Principled Knowledge Extrapolation with GANs

Ruili Feng[1], Jie Xiao[1], Kecheng Zheng[1], Deli Zhao[2], Jingren Zhou[3], Qibin Sun[1], Zheng-Jun Zha[1]

[1]University of Science and Technology of China, Hefei, China

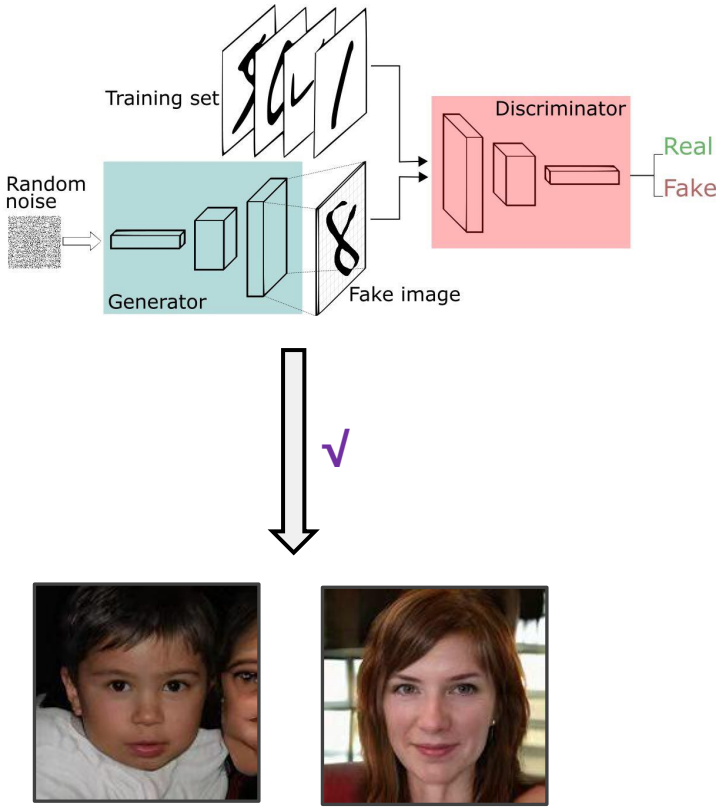
[2]Ant Research, Hangzhou, China

