NEURAL INFORMATION PROCESSING SYSTEMS

Benchmarking Counterfactual Image Generation

Thomas Melistas[1,2,3], Nikos Spyrou[1,2,3] *, Nefeli Gkouti[1,2,3], Pedro Sanchez[3], Athanasios Vlontzos[4,6], Yannis Panagakis[1,2], Giorgos Papanastasiou[2,5], Sotirios A. Tsaftaris[2,3]

[1]National & Kapodistrian University of Athens, Greece, [2]Archimedes/Athena RC, Greece, [3]The University of Edinburgh, UK, [4]Imperial College London, UK, [5]The University of Essex, UK, [6]Spotify

Contact: {th.melistas, n.spyrou, nefeli.gkouti}@athenarc.gr

Motivation: Why we need counterfactuals?

- Plain image editing can be **insufficient**

What if...

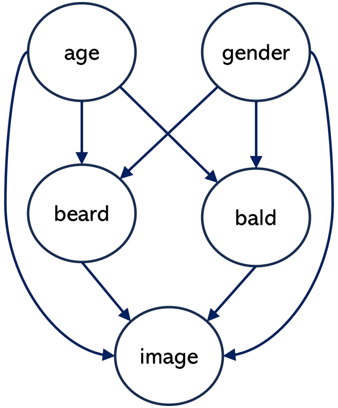
this person was female

this person was young

Image Edit

Motivation: Why we need counterfactuals?

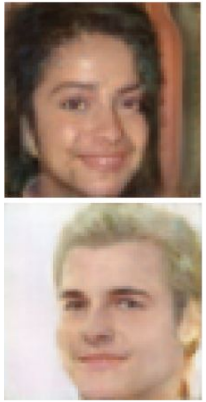
- What if we assume known **causal** relations of attributes?
- Counterfactuals are **causally** plausible*“**What if...** scenarios”*



Causal Graph

this person was female

this person was young



Causal Counterfactual

Motivation: Why is this benchmark important

- **Various** methods for Counterfactual Generation (and more coming..)
- Each method uses its own
    - evaluation metrics
    - datasets
    - causal graphs
    - experimental setup

Deep Structural Causal Models for Tractable Counterfactual Inference. Pawlowski et al. NeurIPS 2020
Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals. Dash et al. WACV 2022
High Fidelity Image Counterfactuals with Probabilistic Causal Models. Ribeiro et al. ICML 2023
Learning to synthesise the ageing brain without longitudinal data Xia et al. Medical Image Analysis 2021
…

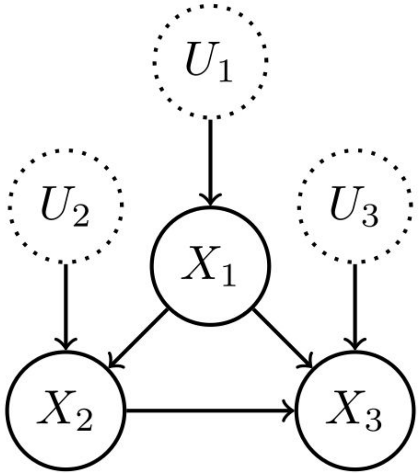
Key Contributions

- We **compare** published methods
  - on **synthetic**, **natural**, **medical** image datasets
  - using different **causal graphs**
  - on common **metrics**
- We provide a **framework** (and codebase) to bring them together
  - under the **Deep-SCM** paradigm
  - **extendable** to new models, datasets, causal graphs

Structural Causal Models (SCM)

A Structural Causal Model $\mathcal{G} := (\mathbf{S}, P(\boldsymbol{u}))$ consists of:

- A collection of structural assignments, called causal mechanisms:
  - $\mathbf{S} = \{f_i\}_{i=1}^{N}$, s.t. $x_i = f_i(u_i, \mathbf{pa}_i)$
- A joint distribution $P(\boldsymbol{u}) = \prod_{i=1}^{N} p(u_i)$ over mutually independent noise variables



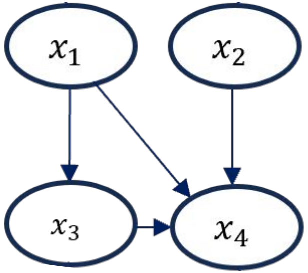
Notation:

- x_i : a random variable (endogenous)
- $\mathbf{pa}_i$: the parents of x_i (its direct causes)
- u_i : exogenous noise (commonly isotropic Gaussian)

