Benchmarking Counterfactual Image Generation

Thomas Melistas

Thomas Melistas*, Nikos Spyrou*, Nefeli Gkouti*, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, Sotirios A. Tsaftaris. "Benchmarking Counterfactual Image Generation." arxiv.org/abs/2403.20287 Under review. 2024

Overview

- Background
  - What is Counterfactual Image Generation?
  - Methods and Models
- Evaluation Metrics
- Benchmarking Setup
- Results

Counterfactual Image Generation

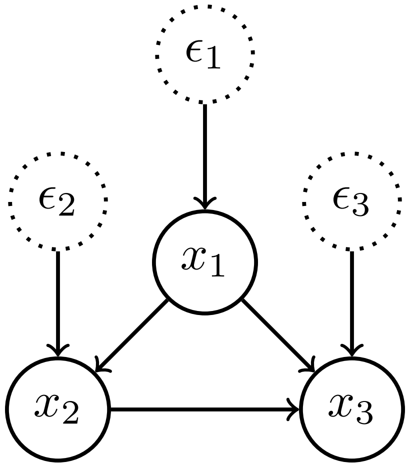
(a) Causal graph (b) Factual (c) Causal (d) Non-causal

Structural Causal Models

An SCM $\mathcal{G} := (\mathbf{S}, p(\boldsymbol{\epsilon}))$ consists of:

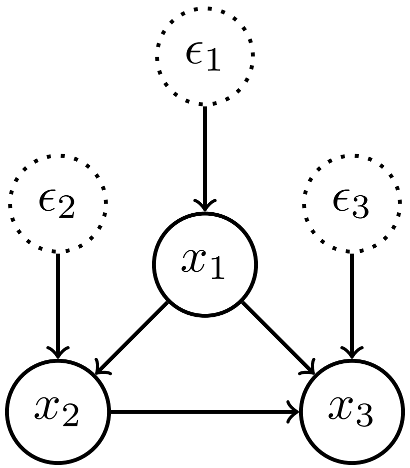
(i) structural assignments $\mathbf{S} = \{f_i\}_{i=1}^N$, s.t. $x_i := f_i(\epsilon_i, \mathbf{pa}_i)$,
(ii) a joint distribution $p(\boldsymbol{\epsilon}) = \prod_{i=1}^N p(\epsilon_i)$ over mutually independent noise variables

x_i: an **endogenous** variable (observed)
$\mathbf{pa}_i$: the parents of x_i (its direct *causes*, endogenous)
ϵ_i: an **exogenous** variable (unobserved)



Structural Causal Models

- Causal relations $\rightarrow$ directed acyclic graph (DAG)
- Acyclic $\rightarrow$ solve recursively for x_i and obtain $\mathbf{x} = \mathbf{f}(\boldsymbol{\epsilon})$
- $\mathbf{x}$: a collection of observable variables, where x_i: image and $\mathbf{pa}_i$: image attributes



Interventions and Counterfactuals

- *Interventional* distributions $P(x_j|do(x_i = y))$:
  - interventions: $x_i = f_i(\epsilon_i, \mathbf{pa}_i) \rightarrow x_i = y$
  - exogenous noise is sampled from the prior $P(\boldsymbol{\epsilon})$
- *Counterfactual* distributions $P(x_{j,x_i=y}|\mathbf{x})$:
  - interventions: as above
  - exogenous noise same with the observation $P(\boldsymbol{\epsilon}|\mathbf{x})$

Interventions and Counterfactuals

Layer	Activity	Semantics	Example
(1) Associational $p(y \mid x)$	Seeing	How would seeing x change my belief in Y?	What does a symptom tell us about the disease?
(2) Interventional $p(y \mid \mathrm{do}(x), z)$	Doing	What happens to Y if I do x?	What if I take aspirin, will my headache be cured?
(3) Counterfactual $p(y_{x'} \mid x, y)$	Imagining	Was it x that caused Y?	Was it the aspirin that stopped my headache?

