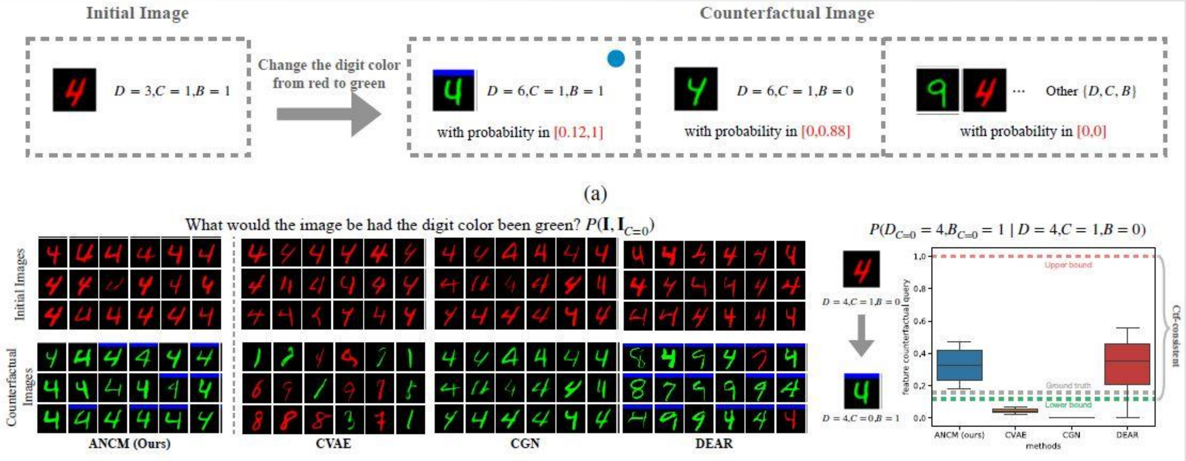
Front Door MNIST Model



Causal Diagram and sample for "Front Door" setting. Bigger digits are likely to be green; red digits are less likely to have a bar on top; there are bigger digits with bars and smaller digits without bars.



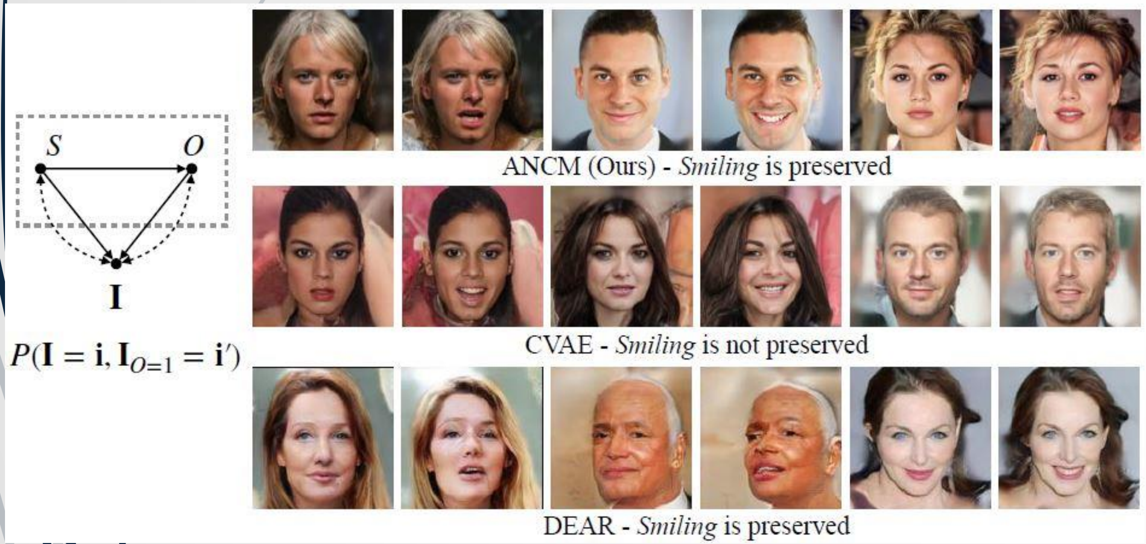
(a) Optimal bound of feature counterfactual queries when editing a green 7 with a bar to a 2.
(b) The counterfactual image generation results when editing a green 7 with a bar to a 2.
(c) Selected feature counterfactual query estimates.



(a) Optimal bound of feature counterfactual queries when editing a red 4 without a bar green.
(b) The counterfactual image generation results when editing a red 4 with a bar to green.
(c) Selected feature counterfactual query estimates.

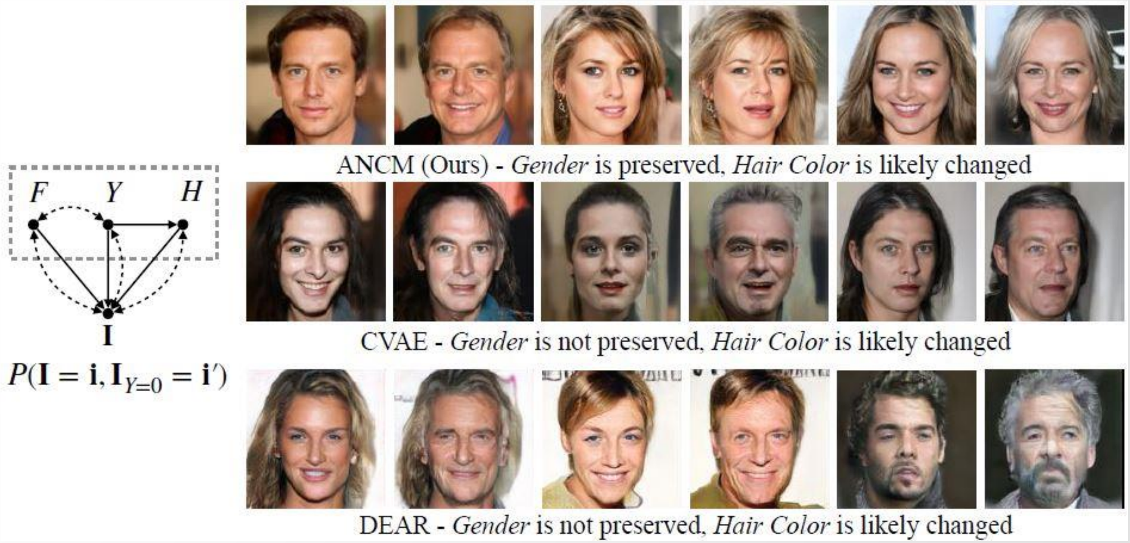
CelebA-HQ Experiments

What would the image be had the person opened their mouth?



Generative factors considered: Smile and Open Mouth with feature set $\boldsymbol{W} = \{S, O\}$

What would the image be had the person been older?



Generative factors considered: Female, Young, and gray Hair with feature set $\boldsymbol{W} = \{F, Y, H\}$.

Thank you for your time

Questions?