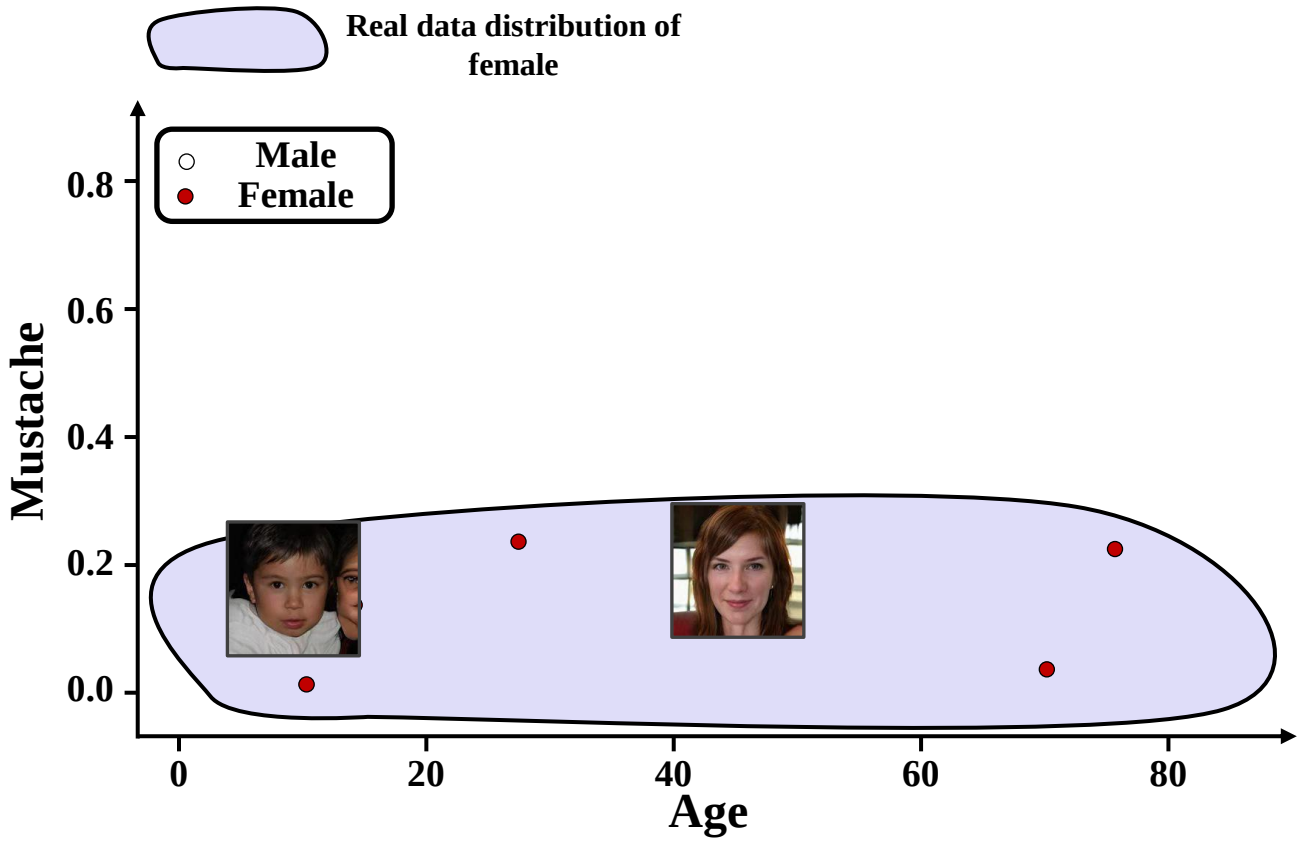
[3]Alibaba Group, Hangzhou, China

ICML 2022, held in Baltimore, Maryland USA

What is Counterfactual Synthesis

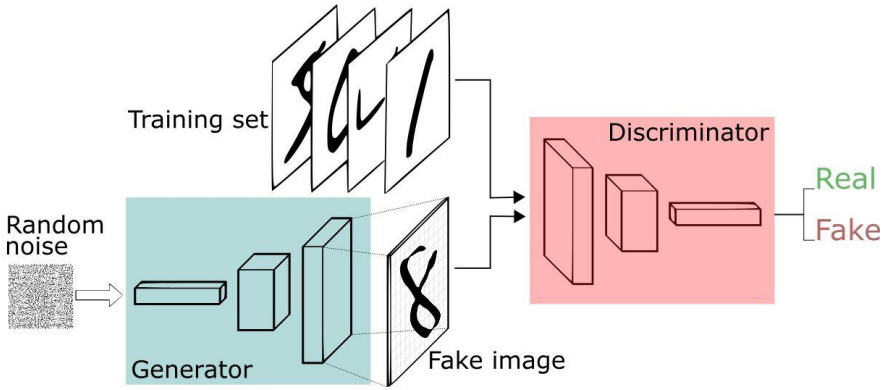
- Generative Networks can well generate high fidelity examples whose distribution is from that of the given training data.