Pearl J. (2009). Causality

Abduction-Action-Prediction

Counterfactuals using SCMs are computed in three steps:
(i) **Abduction**: Infer $P(\boldsymbol{\epsilon}|\mathbf{x})$, the state of the world (exogenous noise) that is compatible with the observation $\mathbf{x}$.
(ii) **Action**: Replace the structural equations $do(x_i = y)$, resulting in a modified SCM $\widetilde{\mathcal{G}} := \mathcal{G}_{\mathbf{x};do(x_i=y)} = (\widetilde{\mathbf{S}}, P(\boldsymbol{\epsilon}|\mathbf{x}))$.
(iii) **Prediction**: Use the modified model to compute $P_{\widetilde{\mathcal{G}}}(\mathbf{x})$.

Using Neural Networks for Abduction

Three categories of mechanisms:

(i) **Invertible, explicit**: Conditional Normalising Flows $\rightarrow$ attributes

(ii) **Amortised, explicit**: Conditional VAEs or Hierarchical VAEs $\rightarrow$ image

(iii) **Amortised, implicit**: Conditional GANs $\rightarrow$ image

Pawlowski et al. Deep structural causal models for tractable counterfactual inference. NeurIPS 2020

Invertible, explicit

- Normalising flows perform mappings between probability densities
- A series of invertible transformations
- For an attribute x_i we utilise a conditional NF $f(\epsilon_i; \mathbf{pa}_i)$ which is invertible: $\epsilon_i = f^{-1}(x_i; \mathbf{pa}_i)$

$$P(x_i|\mathbf{pa}_i) = p(\epsilon_i)\left|\det \nabla_{\epsilon_i} f(\epsilon_i; \mathbf{pa}_i)\right|^{-1}$$

Trippe et al. Conditional density estimation with bayesian normalising flows. 2018

Amortised, explicit (with VAE)

- Encoder: $q_\phi(z|x, pa_x)$ and Decoder: $p_\theta(x|z, pa_x)$, trained with:

  $$\mathrm{ELBO}_\beta(\phi, \theta) = \mathbb{E}_{z\sim q_\phi(z|x,pa_x)}[p_\theta(x|z, pa_x)] - \beta D_{KL}[q_\phi(z|x, pa_x)||p(z)]$$

- The above likelihood and posterior are diagonal gaussians, whose μ and σ we predict with neural networks
- Prior $p(z) \sim N(0, I)$.

Kingma & Welling. Auto-Encoding Variational Bayes. 2014

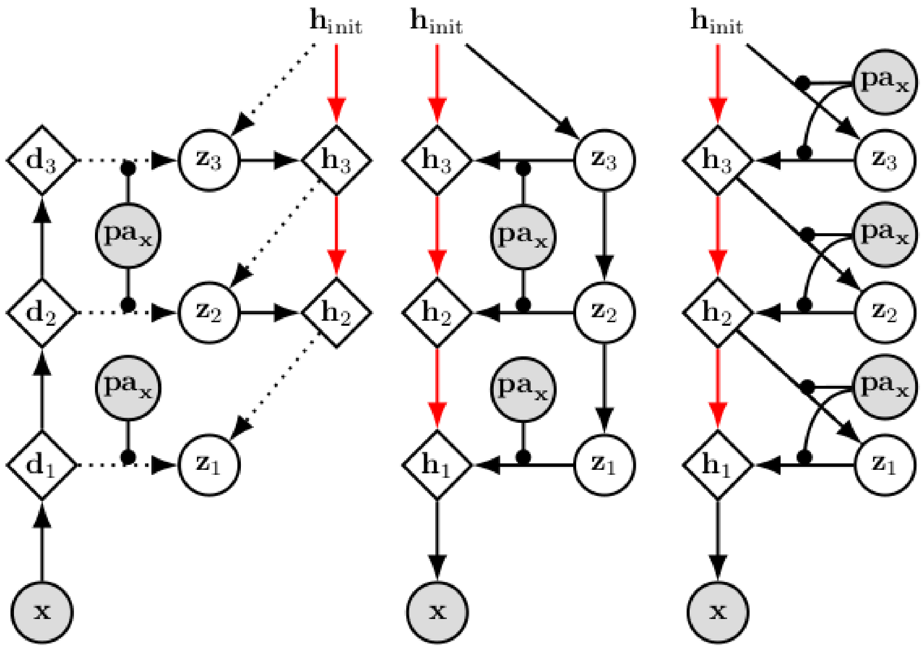
Higgins et al. Beta-vae: Learning basic visual concepts with a constrained variational framework. 2016

Amortised, explicit (with VAE)

- Noise ϵ is decomposed into $z \sim q_\phi(z|x, pa_x)$ and $u \sim N(0, I)$
- To perform counterfactual inference:
  - We sample the latent $z = \mu_\phi(x, pa_x) + \sigma_\phi(x, pa_x) * i$, $i \sim N(0, I)$
  - We sample the counterfactual $x^* = \mu_\theta(z, pa_x^*) + \sigma_\theta(z, pa_x^*) * u$, where $u = \frac{x - \mu_\theta(z, pa_x)}{\sigma_\theta(z, pa_x)}$

Amortised, explicit (HVAE)

- We have L layers of hierarchical latent variables $\boldsymbol{z} = \{z_1, z_2, \ldots, z_L\}$
- $h_i = h_{i+1} + f_\omega^i(z_i, pa_x)$
  $z_i \sim p_\theta(z_i | z_{>i})$,
  $p_\theta(x | z_{1:L}, pa_x) = N(x | \mu_\theta(h_1), \sigma_\theta(h1))$
- $q_\phi(z_{1:L} | x, pa_x) = q_\phi(z_L | x, pa_x)$
  $q_\phi(z_{L-1} | z_L, x, pa_x) \ldots q_\phi(z_1 | z_{i>1}, x, pa_x)$



Sønderby et al, Ladder variational autoencoders. 2016
Ribeiro et al. High Fidelity Image Counterfactuals with Probabilistic Causal Models. 2023

Amortised, implicit

- Encoder E: $z_x = E(x, pa_x)$
- Generator G: $x' = G(z', pa_x)$, where $z' \sim N(0, I)$ or $z' = z_x$
- Discriminator $D(x', z', pa_x)$: generated $\rightarrow$ fake, data $\rightarrow$ real

$$\min_{E,G} \max_{D} V(D, G, E) = \mathbb{E}_{q(x)p(pa_x)}[log(D(x, E(x, pa_x), pa_x))] + \mathbb{E}_{p(z)p(pa_x)}[log(1 - D(G(z, pa_x), z, pa_x))]$$

Finetuning the encoder:

$$L_x = \mathbb{E}_{x \sim q(x)} \|x - G(E(x, pa_x), pa_x)\|_2$$

$$L_z = \mathbb{E}_{z \sim p(z)} \|z - E(G(z, pa_x), pa_x)\|$$

To produce counterfactuals:

$$x^* = G(E(x, pa_x), pa_x^*))$$

Dumoulin et al, Adversarially learned inference. 2017

Dash et al: Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. 2022

Metrics

We use 4 metrics to evaluate the generated counterfactuals

- **Composition**
- **Effectiveness**
- **Realism** (FID)
- **Minimality** (CLD)

Composition

- Conceptually: the image and its attributes do not change without intervention
- *If we force a variable X to a value x it would have without the intervention, it should have no effect on the other variables*
- Therefore, a *null-intervention* applied m times: $f_\emptyset^m$ should change no variable
- $\text{composition}^m(x, pa_x) = d(x, f_\emptyset^m(x, pa_x))$,
  where $d(\,.\,,\,.\,)$ is a suitable distance metric

Galles & Pearl: An axiomatic characterization of causal counterfactuals. 1998

Monteiro et al. Measuring axiomatic soundness of counterfactual image models. 2022

Effectiveness

- Conceptually: how successful was the performed intervention
- *If we force a variable X to have the value x, then X will take on the value x*
- We train an anti-causal predictor g_θ^i on observations, for each pa_x^i
- $\text{effectiveness}_i(x^*, {pa^*}_x^i) = d(g_\theta^i(x^*), {pa^*}_x^i)$,
  where $d(.\,,.)$: classification metric for categorical variables, regression for continuous

Galles & Pearl: An axiomatic characterization of causal counterfactuals. 1998

Monteiro et al. Measuring axiomatic soundness of counterfactual image models. 2022