SCM-based Interventional Counterfactuals

Pearl's Framework:

- Abduction
- Action
- Prediction

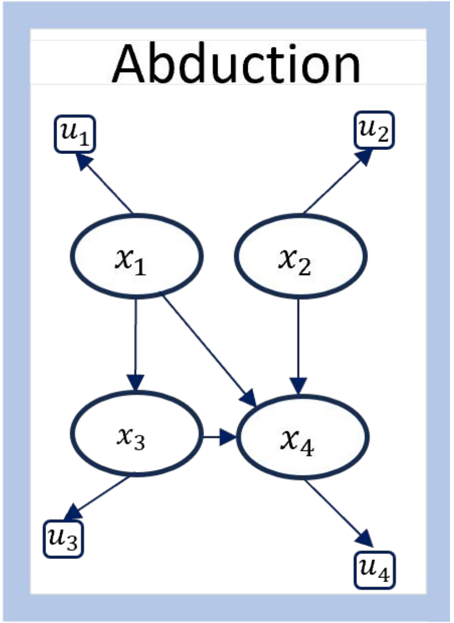


Causation, action, and counterfactuals. Pearl. Theoretical Aspects of Rationality and Knowledge 1996

SCM-based Interventional Counterfactuals

Pearl's Framework:

- Abduction
- Action
- Prediction



Infer the exogenous noise u_i for each endogenous variable x_i

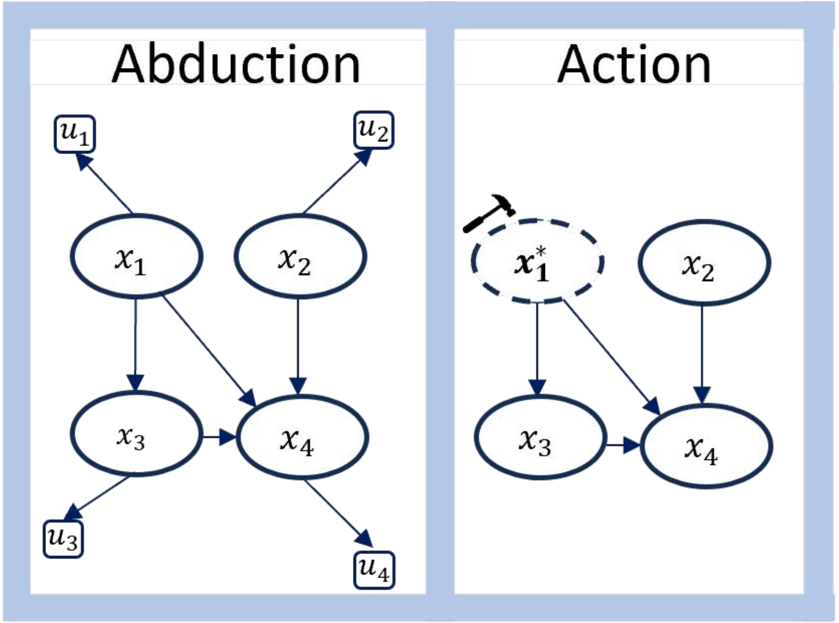
Causation, action, and counterfactuals. Pearl. Theoretical Aspects of Rationality and Knowledge 1996

SCM-based Interventional Counterfactuals

Pearl's Framework:

- Abduction
- Action
- Prediction

Intervene on a variable *do*(x_i*), forcing its value to change



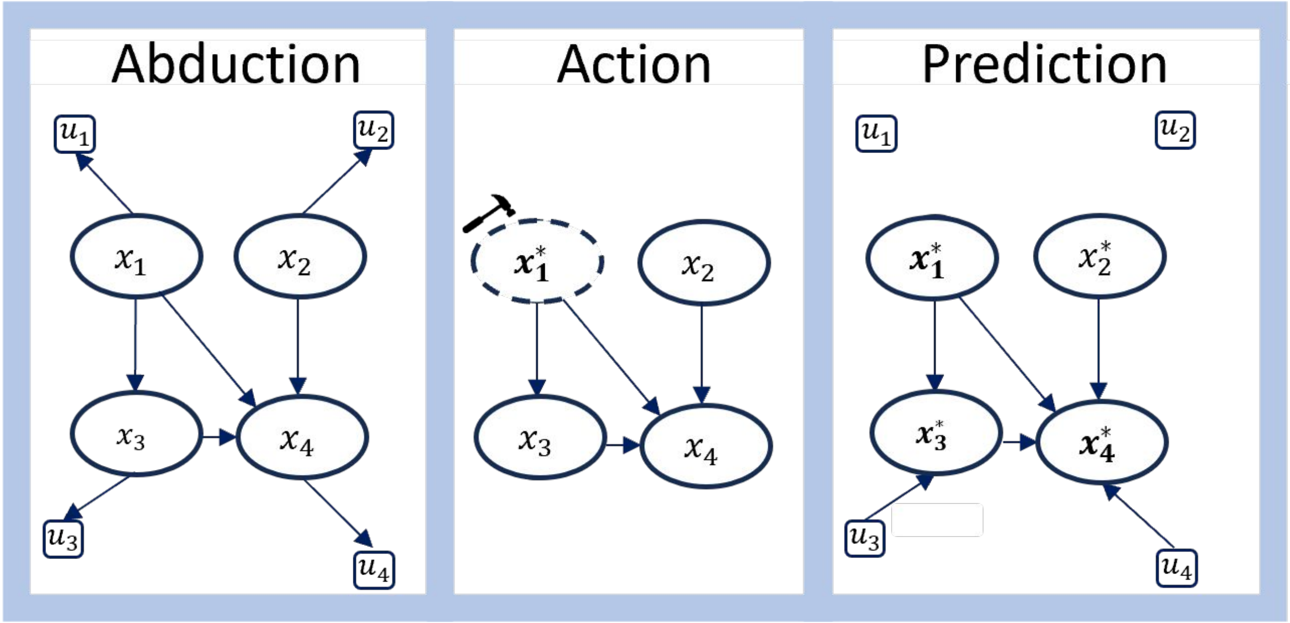
Causation, action, and counterfactuals. Pearl. Theoretical Aspects of Rationality and Knowledge 1996

SCM-based Interventional Counterfactuals

Pearl's Framework:

- Abduction
- Action
- Prediction

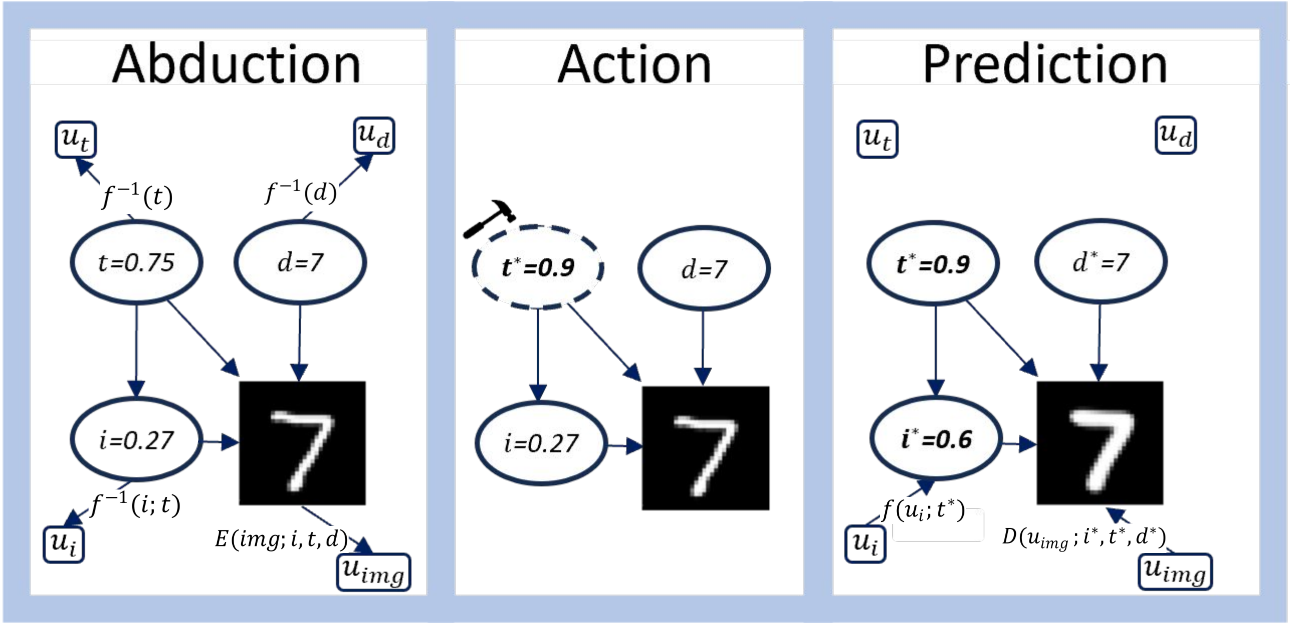
Use the modified model to produce the counterfactual