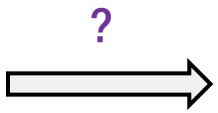




What is Counterfactual Synthesis

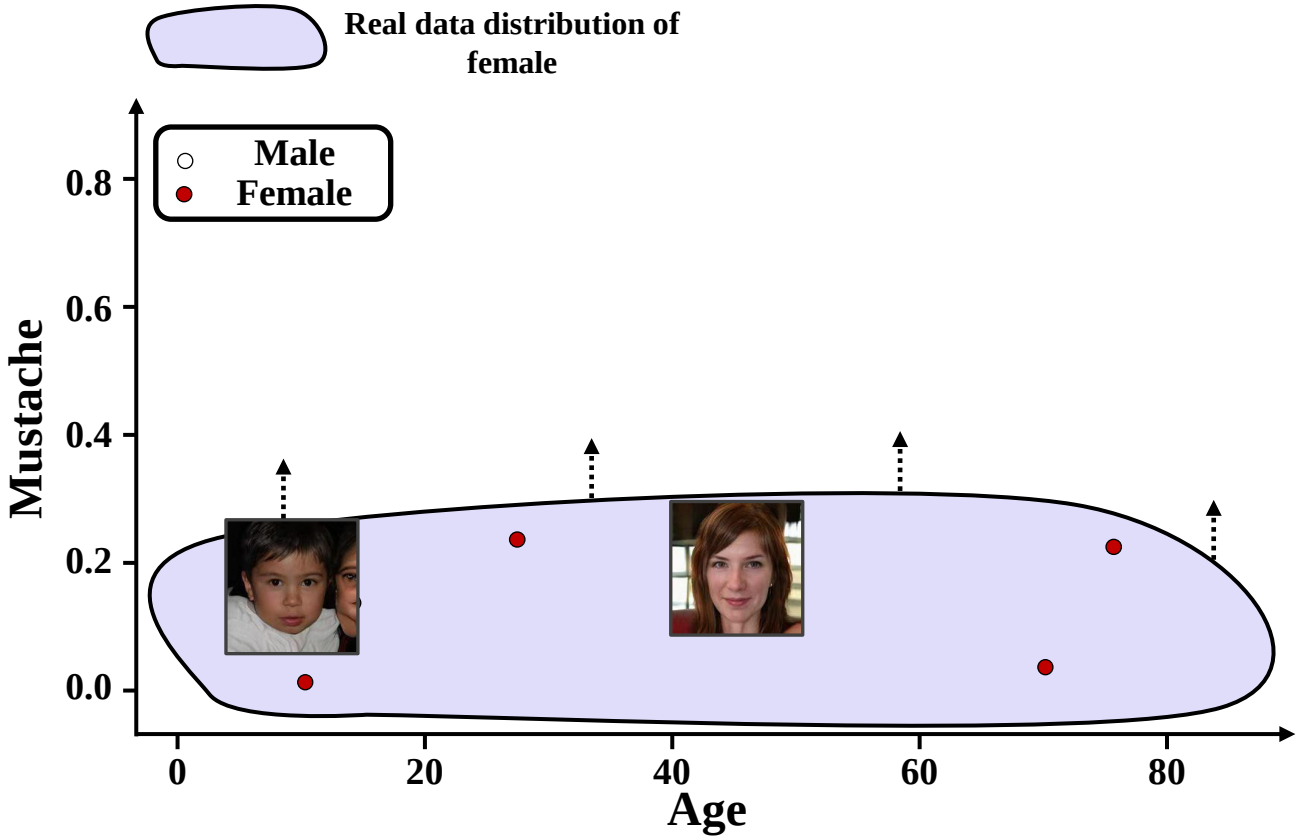
- Current methods fail to generate high fidelity counterfactual results;
- Real data distribution excludes the cases of children or female in mustache.