Fréchet Inception Distance (FID)

- Conceptually: Counterfactual image quality
- Measures the similarity of counterfactual images to observational data
- We use Inception-v3 trained on ImageNet to extract features
- Defined as:

$$d^2((m_q, C_q), (m_p, C_p)) = ||m_q - m_p||_2^2 + \mathrm{Tr}(C_q + C_p - 2(C_q C_p)^{\frac{1}{2}})$$

where q: counterfactual features distribution, p: factual features distribution Fréchet distance d is defined as the distance between the Gaussian with mean (m_q, C_q) obtained from q and the Gaussian with mean (m_p, C_p) obtained from p.

Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017

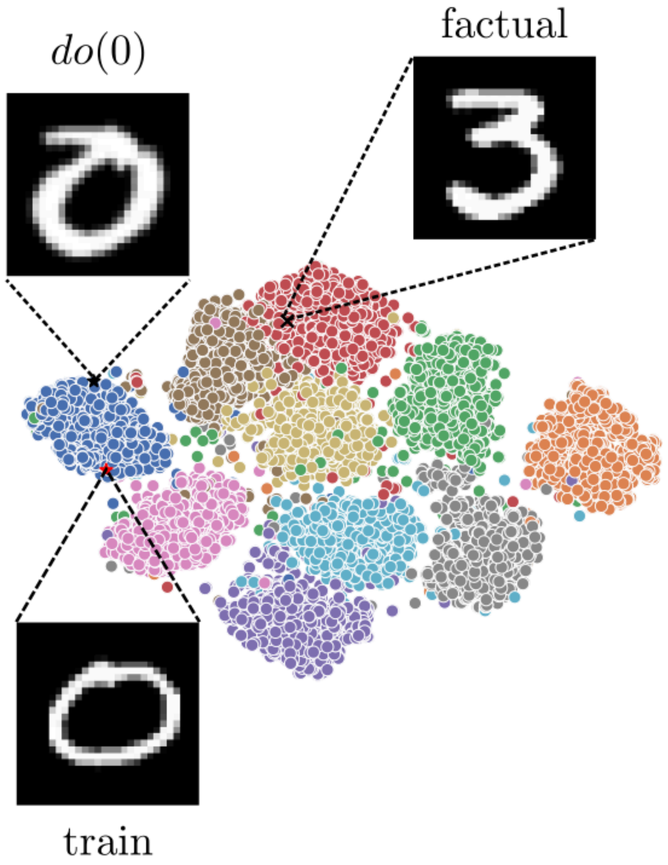
Counterfactual Latent Divergence (CLD)

- Conceptually: Counterfactual only differs in the intervened parent attribute
- Formally:

$$\mathrm{CLD} = \log(w_1 \exp P(S_{x^*} \leq div) + w_2 \exp P(S_x \geq div))$$

,where: $div = d(x, x*)$, $S_x = \{d(x, x')|pa_{x'} = pa_x\}$,
$S_{x^*} = \{d(x, x')|pa_{x'} = pa_{x^*}\}$

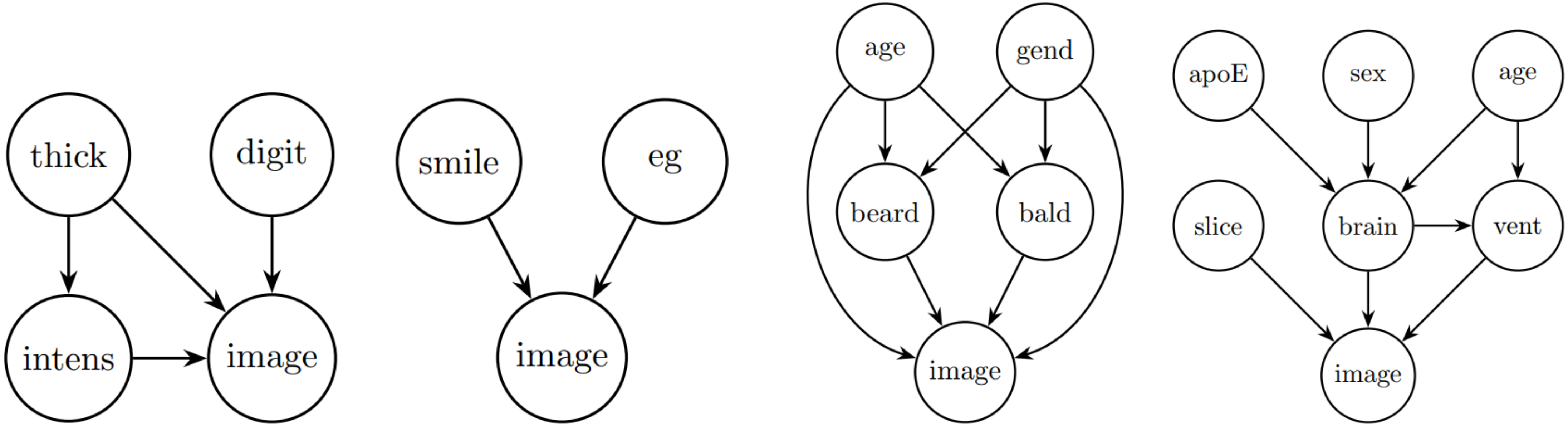
- $d(.\,,.)$: KL-divergence between the latents given by an unconditional VAE



Sanchez & Tsaftaris: Diffusion Causal Models for Counterfactual Estimation. 2022

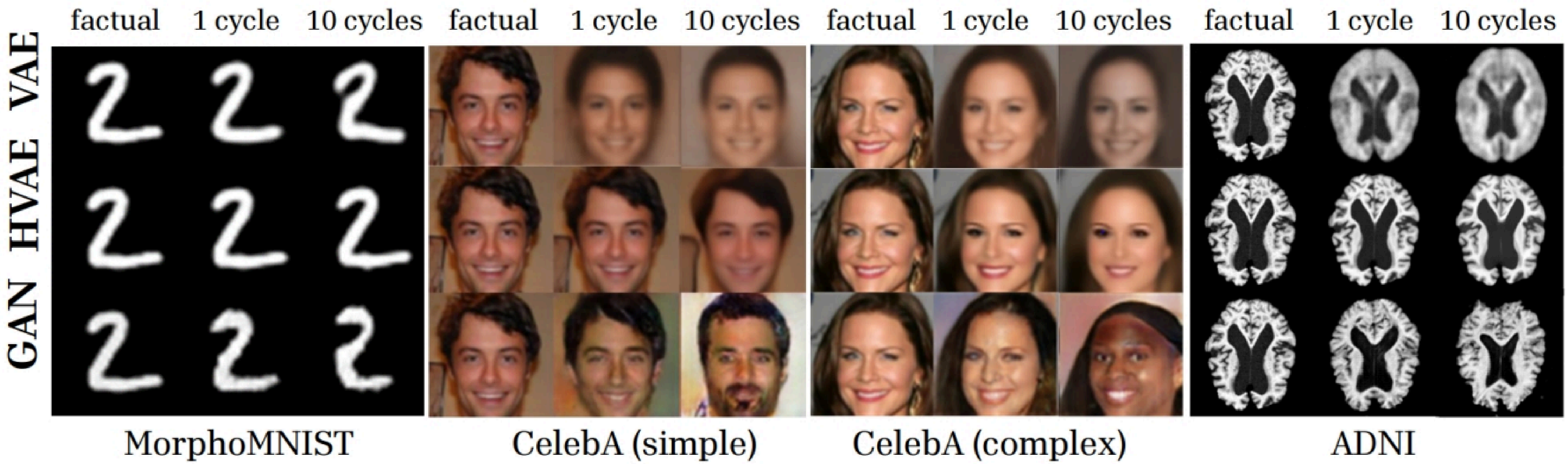
Datasets used for benchmarking

- **MorphoMNIST**: Synthetic dataset of digits (32×32)
- **CelebA**: Human faces (64×64) $\rightarrow$ simple & complex causal graph
- **ADNI**: 2D slices of brain MRIs (192×192)



(a) MorphoMNIST (b) CelebA (simple) (c) CelebA (complex) (d) ADNI

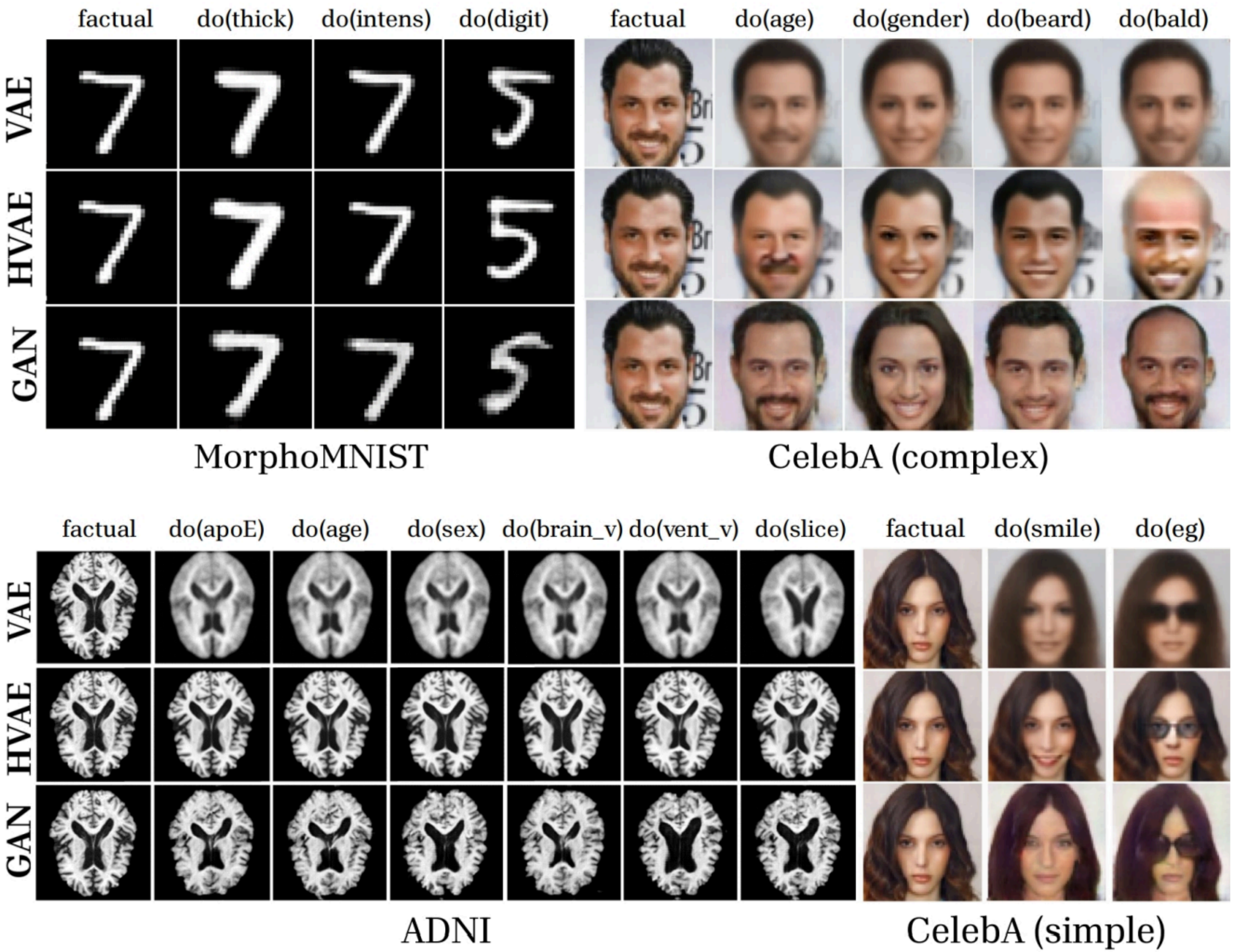
Composition



Composition

Model	l_1 image space ↓		LPIPS ↓	
	1 cycle	10 cycles	1 cycle	10 cycles
MorphoMNIST				
VAE	$2.600_{1.454}$	$7.698_{4.199}$	$0.026_{0.003}$	$0.075_{0.003}$
HVAE	$\mathbf{0.438_{0.178}}$	$\mathbf{1.550_{0.541}}$	$\mathbf{0.006_{0.001}}$	$\mathbf{0.024_{0.003}}$
GAN	$3.807_{1.822}$	$11.697_{5.750}$	$0.049_{0.004}$	$0.184_{0.008}$
CelebA (simple/complex)				
VAE	$127.8_{18.1}/\mathbf{121.2_{13.2}}$	$\mathbf{123.4_{22.1}}/127.9_{21.1}$	$0.295_{0.004}/0.282_{0.061}$	$0.424_{0.005}/0.412_{0.091}$
HVAE	$129.4_{11.6}/122.7_{10.4}$	$138.8_{23.4}/\mathbf{124.6_{16.4}}$	$\mathbf{0.063_{0.003}}/\mathbf{0.122_{0.033}}$	$\mathbf{0.200_{0.008}}/\mathbf{0.240_{0.053}}$
GAN	$\mathbf{115.3_{22.0}}/127.6_{17.3}$	$128.4_{21.4}/131.8_{23.3}$	$0.290_{0.003}/0.276_{0.074}$	$0.462_{0.005}/0.490_{0.121}$
ADNI				
VAE	$18.882_{1.786}$	$30.250_{3.389}$	$0.306_{0.008}$	$0.384_{0.006}$
HVAE	$\mathbf{3.384_{0.367}}$	$\mathbf{7.456_{0.622}}$	$\mathbf{0.101_{0.012}}$	$\mathbf{0.156_{0.014}}$
GAN	$24.261_{1.821}$	$32.794_{3.578}$	$0.268_{0.009}$	$0.323_{0.007}$

Qualitative Results



Effectiveness

Model	Thickness (t) MAE ↓			Intensity (i) MAE ↓			Digit (y) Acc. ↑		
	$do(t)$	$do(i)$	$do(y)$	$do(t)$	$do(i)$	$do(y)$	$do(t)$	$do(i)$	$do(y)$
VAE	$0.109_{0.01}$	$0.333_{0.02}$	$0.139_{0.01}$	$3.15_{0.26}$	$5.33_{0.29}$	$4.64_{0.35}$	$\mathbf{0.989_{0.01}}$	$\mathbf{0.988_{0.01}}$	$0.775_{0.02}$
HVAE	$\mathbf{0.086_{0.09}}$	$\mathbf{0.224_{0.03}}$	$\mathbf{0.117_{0.01}}$	$\mathbf{1.99_{0.18}}$	$\mathbf{3.52_{0.26}}$	$\mathbf{2.10_{0.17}}$	$0.985_{0.01}$	$0.935_{0.02}$	$\mathbf{0.972_{0.02}}$
GAN	$0.228_{0.01}$	$0.680_{0.02}$	$0.393_{0.02}$	$9.43_{0.59}$	$15.14_{0.99}$	$12.39_{0.57}$	$0.961_{0.02}$	$0.966_{0.01}$	$0.451_{0.024}$

CelebA (simple)				
Model	**Smiling (s) F1 ↑**		**Eyeglasses (e) F1 ↑**	
	$do(s)$	$do(e)$	$do(s)$	$do(e)$
VAE	$0.897_{0.02}$	$0.987_{0.01}$	$0.938_{0.05}$	$0.810_{0.02}$
HVAE	$\mathbf{0.998_{0.01}}$	$\mathbf{0.997_{0.01}}$	$0.883_{0.06}$	$\mathbf{0.981_{0.02}}$
GAN	$0.819_{0.02}$	$0.873_{0.01}$	$\mathbf{0.957_{0.03}}$	$0.891_{0.01}$