Causation, action, and counterfactuals. Pearl. Theoretical Aspects of Rationality and Knowledge 1996

Counterfactual Image Generation

MorphoMNIST
t: thickness
i: intensity
d: digit



Deep Structural Causal Models (Deep-SCM)

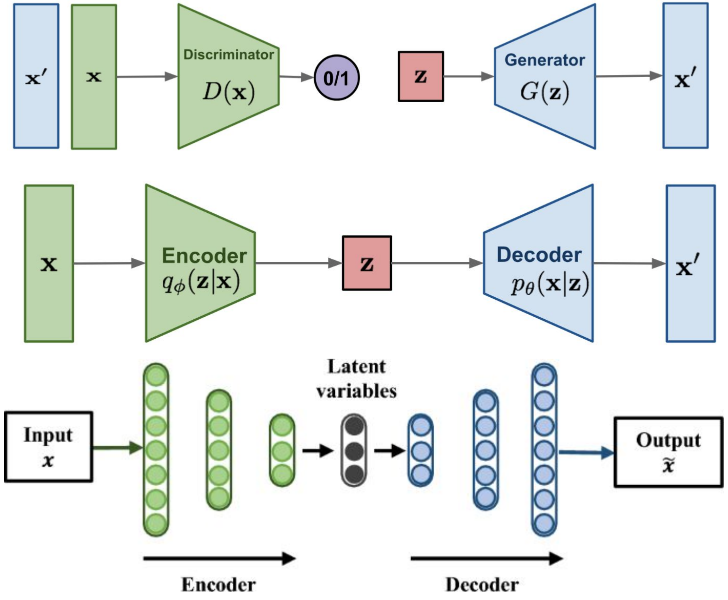
Parent attributes → **Conditional Normalizing Flow**

Conditional Generative models

GAN

VAE

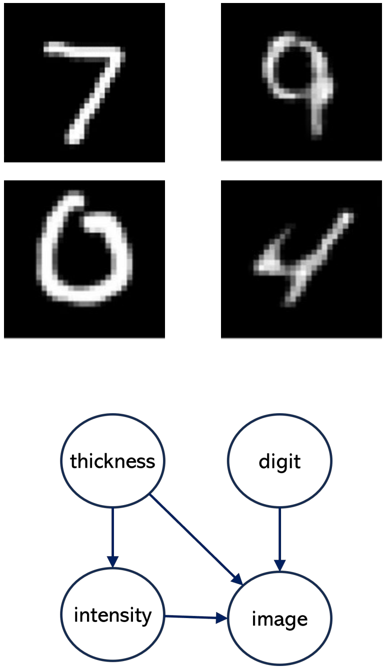
HVAE



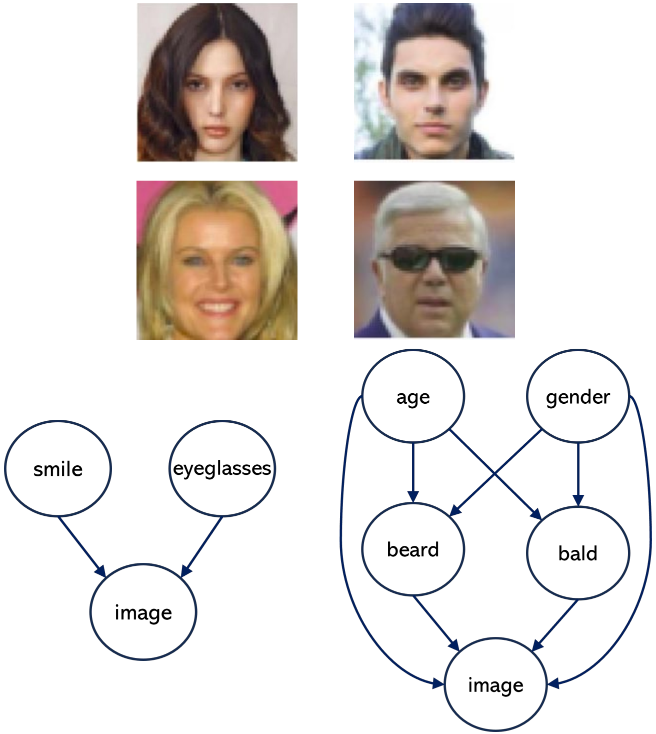
Deep Structural Causal Models for Tractable Counterfactual Inference. Pawlowski et al. NeurIPS 2020