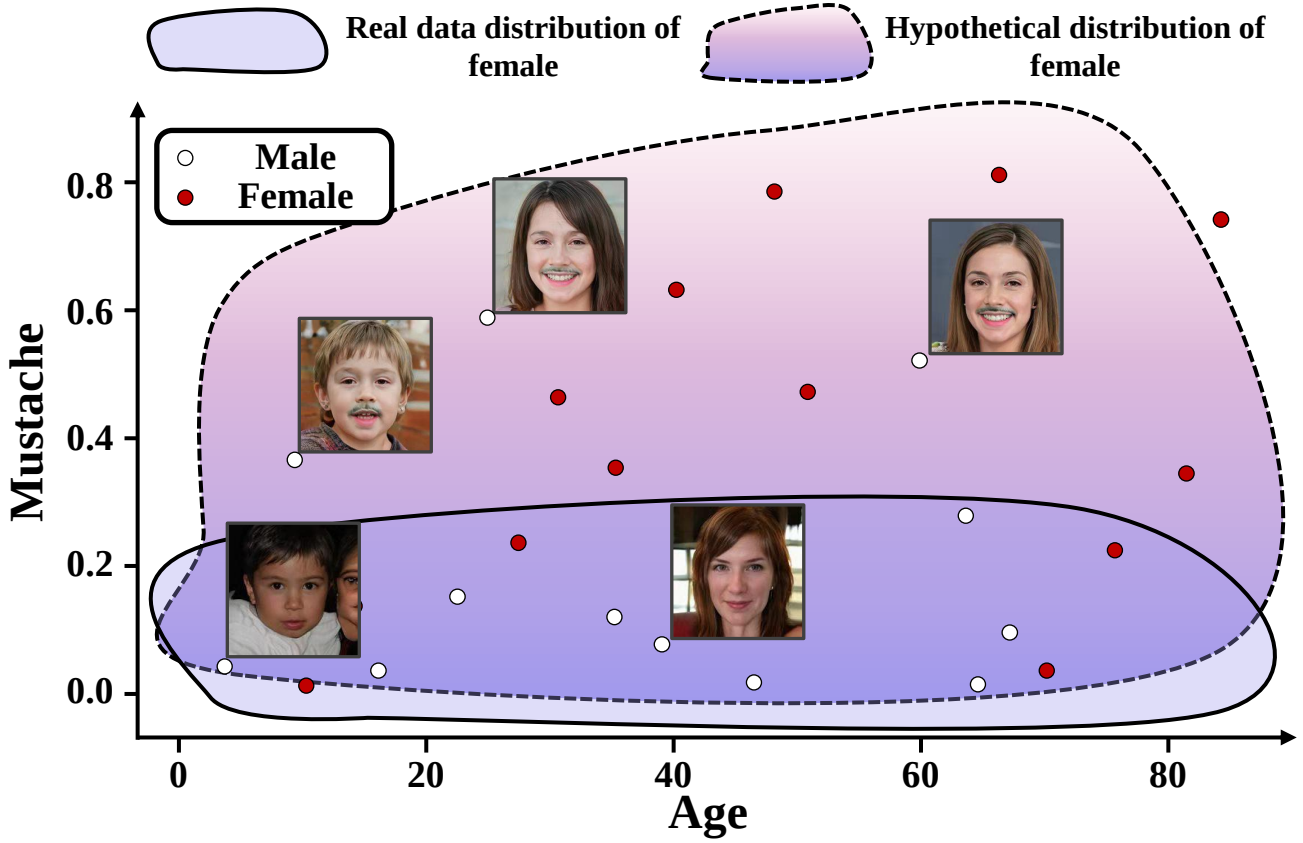
What is Counterfactual Synthesis

- Counterfactual Synthesis desirably extrapolates to those **counterfactual** cases, but keeps the other aspects unchanged.



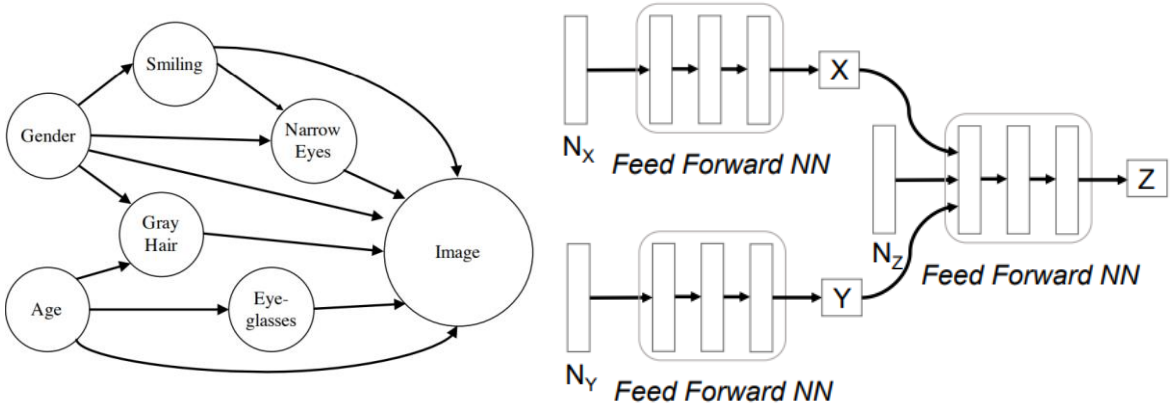
What is Counterfactual Synthesis

- Counterfactual Synthesis desirably extrapolates to those **counterfactual** cases, but keeps the other aspects unchanged.

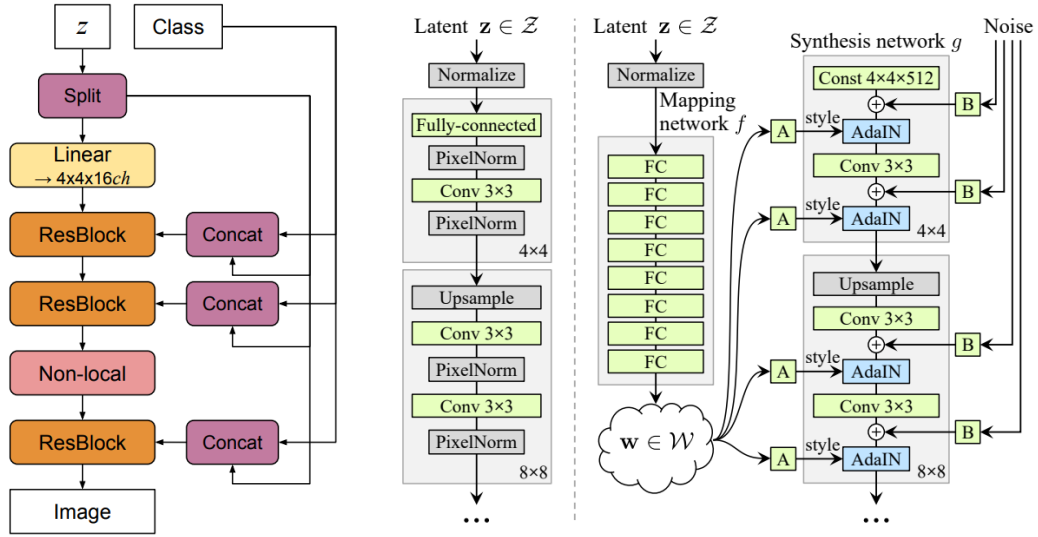


Structural Causal Model for Counterfactual Synthesis

- Prior SCM to identify causalities
  - many potential factors, obscure entanglements, strong ignorability



- SCM cannot easily leverage state-of-the-art generative models (i.e., BigGAN, StyleGAN)



Kocaoglu, M. and et, al. CausalGAN: Learning causal implicit generative models with adversarial training. In International Conference on Learning Representations, 2018.
Karras, T. and et, al. 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.
Andrew Brock and et, al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. . In International Conference on Learning Representations, 2019.

Principled GAN Knowledge Extrapolation

- Given:
  - data domain: $\mathcal{X}$
  - data distribution: $\mathbb{P}_{\mathcal{X}}$
- Assume:
  - pretrained generator: $G_{\theta}{}_{\mathcal{X}}: \mathcal{Z} \to \mathcal{X}$
  - posterior distribution for a knowledge l: $\mathbb{P}_l(x) = \mathbb{P}(l|x)$
- Task: infer $\mathbb{P}_H$, with only difference from $\mathbb{P}_{\mathcal{X}}$ in knowledge l

Principled GAN Knowledge Extrapolation

- Solution: by adversarial training:

$$\min_{G_\theta} \max_{D_\phi} \mathbb{E}_{x\sim\mathbb{P}_H}\left[\log D_\phi(x)\right] + \mathbb{E}_{x\sim\mathbb{P}_{G_\theta}}\left[\log\left(1 - D_\phi(x)\right)\right] \quad (1)$$

  - Goodfellow et al. have shown that Eq. (1) converges to $\mathbb{P}_{G_\theta} = \mathbb{P}_H$!
  - Problem: we do not have samples from $\mathbb{P}_H$, how to train discriminator?
  - If we know the optimal discriminator, then we only need to solve

$$\min_{G_\theta} \mathbb{E}_{x\sim\mathbb{P}_{G_\theta}}[\log(1 - D_{\phi^*}(x))] \quad (2)$$

    which do not need samples from $\mathbb{P}_H$!

Indiscernibility Space

- **Indistinguishable assumption**

  $\mathbb{P}_{\mathcal{X}}$ *is indistinguishable from* $\mathbb{P}_H$ *except for the altered knowledge* l

- **Definition**

  the indiscernibility space $\mathcal{I}^l$ *for knowledge* l*:*
  $\mathcal{I}^l = \{\theta : \mathbb{P}_{G_\theta} is\ indistinguishable\ from\ \mathbb{P}_H\ except\ for\ knowledge\ l\}$

- By Indiscernibility Space, solving (1) is equivalent to solving

$$\min_{G_\theta \in \mathcal{I}^l} \max_{D_\phi} \mathbb{E}_{x \sim \mathbb{P}_H}\left[\log D_\phi(x)\right] + \mathbb{E}_{x \sim \mathbb{P}_{G_\theta}}\left[\log\left(1 - D_\phi(x)\right)\right]$$

Optimal Discriminator

Theorem. *If* $G_\theta \in \mathcal{I}^l$, *then the optimal discriminator of problem* $\max_{D_\phi} V(D_\phi, G_\theta)$ *is* $D_{\phi^*}(x) = \mathbb{P}_l(x)$ *for some probability distribution* $\mathbb{P}_l$ *of knowledge l.*

Problem (2) is equivalent to

$$\min_{\theta \in \mathcal{I}^l} \mathbb{E}_{x \sim \mathbb{P}_{G_\theta}} [\log \mathbb{P}_{\bar{l}}(x)] = -H\left(\mathbb{P}_{G_\theta}, \mathbb{P}_{\bar{l}}\right), \qquad (3)$$

where $\mathbb{P}_{\bar{l}} = 1 - \mathbb{P}_l$, H is the cross entropy function.

Solution to Generator

Algorithm: repeatedly update θ as

$$\theta^{k+1} = \epsilon \mathrm{Tr}\left(\nabla_\theta H\left(\mathbb{P}_{G_{\theta^k}}, \mathbb{P}_{\bar{l}}\right), \lambda\right),$$

where $\mathrm{Tr}(v, \lambda) = \begin{cases} v_i = 0, & if\ |\nabla_\theta H_i| \leq \lambda, \\ v_i = 1, & if\ \nabla_\theta H_i < 0, 0 \leq \lambda < |\nabla_\theta H_i|, \\ v_i = -1, & if\ \nabla_\theta H_i > 0, 0 \leq \lambda < |\nabla_\theta H_i|, \end{cases}$

and $\theta^0 = \theta$, then G_{θ^K} generates a distribution that approximates the counterfactual hypothesis $\mathbb{P}_H$!

Principal Knowledge Descent

- **Theorem.** *Let Δ be the descent value of the objective (3) by implementing Algorithm 1, and δ be the change of the other knowledge, i.e.,*

$$\Delta = H\left(\mathbb{P}_{G_{\theta K},}\, \mathbb{P}_{\bar{l}}\right) - H\left(\mathbb{P}_{G_{\theta},}\, \mathbb{P}_{\bar{l}}\right),$$
$$\delta = \mathbb{E}_{x\sim\mathbb{P}_X}\left[\left\| f^r_{G_{\theta}K}(x) - f^r_X(x)\right\|\right],$$

where K is the iteration turns. Assume that $\mathbb{P}_{G_{\theta}x} = \mathbb{P}_X$, ε is small enough, and $L = \sup_{\left\|\theta-\theta^x\right\|_\infty < K\varepsilon} \left\|\nabla_\theta f^r_{G_\theta}\right\|_\infty$. There is $\lambda_{max} < 0$ such that $\forall\lambda \in (0, \lambda_{max})$, we have $\Delta > 0$ and

$$\frac{\Delta}{\delta} \geq \frac{\lambda}{L} + o(1).$$

Findings and Results

- Counterfactual Synthesis

- Data domain: ImageNet
- Pretrained generator: BigGAN256
- Posterior distribution: ResNet50

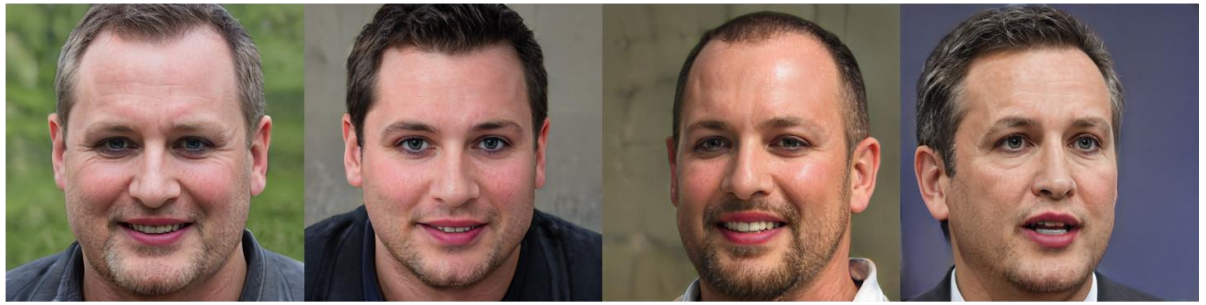
Findings and Results

- Counterfactual Synthesis

Domain: Female **Counterfact: Mustache**

Domain: Child **Counterfact: Gray Hair**

Domain: Child **Counterfact: Mustache**



Domain: Male **Counterfact: Lipstick**

- Data domain: FFHQ
- Pretrained generator: StyleGAN2
- Posterior distribution: ResNet50 trained on CelebA-HQ