CelebA (complex)								
	Age (a) F1 ↑				**Gender (g) F1 ↑**			
	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
VAE	$0.35_{0.04}$	$0.782_{0.02}$	$0.816_{0.02}$	$0.819_{0.02}$	$0.977_{0.01}$	$0.909_{0.02}$	$0.959_{0.02}$	$\mathbf{0.973_{0.01}}$
HVAE	$\mathbf{0.654_{0.1}}$	$\mathbf{0.893_{0.04}}$	$\mathbf{0.908_{0.03}}$	$\mathbf{0.899_{0.03}}$	$\mathbf{0.988_{0.02}}$	$0.949_{0.03}$	$\mathbf{0.994_{0.01}}$	$0.95_{0.03}$
GAN	$0.413_{0.04}$	$0.71_{0.02}$	$0.818_{0.02}$	$0.799_{0.01}$	$0.952_{0.01}$	$\mathbf{0.982_{0.01}}$	$0.92_{0.01}$	$0.961_{0.01}$
	Beard (br) F1 ↑				**Bald (bl) F1 ↑**			
	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
VAE	$0.944_{0.01}$	$0.828_{0.03}$	$0.296_{0.05}$	$\mathbf{0.945_{0.02}}$	$\mathbf{0.023_{0.03}}$	$0.496_{0.05}$	$0.045_{0.04}$	$0.412_{0.03}$
HVAE	$\mathbf{0.952_{0.03}}$	$\mathbf{0.951_{0.03}}$	$\mathbf{0.441_{0.11}}$	$0.916_{0.04}$	$0.02_{0.05}$	$\mathbf{0.86_{0.05}}$	$0.045_{0.07}$	$\mathbf{0.611_{0.04}}$
GAN	$0.908_{0.01}$	$0.838_{0.02}$	$0.233_{0.03}$	$0.907_{0.01}$	$0.021_{0.02}$	$0.82_{0.02}$	$\mathbf{0.055_{0.02}}$	$0.492_{0.02}$

Model	Brain volume (b) MAE ↓			Ventricular volume (v) MAE ↓			Slice (s) F1 ↑		
	$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$
VAE	$0.17_{0.03}$	$0.15_{0.06}$	$0.15_{0.06}$	$0.08_{0.05}$	$0.20_{0.04}$	$0.08_{0.05}$	$\mathbf{0.52_{0.15}}$	$\mathbf{0.48_{0.15}}$	$\mathbf{0.46_{0.10}}$
HVAE	$\mathbf{0.09_{0.03}}$	$\mathbf{0.12_{0.06}}$	$\mathbf{0.13_{0.06}}$	$\mathbf{0.06_{0.04}}$	$\mathbf{0.04_{0.01}}$	$\mathbf{0.06_{0.04}}$	$0.38_{0.15}$	$0.41_{0.16}$	$0.41_{0.11}$
GAN	$0.17_{0.02}$	$0.16_{0.07}$	$0.16_{0.06}$	$0.12_{0.02}$	$0.22_{0.03}$	$0.12_{0.03}$	$0.14_{0.03}$	$0.16_{0.03}$	$0.05_{0.02}$

Realism (FID) & Minimality (CLD)

Model	MorphoMNIST		CelebA (simple/complex)		ADNI	
	FID ↓	CLD ↓	FID ↓	CLD ↓	FID ↓	CLD ↓
VAE	10.124	**0.268**	66.412/59.393	0.301/**0.299**	278.245	0.352
HVAE	**5.362**	0.272	**22.047**/35.712	**0.295**/0.305	**74.696**	**0.347**
GAN	35.568	0.286	31.560/**27.861**	0.38/0.304	113.749	0.353

Conclusions

- HVAE outperforms other models across metrics and datasets
- GAN counterfactuals more realistic than VAE but far from factuals for complex datasets
- Amortised implicit mechanisms $\rightarrow$ better abduction,
  but hierarchical latents are important

Future Work

- Extend to other
  - Generative models for the image mechanism (i.e. Diffusion Models)
  - Counterfactual paradigms (i.e. Deep Twin Networks, Backtracking Counterfactuals)
- Metrics
  - Limit bias of used model-dependent metrics (i.e. effectiveness, CLD)
  - Come up with new metrics

Sanchez & Tsaftaris: Diffusion Causal Models for Counterfactual Estimation. 2022
Vlontzos et al. Estimating categorical counterfactuals via deep twin networks. 2023
Kladny et al. Deep backtracking counterfactuals for causally compliant explanations. 2024

$$Thank \quad you \quad for \quad your \quad attention!!!$$