Datasets

MorphoMNIST



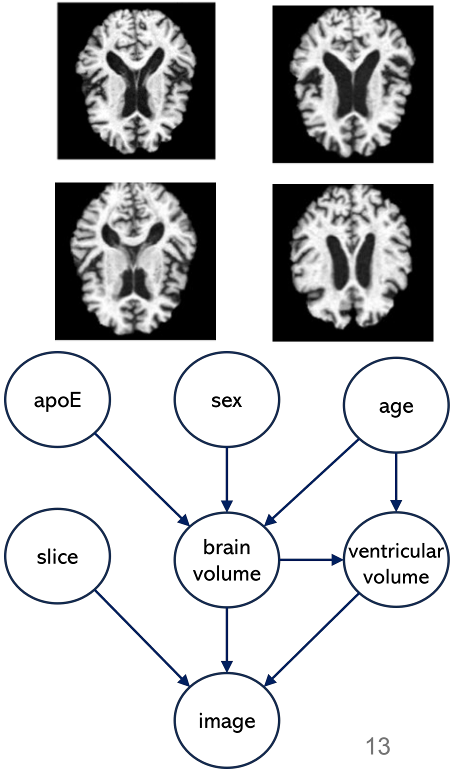
CelebA



Simple graph

Complex graph

ADNI



Evaluation of image counterfactuals

- Desired properties of image counterfactuals
  - **successful** intervention

Factual

do(Smile = True)

Counterfactual

Evaluation of image counterfactuals

- Desired properties of image counterfactuals
  - **successful** intervention
  - **minimal** changes (preserve the identity)

Factual

do(Smile = True)

Counterfactual

Evaluation of image counterfactuals

- Desired properties of image counterfactuals
    - **successful** intervention
    - **minimal** changes (preserves the identity)

- **No access** to **ground truth** counterfactuals

- 4 evaluation **metrics**:
    - Composition
    - Effectiveness
    - Realism
    - Minimality

An axiomatic characterization of causal counterfactuals. Galles & Pearl. Foundations of Science 1998

Metrics: Composition

Factual

do(null)

Successful

Unsuccessful

- The image **should not change** when performing a **null-intervention**.
- It measures the ability of the mechanism to **reconstruct** the original image.

Metrics: Effectiveness

Factual

do(Smile)

Successful

Unsuccessful

- Determines if the intervention was **successful**.
- Utilizes **classifiers/regressors** trained on the data distribution

Measuring axiomatic soundness of counterfactual image models. Miguel Monteiro et al. ICLR 2023

Metrics: Realism

Factual

do(Smile)

Successful

Unsuccessful

- It measures counterfactual image quality **by capturing its similarity** to the **factual**.
- To evaluate realism quantitatively, we leverage **FID**.

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Martin Heusel et al., Neurips 2017

Metrics: Minimality

Factual

do(Eyeglasses)

Successful

Unsuccessful

- Counterfactual leaves non-intervened attributes **unaffected**
- Counterfactual Latent Divergence (**CLD**)
  - calculates the "distance" between the counterfactual and factual images in a latent space

Who is the winner?

	Composition	Effectiveness	Realism	Minimality
MorphoMNIST	**HVAE**	**HVAE**/VAE	**HVAE**	VAE
CelebA (simple)	**HVAE**	**HVAE**/GAN	**HVAE**	**HVAE**
CelebA (complex)	**HVAE**	**HVAE**/GAN	GAN	VAE
ADNI	**HVAE**	**HVAE**/VAE	**HVAE**	**HVAE**

Key Takeaways

- A **unified framework** for rigorous benchmarking of diverse models, datasets, and causal graphs in **counterfactual image generation**

- An easy-to-use Python package
  https://github.com/gulnazaki/counterfactual-benchmark

- Interested in our work?
  Visit our page: https://gulnazaki.github.io/counterfactual-benchmark

ARCHIMEDES
Funded by the European Union
NextGenerationEU
CHAI
Royal Academy of Engineering
UKRI
Engineering and Physical Sciences Research Council
THE UNIVERSITY of EDINBURGH
School of Engineering
Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN