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]

Abstract

Human can extrapolate well, generalize daily knowledge into unseen scenarios, raise and answer counterfactual questions. To imitate this ability via generative models, previous works have extensively studied explicitly encoding Structural Causal Models (SCMs) into architectures of generator networks. This methodology, however, limits the flexibility of the generator as they must be carefully crafted to follow the causal graph, and demands a ground truth SCM with strong ignorability assumption as prior, which is a nontrivial assumption in many real scenarios. Thus, many current causal GAN methods fail to generate high fidelity counterfactual results as they cannot easily leverage state-of-the-art generative models. In this paper, we propose to study counterfactual synthesis from a new perspective of knowledge extrapolation, where a given knowledge dimension of the data distribution is extrapolated, but the remaining knowledge is kept indistinguishable from the original distribution. We show that an adversarial game with a closed-form discriminator can be used to address the knowledge extrapolation problem, and a novel principal knowledge descent method can efficiently estimate the extrapolated distribution through the adversarial game. Our method enjoys both elegant theoretical guarantees and superior performance in many scenarios.

1. Introduction

Human beings exhibit remarkable ability of cognitive extrapolation (Ehrlich, 2005; Beck et al., 2006) in a variety of aspects. For example, we can accurately extrapolate the motion of objects (Ehrlich, 2005), imagine unseen objects (Kocaoglu et al., 2018), raise and answer counterfactual questions (Beck et al., 2006). There is a temptation to wonder whether Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) can generalize well to examples whose distribution is arbitrarily far from that of the given training data (or even counterfactual examples), as shown in Fig. 1. Specifically, this knowledge extrapolation ability of GANs can be reflected as synthesizing out-of-distribution examples and manipulating semantic features to constitute counterfactual combinations.

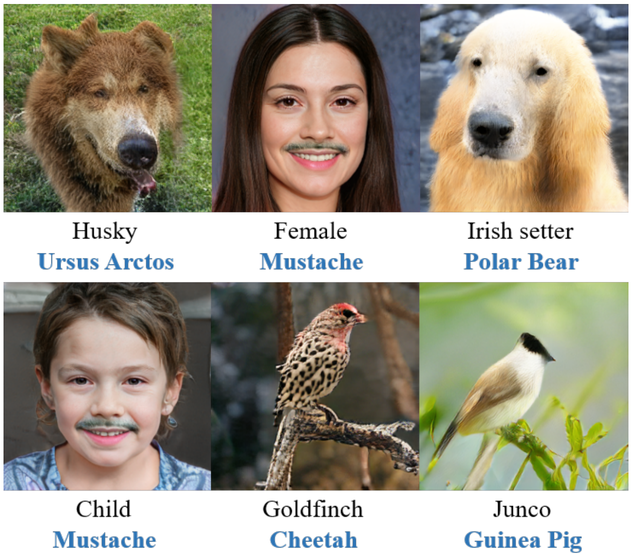


Figure 1. Knowledge extrapolation. All the above objects are counterfactual, rarely existent in real world. The origin domains of them are written benzene, and the extrapolated knowledge is marked in **purple**.

Counterfactual synthesis (Kocaoglu et al., 2018; Sauer & Geiger, 2020; Nemirovsky et al., 2020; Yang et al., 2021; Averitt et al., 2020; Thiagarajan et al., 2021) is one of the most promising tasks to achieve the general goal of knowledge extrapolation in GANs. For counterfactual synthesis, if brusquely ignoring differences in details, most existing methods follow the framework of the same fashion — directly modeling a Structural Causal Model (SCM) (Pearl, 2009a) in well-designed architectures of generator networks. Factors of interest in the causal graph are designed as labels to control the synthesis of the generator. However, this type of approaches is somewhat inflexible. Specifically, it demands a prior SCM to identify all the causalities in the data

[1] University of Science and Technology of China, Hefei, China [2] Ant Research, Hangzhou, China [3] Alibaba Group, Hangzhou, China. Correspondence to: Zheng-Jun Zha <zhazj@ustc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

generation process. Theoretically, constructing a prior SCM with photo-realistic synthesis effect can be embarrassingly difficult if many potential factors and obscure entanglements are involved (Pearl, 2009a; Sekhon, 2008; Holland, 1986). In addition, the rigorous inference of causal effects needs the strong assumptions like 'strong ignorability' (*i.e.*, assuming no unobserved confounders) (Sekhon, 2008; Holland, 1986), which is hard to verify in practice (Pearl, 2009b;a). Therefore, these methods are limited in many knowledge extrapolation scenarios because it is inconvenient to deduce a prior SCM. In addition to the demand of prior SCMs, these approaches also put constraints on the network design that has to be consistent with the prior SCM. Concretely, they need to construct generators that each network component corresponds to a causal graph factor to yield causal interventions (Pearl, 2009a; Holland, 1986). While new GAN architectures adapt rapidly, it is of great challenge to directly apply those methods to state-of-the-art GANs (*e.g.*, StyleGANs (Karras et al., 2019; 2020b) or BigGAN (Brock et al., 2018)), thus limiting their overall performance regarding high-fidelity counterfactual synthesis.

In this paper, we propose a principled knowledge extrapolation method to circumvent the two embarrassing flaws of current methods, and provide a new viewpoint for knowledge extrapolation of GANs. Instead of modeling prior causal graphs in generators, here we turn to a simple hypothesis that the original distribution and extrapolated distribution are indistinguishable, except for the extrapolated knowledge dimension. For example, if there is a hypothetical distribution in which all people (including women and children) could wear beards, then the information to distinguish this hypothetical distribution from real world data only exists in the distribution of beards. Thus the beard-irrelevant information (*i.e.*, gender, age or other knowledge dimensions) cannot contribute to distinguish these two distributions, as the remaining knowledge is the same between these two distributions. Under this assumption, we prove that an adversarial learning strategy (Goodfellow et al., 2014) can approximate the hypothetical distribution with a closed-form discriminator when we manage to preserve the irrelevant knowledge unchanged during the adversarial game. To achieve this goal, a novel Principal Knowledge Descent (PKD) method is proposed to solve a sparse paradigm in the parameter space that starts from the pretrained generator distribution to an approximation of the hypothetical distribution. The sparse paradigm will involve only the most related parameters to our knowledge of interest, thus posing negligible influence to the other knowledge.

In conclusion, the contributions of this paper include:

- We propose a new principled GAN knowledge extrapolation method that is flexible to use and can be easily adapted to state-of-the-art GAN architectures;
- We design a novel sparse descent strategy to efficiently estimate the extrapolated distribution based on our theory;
- The proposed method is the first to successfully synthesize high-fidelity counterfactual results in various image data domains.

2. Related Work

The prevalence of GAN (Goodfellow et al., 2014) has aroused researchers' ambitions to utilize various novel GANs to synthesize counterfactual data under the guidance of prior structural causal models. CausalGAN (Kocaoglu et al., 2018) proposes to learn a causal implicit model through adversarial training with a given causal graph for facial attribute disentanglement. CounterGAN (Nemirovsky et al., 2020) employs a residual generator to improve counterfactual realism and actionability compared to regular GANs. Counterfactual Generative Network (CGN) (Sauer & Geiger, 2020) suggests to decouple the ImageNet generation into four aspects of the shape, texture, background, and composer. CGN explicitly models the causality in the four aspects to yield counterfactual combinations of them, like triumphal arch with the elephant texture. CausalVAE (Yang et al., 2021) employs structural causal layer to encode prior causalities. Thiagarajan et al. exploits deep image priors from a U-Net (Ronneberger et al., 2015) and a classification model to synthesize counterfactual images. Those methods provide compelling insights to the causal explanation of black-box generative models, but put extra limitations on generator architectures, thus generally yielding much less plausible synthesis than state-of-the-art GAN models. Also, these methods rely on prior causal models, which grossly limit their generalization to other GAN architectures.

3. Method

Given a data domain $\mathcal{X}$ and a data distribution $\mathbb{P}_{\mathcal{X}}$, we assume that there is already a pretrained generator network $\boldsymbol{G}_{\boldsymbol{\theta}^{\mathcal{X}}} : \mathcal{Z} \to \mathcal{X}$ that captures the data distribution, and a posterior probability $\mathbb{P}_l(\boldsymbol{x}) = \mathbb{P}(l|\boldsymbol{x})$ [1] for a knowledge of interest l. The pretrained generator network transports the prior distribution $\mathbb{P}_{\mathcal{Z}}$ (which is usually the standard Gaussian) on the latent space $\mathcal{Z}$ into the data distribution $\mathbb{P}_{\mathcal{X}}$, *i.e.*, $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^{\mathcal{X}}}} = \mathbb{P}_{\mathcal{X}}$ (Goodfellow et al., 2014), with $\boldsymbol{\theta}^{\mathcal{X}}$ being its parameters at convergence. The generator can be obtained from a pretrained GAN (Goodfellow et al., 2014), VAE (Kingma & Welling, 2013), or other smooth parametric methods that yield generative components (Kingma & Dhariwal, 2018; Dinh et al., 2016). The posterior probability $\mathbb{P}_l$ can be obtained through classification or regression neural networks on knowledge l, or other smooth parametric

[1] This paper uses $\mathbb{P}$ to denote the probability density.

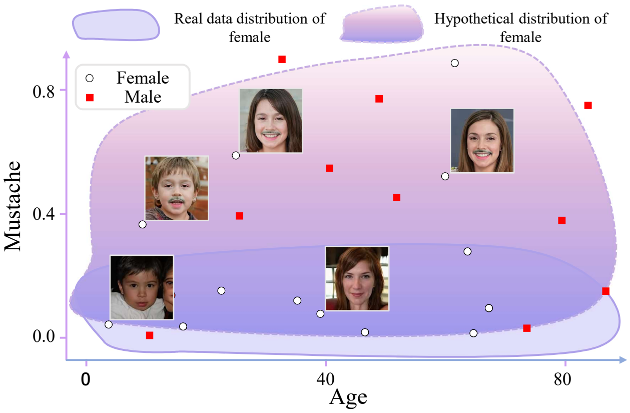


Figure 2. Illustration of the hypothetical distribution. Real data distribution excludes the cases of children or female in mustache. The hypothetical distribution extrapolates to those counterfactual cases, but keeps the other aspects unchanged, especially the confounder factors (factors that influence both the dependent variable (face image) and independent variable (mustache), causing a spurious association) like gender or age of the faces.

methods that yield posterior estimation of l. For example, a most common case of counterfactual synthesis is a GAN that generates facial images, together with classifiers that identify the posteriors of semantic attributes such as 'mustache', 'age', 'gender', etc. (Kocaoglu et al., 2018).

Our task here is to infer a hypothetical distribution $\mathbb{P}_H$, where it only differs from the real data in the knowledge of interest, and is indistinguishable from the real data distribution $\mathbb{P}_\mathcal{X}$ among all the remaining knowledge. To capture this hypothetical distribution, we want to get the parametric value $\boldsymbol{\theta}^H$ such that $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^H}} = \mathbb{P}_H$. Fig. 2 illustrates the example of 'mustache'. If the real data are facial images and the knowledge of interest is 'mustache', then we want to obtain a generator that is capable of synthesizing images where all people, including women and children, can wear the 'mustache' while all the other semantic knowledge is maintained realistic and plausible. The hypothetical distribution extrapolates the real data distribution along the dimension of knowledge of interest and is counterfactual in the real world. Here the key difference between our method and previous methods is that we bypass SCM to directly yield counterfactual synthesis results with given pretrained generators, and eliminate the limitation to generator architectures so that we can directly apply our method to any given generative models.

Here we propose a different perspective to solve the hypothetical distribution $\mathbb{P}_H$ from adversarial training with GANs. Theoretically, the adversarial training of GAN models will terminate at the global optimum of the generated distribution equal to the data distribution (Goodfellow et al., 2014). Thus, if we aim to alter the data distribution to the hypothetical distribution, we may formulate an adversarial game that will halt at the generated distribution equal to the hypothetical distribution. So we propose to solve

$$\begin{aligned}\min_{\boldsymbol{G}_{\boldsymbol{\theta}}} \max_{\boldsymbol{D}_{\boldsymbol{\phi}}} V(\boldsymbol{D}_{\boldsymbol{\phi}}, \boldsymbol{G}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_H}\left[\log(\boldsymbol{D}_{\boldsymbol{\phi}}(\boldsymbol{x}))\right] \\ + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}}\left[\log(1-\boldsymbol{D}_{\boldsymbol{\phi}}(\boldsymbol{x}))\right],\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are parameters, and $\boldsymbol{D}_{\boldsymbol{\phi}}$ is the discriminator network. The problem here is that $\mathbb{P}_H$ is merely hypothetical, meaning that we do not have any sample from it at hand, so evaluating the value of the term $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_H}\left[\log(\boldsymbol{D}_{\boldsymbol{\phi}}(\boldsymbol{x}))\right]$ of $V(\boldsymbol{D}_{\boldsymbol{\phi}}, \boldsymbol{G}_{\boldsymbol{\theta}})$, or gradients of it seems to be impossible with typical Monte Carlo methods (Hammersley, 2013). In the next section, a novel adversarial extrapolation in the knowledge dimension will be deduced to address this problem.

3.1. Adversarial Extrapolation in Indiscernibility Space

The typical algorithm to solve Problem (1) is to alternately optimize the generator and discriminator to attain a Nash equilibrium (Goodfellow et al., 2014). For the hypothetical distribution, however, we are allowed to simplify the training considerably with an indistinguishable assumption.

Assumption 3.1 (Indistinguishable assumption)**.** The real data distribution $\mathbb{P}_\mathcal{X}$ is indistinguishable from the hypothetical distribution $\mathbb{P}_H$ except for the altered knowledge l.

Definition 3.2 (Indiscernibility Space)**.** We denote the collection of all the parameters that can induce generated distribution satisfying Assumption 3.1 as the indiscernibility space $\mathcal{I}^l$ of knowledge l, *i.e.*, $\mathcal{I}^l = \{\boldsymbol{\theta} : \mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$ *is indistinguishable from the hypothetical distribution* $\mathbb{P}_H$ *except for the knowledge of interest* $l\}$.

Apparently, we have $\boldsymbol{\theta}^\mathcal{X}, \boldsymbol{\theta}^H \in \mathcal{I}^l$, and $\boldsymbol{\theta}^H$ is the global optimum of Problem (1). Thus, solving Problem (1) is equivalent to solving it inside the indiscernibility space $\mathcal{I}^l$:

$$\begin{aligned}\min_{\boldsymbol{G}_{\boldsymbol{\theta}}\in\mathcal{I}^l} \max_{\boldsymbol{D}_{\boldsymbol{\phi}}} V(\boldsymbol{D}_{\boldsymbol{\phi}}, \boldsymbol{G}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_H}\left[\log(\boldsymbol{D}_{\boldsymbol{\phi}}(\boldsymbol{x}))\right] \\ + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}}\left[\log(1-\boldsymbol{D}_{\boldsymbol{\phi}}(\boldsymbol{x}))\right].\end{aligned} \tag{2}$$

Investigate Assumption 3.1 from knowledge l. We now consider to solve the the optimal discriminator for generators inside the Indiscernibility Space of Problem (2). Assume $\boldsymbol{\theta} \in \mathcal{I}^l$. Distinguishing which distribution between $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$ and $\mathbb{P}_H$ a sample $\boldsymbol{x}$ is more likely to be sampled from can leverage the ratio $\frac{\mathbb{P}_H(\boldsymbol{x})}{\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{x})}$. If this ratio is larger than one, then $\boldsymbol{x}$ is more likely from $\mathbb{P}_H$, otherwise from $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$. As $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$ is indistinguishable from $\mathbb{P}_H$ except for l, this ratio should be purely decided by the posterior distribution of knowledge l on sample $\boldsymbol{x}$. Namely, there is some posterior distribution $\mathbb{P}_l(\boldsymbol{x}) = \mathbb{P}(l|\boldsymbol{x})$ of knowledge l, such that

$$\frac{\mathbb{P}_H(\boldsymbol{x})}{\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{x})} = \frac{\mathbb{P}_l(\boldsymbol{x})}{1-\mathbb{P}_l(\boldsymbol{x})}. \tag{3}$$

Remark 3.3. Eq. (3) means that distinguishing $\boldsymbol{x}$ amounts to distinguishing knowledge l. When sampling from $\mathbb{P}_H$, knowledge l is more likely to appear. While sampling from $\mathbb{P}_{\boldsymbol{G_\theta}}$, knowledge l is more likely to be ignored. Thus, $\mathbb{P}_H$ also extrapolates knowledge l beyond the original distribution of $\mathbb{P}_{\boldsymbol{G_\theta}}$.

An essential property of the indiscernibility space then immediately follows, that when constraining the generator in the indiscernibility space, the optimal discriminator admits a closed-form solution. Rigorously, we have the following:

Theorem 3.4. *If $\boldsymbol{G_\theta} \in \mathcal{I}^l$, then the optimal discriminator of problem $\max_{\boldsymbol{D_\phi}} V(\boldsymbol{D_\phi}, \boldsymbol{G_\theta})$ is $\boldsymbol{D_{\phi^*}}(\boldsymbol{x}) = \mathbb{P}_l(\boldsymbol{x})$ for some probability distribution $\mathbb{P}_l$ of knowledge l. Thus, Problem (2) is equivalent to*

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathcal{I}^l} V(\boldsymbol{D_{\phi^*}}, \boldsymbol{G_\theta}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_H}\left[\log(\mathbb{P}_l(\boldsymbol{x}))\right] \\ &+ \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G_\theta}}}\left[\log(1-\mathbb{P}_l(\boldsymbol{x}))\right].\end{aligned} \tag{4}$$

Let $\mathbb{P}_{\bar{l}} = 1 - \mathbb{P}_l$ denote the probability that l does not occur in sample $\boldsymbol{x}$. As $\boldsymbol{G_\theta}$ is not involved in $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_H}\left[\log(\mathbb{P}_l(\boldsymbol{x}))\right]$, we only need to solve

$$\min_{\boldsymbol{\theta}\in\mathcal{I}^l} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G_\theta}}}\left[\log(\mathbb{P}_{\bar{l}}(\boldsymbol{x}))\right] = -H(\mathbb{P}_{\boldsymbol{G_\theta}}, \mathbb{P}_{\bar{l}}), \tag{5}$$

where H is the cross entropy function (Shore & Johnson, 1981; De Boer et al., 2005; Murphy, 2012). Solving this problem means to maximize the information of knowledge l while keeping the generated distribution indistinguishable from the real data on the remaining parts. Intuitively, this procedure just points to the desired hypothetical distribution as is claimed by Theorem 3.4.

Investigate Assumption 3.1 from the remaining knowledge. Here the main difficulty is to handle the constraint $\boldsymbol{\theta} \in \mathcal{I}^l$ in Eq. (5). To this end, we review the Indistinguishable Assumption 3.1 from the perspective of the remaining knowledge. An equivalent statement to "$\mathbb{P}_{\boldsymbol{G_\theta}}$ and $\mathbb{P}_H$ are indistinguishable except for the knowledge of interest l" is that, the remaining knowledge r is not changed from $\mathbb{P}_{\boldsymbol{G_\theta}}$ to $\mathbb{P}_H$, and all the changes occur for the knowledge of interest l. Let $f^r_{\boldsymbol{G_\theta}}(\boldsymbol{x}), f^r_H(\boldsymbol{x}), f^r_{\mathcal{X}}(\boldsymbol{x})$ be the remaining knowledge on sample $\boldsymbol{x}$ of distribution $\mathbb{P}_{\boldsymbol{G_\theta}}, \mathbb{P}_H, \mathbb{P}_{\mathcal{X}}$, respectively. We should have an equivalent definition to the indiscernibility space

$$\mathcal{I}^l = \{\boldsymbol{\theta} : \forall \boldsymbol{x} \in \mathcal{X}, f^r_{\boldsymbol{G_\theta}}(\boldsymbol{x}) = f^r_H(\boldsymbol{x}) = f^r_{\mathcal{X}}(\boldsymbol{x})\}. \tag{6}$$

Remark 3.5. We omit the discussion of the exact form of the remaining knowledge f^r in order to bypass the demand of SCMs or strong ignorability that all confounders are known. Thus, our method does not rely on the exact factor factorization to the data distribution.

Thus, the final objective can be transformed into

$$\min_{f^r_{\boldsymbol{\theta}} = f^r_{\mathcal{X}}} -H(\mathbb{P}_{\boldsymbol{G_\theta}}, \mathbb{P}_{\bar{l}}). \tag{7}$$

Algorithm 1 Principal Knowledge Descent (PKD)

```
**Input:** Maximum number of iteration $K$, pretrained generator $\boldsymbol{G_{\theta^{\mathcal{X}}}}$ which captures the data distribution with $\mathbb{P}_{\boldsymbol{G_{\theta^{\mathcal{X}}}}} = \mathbb{P}_{\mathcal{X}}$, prior distribution $\mathcal{N}(\mathcal{O}, \mathcal{I})$ of the generator, posterior distribution $\mathbb{P}_l(\boldsymbol{x})$ for knowledge $l$, step size $\epsilon$, batch size $m$, and hyper-parameter $\lambda > 0$.
**Set:** $k = 0$ and $\boldsymbol{\theta}^k = \boldsymbol{\theta}^{\mathcal{X}}$.
**repeat**
  Randomly sample latent codes $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ from prior distribution $\mathcal{N}(\mathcal{O}, \mathcal{I})$.
  Compute $\vec{\boldsymbol{n}} = \nabla_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^m \log(1 - \mathbb{P}_l(\boldsymbol{G_{\theta^k}}(\boldsymbol{z}_i)))$;
  Compute $\boldsymbol{I} = (|\vec{\boldsymbol{n}}| > \lambda)_b, \boldsymbol{sgn} = (\vec{\boldsymbol{n}} > 0)_b - (\vec{\boldsymbol{n}} < 0)_b$, where $(\cdot)_b$ is element-wise Boolean operation;
  Update $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \epsilon \boldsymbol{sgn} * \boldsymbol{I}, k = k + 1$, where $*$ denotes element-wise multiplication;
**until** $k = K$.
**Output:** An extrapolated generator $\boldsymbol{G_{\theta^K}}$ that induces distribution $\mathbb{P}_{\boldsymbol{G^K_\theta}}$ to estimate the hypothetical distribution $\mathbb{P}_H$.
```

However, we still have the constraint $f^r_{\boldsymbol{\theta}} = f^r_{\mathcal{X}}$ unknown as we do not know the exact form of $f^r_{\boldsymbol{\theta}}$. Hopefully, the solution to this problem can be efficiently estimated. In the next section, we will study the associated numerical approximation.

Remark 3.6. If the constraint $\boldsymbol{\theta} \in \mathcal{I}^l$ is dropped, then the cross entropy achieves its optimal value if and only if $\mathbb{P}_{\boldsymbol{G_\theta}} = \mathbb{P}_{\bar{l}}$ (Murphy, 2012). This is a degenerate case that the generator may even not yield the valid synthesis, and lose all knowledge except the one of interest l. Thus, enforcing the optimization inside the indiscernibility space is a decisive condition for knowledge extrapolation.

3.2. Principal Knowledge Descent

In this section, we study how to numerically solve Problem (7). We show that its solution can be approximated through a series of principal knowledge descent with sparse and convex regularization.

Given the current parameter $\boldsymbol{\theta}$, here we want to find a direction $\Delta\boldsymbol{\theta}$ such that

- the knowledge of interest is altered accordingly, meaning that the cross entropy $-H(\mathbb{P}_{\boldsymbol{G_{\theta+\Delta\theta}}}, \mathbb{P}_{\bar{l}})$ is optimized;
- the other knowledge is unchanged, *i.e.*, $|f^r_{\boldsymbol{G_{\theta+\Delta\theta}}} - f^r_{\mathcal{X}}|$ is as small as possible.

Such a direction can preserve $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ to stay in $\mathcal{I}^l_{\mathcal{X}}$, and decline the value of the objective (5) if $\boldsymbol{\theta} \in \mathcal{I}^l_{\mathcal{X}}$. We call this direction as the principal knowledge descent direction. With this method, we can compute a path starting from $\boldsymbol{\theta}^{\mathcal{X}}$, and along this path, the other knowledge is kept intact, but the cross entropy criterion increases drastically. Then any point at this path corresponds to a certain degree of altering the knowledge of interest l. Now we discuss how to compute the principal knowledge descent direction.

Remaining Knowledge Penalty. Suppose $\boldsymbol{\theta} \in \mathcal{I}^l_{\mathcal{X}}$, then we have $f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}} = f^r_{\mathcal{X}}$ according to Eq. (6). Thus the change of the other knowledge under a small perturbation $\Delta\boldsymbol{\theta}$ can be written as

$$\begin{aligned} &|f^r_{\boldsymbol{G}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}} - f^r_{\mathcal{X}}| = |f^r_{\boldsymbol{G}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}} - f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}| \\ &\approx |(\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}})^T \Delta\boldsymbol{\theta}| \le \|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}\|_\infty \|\Delta\boldsymbol{\theta}\|_1 \\ &\le L\|\Delta\boldsymbol{\theta}\|_1, \end{aligned} \tag{8}$$

provided that $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$ is continuously differentiable, and L is an upper bound for $\|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}\|_\infty$ on the indiscernibility space. The penult inequality stems from the Cauchy inequality, which offers an estimation to the upper bound of the other knowledge alteration. Thus, if we constrain the ℓ_1 norm of $\Delta\boldsymbol{\theta}$, then we can constrain the overall change of the other knowledge.

Linear Principal Part. On the other hand, the descent value caused by updating $\boldsymbol{\theta}$ with $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ is

$$\begin{aligned} &- H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}}) + H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}}) \\ =& - \nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta} + o(\|\Delta\boldsymbol{\theta}\|_\infty). \end{aligned} \tag{9}$$

Thus, if we limit $\|\Delta\boldsymbol{\theta}\|_\infty \le \epsilon \ll 1$, then maximizing the descent value $-H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}}) + H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}})$ is equivalent to minimizing the linear principal part $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}$.

One Step Objective. In conclusion, we propose to get the descent direction of a single step from combining the remaining knowledge penalty and the linear principal part, *i.e.*

$$\min_{-\epsilon \preceq \Delta\boldsymbol{\theta} \preceq \epsilon} \nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta} + \lambda\|\Delta\boldsymbol{\theta}\|_1, \tag{10}$$

where $0 < \epsilon \ll 1$ is the step size, $\lambda = L\lambda_0 > 0$ containing two factors of an estimation L to the upper bound of $\|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}\|_\infty$ and a hyper-parameter λ_0 to adjust the weight of the regularization term $\|\Delta\boldsymbol{\theta}\|_1$, and $-\epsilon \preceq \Delta\boldsymbol{\theta} \preceq \epsilon$ means that each entry of $\Delta\boldsymbol{\theta}$ lies in $[-\epsilon, \epsilon]$. The first term $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}$ in (10) maximizes the descent value $(H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}}) - H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}}))$ induced by $\Delta\boldsymbol{\theta}$, moving $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}$ closer to $\mathbb{P}_{\bar{l}}$ as it is the cross entropy term. The second term $\lambda\|\Delta\boldsymbol{\theta}\|_1$ penalizes the overall change $\|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}\|_\infty \|\Delta\boldsymbol{\theta}\|_1$ induced to the other knowledge.

Sparsity of the Solution. The ℓ_1 penalty of $\|\Delta\boldsymbol{\theta}\|_1$ is well known to enforce sparsity to the elements of the solution (Santosa & Symes, 1986; Tibshirani, 1996). Namely, only a small number of elements of $\Delta\boldsymbol{\theta}$ will be non-zeros, meaning that the principal knowledge descent will only involve a fraction of parameters, while the most parameters of the generator will remain unchanged. The typical knowledge, like the expression, age, or gender of facial images, may only contain information of low dimensions (Penev & Sirovich, 2000; Härkönen et al., 2020; Shen et al., 2020). It will be obviously excessive for extracting a single knowledge domain in the whole parameter space. Accompanying with the excessively used parameters are two well-known challenges: over-fitting (Anderson & Burnham, 2004) and numerical instability (Hildebrand, 1987). The expressive capability of millions of parameters of the generator is powerful, thereby easily overfitting the bias of the classification or regression network we use as the optimal discriminator. Another issue is the numerical instability. As we have mentioned, the valid parameters relevant to a certain semantic feature should be sparse due to low-dimensional semantic information. Thus, updating all the parameters simultaneously may bring unexpected changes like entanglement of attributes. For example, changing 'age' leads to adding 'beard' in facial images.

Closed-Form Solution. An intriguing property pertaining to Problem (10) is that this convex optimization problem has a closed-form solution by virtue of strong duality and Karush–Kuhn–Tucker conditions (Boyd et al., 2004).

Theorem 3.7. *There is $\lambda_{max} > 0$ such that $\forall\lambda \in (0, \lambda_{max})$, Problem (10) admits a non-empty solution set. Specifically, a special solution can be attained by*

$$\begin{cases} \Delta\boldsymbol{\theta}_i = 0, & \text{if } |\nabla_{\boldsymbol{\theta}} H_i| \le \lambda, \\ \Delta\boldsymbol{\theta}_i = \epsilon, & \text{if } \nabla_{\boldsymbol{\theta}} H_i < 0, 0 \le \lambda < |\nabla_{\boldsymbol{\theta}} H_i|, \\ \Delta\boldsymbol{\theta}_i = -\epsilon, & \text{if } \nabla_{\boldsymbol{\theta}} H_i > 0, 0 \le \lambda < |\nabla_{\boldsymbol{\theta}} H_i|, \end{cases} \tag{11}$$

where $\Delta\boldsymbol{\theta}_i$ and $\nabla_{\boldsymbol{\theta}} H_i = [\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})]_i$ are the i-th element of $\Delta\boldsymbol{\theta}$ and $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})$, respectively.

Principal Knowledge Descent. We conclude our algorithm to solve Problem (5) in Algorithm 1. This algorithm manages to minimize the objective of Problem (5) while trying to maintain the remaining knowledge distribution. Rigorously, we have the following theorem.

Theorem 3.8. *Let Δ be the descent value of the objective (5) by implementing Algorithm 1, and δ be the change of the other knowledge, i.e.,*

$$\Delta = H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^K}}, \mathbb{P}_{\bar{l}}) - H(\mathbb{P}_{\boldsymbol{G}_{\mathcal{X}}}, \mathbb{P}_{\bar{l}}), \tag{12}$$

$$\delta = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\mathcal{X}}} \left[\left| f^r_{\boldsymbol{G}_{\boldsymbol{\theta}^K}}(\boldsymbol{x}) - f^r_{\mathcal{X}}(\boldsymbol{x}) \right| \right], \tag{13}$$

where K is the iteration turns. Assume that $\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^{\mathcal{X}}}} = \mathbb{P}_{\mathcal{X}}$, ϵ is small enough, and $L = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\mathcal{X}}\|_\infty \le K\epsilon} \|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}\|_\infty$. There is $\lambda_{max} > 0$ such that $\forall\lambda \in (0, \lambda_{max})$, we have $\Delta > 0$ and

$$\frac{\Delta}{\delta} \ge \frac{\lambda}{L} + o(1). \tag{14}$$

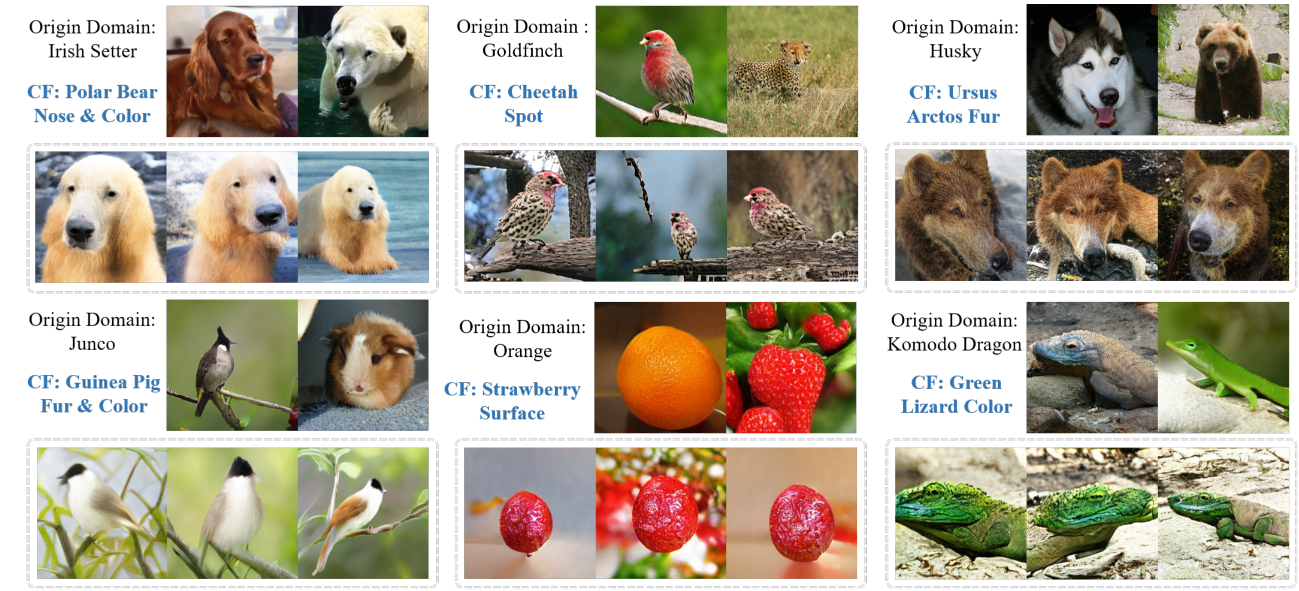


Figure 3. Knowledge extrapolation of BigGAN on ImageNet. Odd rows report the original data domain and extrapolated counterfactual knowledge (marked with **CF:** and **purple** color), while even rows report the counterfactual results generated by the proposed method. To the best of our knowledge, this is the first work that yields high-fidelity photo-realistic counterfactual synthesis of various image domains.

This theorem tells us that choosing a large λ can yield small variation of the remaining knowledge. Recall that $\lambda = \lambda_0 L$. This theorem also implies that λ_0 controls the principal ratio of Algorithm 1 — the ratio of change in the knowledge of interest and the other knowledge.

3.3. Dirac Knowledge Extrapolation

While our method is designed for distribution-wise extrapolation, we can also use it for single image extrapolation by setting the data distribution as the Dirac distribution (Arfken & Weber, 1999)

$$\mathbb{P}_{\mathcal{X}}(\boldsymbol{x}) = \delta_{\boldsymbol{x}_0}(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} = \boldsymbol{x}_0, \\ 0 & \boldsymbol{x} \neq \boldsymbol{x}_0. \end{cases} \tag{15}$$

We may then change the prior distribution $\mathcal{N}(\mathcal{O}, \mathcal{I})$ in Algorithm 1 with $\delta_{\boldsymbol{G}_{\boldsymbol{\theta}\mathcal{X}}^{-1}(\boldsymbol{x}_0)}(\boldsymbol{z})$ to implement the single image knowledge extrapolation. While the Dirac distribution is not smooth, directly implementing it may cause numerical instability. To enhance numerical stability, we will instead change the prior distribution $\mathcal{N}(\mathcal{O}, \mathcal{I})$ in Algorithm 1 with $\mathcal{N}(\boldsymbol{G}_{\boldsymbol{\theta}\mathcal{X}}^{-1}(\boldsymbol{x}) + \mathcal{O}, \xi\mathcal{I})$, where $0 < \xi \ll 1$ is a small number to approximate the Dirac distribution.

4. Findings and Results

In this section, we present several discoveries from our proposed Principal Knowledge Descent (PKD) method and extrapolation results of state-of-the-art GANs, including BigGAN256-Deep (Brock et al., 2018), StyleGAN2 (Karras et al., 2019; 2020b) on FFHQ faces (Karras et al., 2019), and StyleGAN2-ADA (Karras et al., 2020a) on BreCaHAD (Aksac et al., 2019) which contains breath cancer slices. The details of the experiments are reported in Appendix Sec. B, including choices of the hyper-parameters λ, ϵ, K, m, sources of all pretrained models (*i.e.*, generators and posterior estimation models), and dataset information.

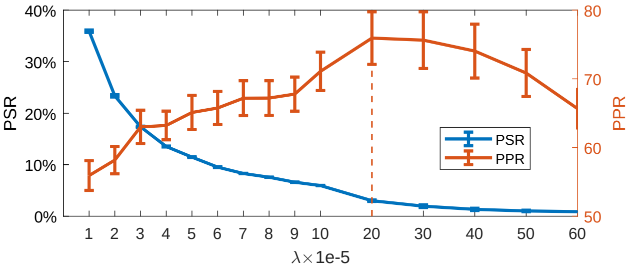


Figure 4. Pixel Principal Ratio (PPR) and Parameter Sparsity Ratio (PSR) of PKD under different λ. We find a turning point for the PPR metric, which also corresponds to the minimum volume of parameters to capture the knowledge of interest.

Metric	FID	IS	Path Length
Original GAN	**5.31**	**4.36**	185.59
PKD-Mustache	7.17	4.10	187.50
PKD-Lipstick	7.26	4.26	**175.20**
PKD-GrayHair	7.04	4.18	198.24

Table 1. Numerical metrics for the quality of extrapolated distributions of PKD method on the FFHQ domain. Basically, we find that the decline of synthesis quality is negligible.

λ-Sparsity For a given data domain and generator model, we find that the volume of parameters that are active to a knowledge of interest is an interesting property. To investigate it, we randomly sample 500 images in the StyleGAN2

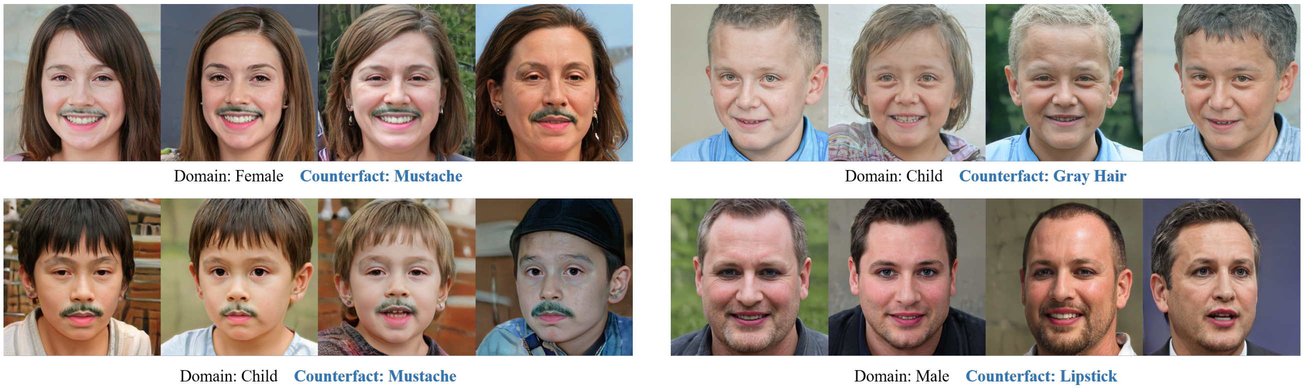

Domain: Female Counterfact: Mustache
Domain: Child Counterfact: Gray Hair
Domain: Child Counterfact: Mustache
Domain: Male Counterfact: Lipstick

Figure 5. Knowledge extrapolation of StyleGAN2 on FFHQ dataset. Odd rows report the original data domain and extrapolated counterfactual knowledge (marked in **purple**), while even rows report the counterfactual results generated by the proposed method.

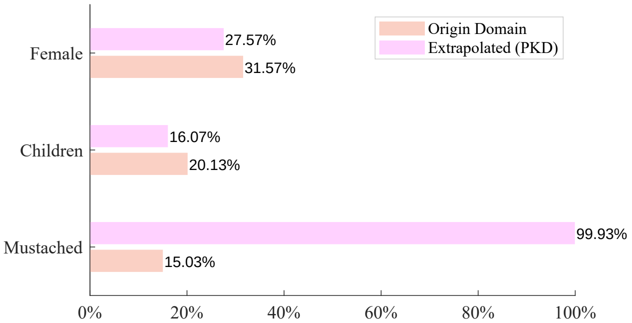


Figure 6. Statistics of the original distribution and extrapolated distribution after extrapolating 'mustache' on FFHQ faces. PKD increases the ratio of faces with 'mustache' from 15.03% to 99.93%, while only inducing less than 5% faces becoming 'male' or 'mature'. Considering the subtle case where 'mustache' alone can confuse age or gender, the consequence of entanglement is trivial. Thus, almost all faces in the extrapolated distribution are mustached, while the other statistics like the ratio of children and female are well preserved.

generator of FFHQ domain, and conduct Dirac Knowledge Extrapolation to those data points with different hyper-parameter λ. We report the results of two metrics: *Pixel Principal Ratio* (PPR) and *Parameter Sparsity Ratio* (PSR). PPR measures the ratio of changes induced by PKD between the cross entropy $H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}, \mathbb{P}_{\bar{l}})$ and the pixel value of the image, *i.e.*,

$$\mathrm{PPR} = \frac{|\log(\mathbb{P}_l(\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{z}))) - \log(\mathbb{P}_l(\boldsymbol{G}_{\boldsymbol{\theta}^{\mathcal{X}}}(\boldsymbol{z})))|}{\frac{1}{hwc}\|\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{z}) - \boldsymbol{G}_{\boldsymbol{\theta}^{\mathcal{X}}}(\boldsymbol{z})\|_2^2}, \quad (16)$$

where h, w, c are height, width, and channel numbers of the image domain. PSR measures the ratio of parameters that correspond to non-zero updating during PKD. The remaining parameters can be viewed as inactive to the extrapolated knowledge. We report the mean values and standard deviations of these two measurements in Fig. 4. As indicated by Theorem 3.8, PPR monotonically increases before λ_{max} of 2e-4. It suggests that before λ_{max}, increasing λ can further exclude redundant parameters in the PKD process. After λ_{max}, however, increasing λ will also hurt the extraction of the knowledge of interest. On the other hand, PSP monotonically decreases as λ increases, which is the consequence of sparsity enforced by the ℓ_1 penalty. Thus, the turning point of PPR should give the minimum ratio of parameters that are active to the knowledge of interest. Thus, in all the previous experiments, we set the hyper-parameter λ to a value that is slightly smaller than λ_{max} to secure better performance.

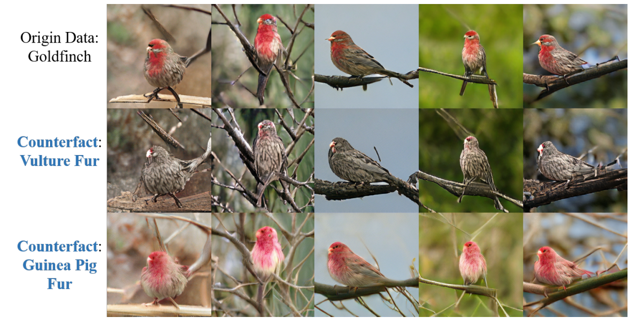


Figure 7. Dirac knowledge extrapolation to ImageNet animals. Top row shows the original data, middle and bottom rows show the extrapolated data, with extrapolated knowledge marked in **purple**.

PKD offers a new methodology for counterfactual synthesis. We report knowledge extrapolation results of our PKD method on ImageNet (Deng et al., 2009) data domain and FFHQ (Karras et al., 2019) face data domain in Fig. 3 and 5, respectively. In ImageNet domain, we use a pretrained BigGAN256-Deep model as the pretrained generator and ResNet50 (He et al., 2016) classifiers as the posterior distribution for knowledge of interest. We infer the results of counterfactual combination of knowledge among different ImageNet categories. As displayed in Fig. 3, we successfully synthesize non-existent species such as goldfinches with cheetah spot, huskies with ursus arctos fur, oranges with strawberry surface, *etc*. In FFHQ facial images domain, we use the StyleGAN2 model as the pretrained generator, and ResNet50 classifiers trained on CelebA-HQ (Karras et al., 2018) annotations for facial attributes like 'mustache', 'lipstick', 'gray hair' as the posterior distribution for knowl-

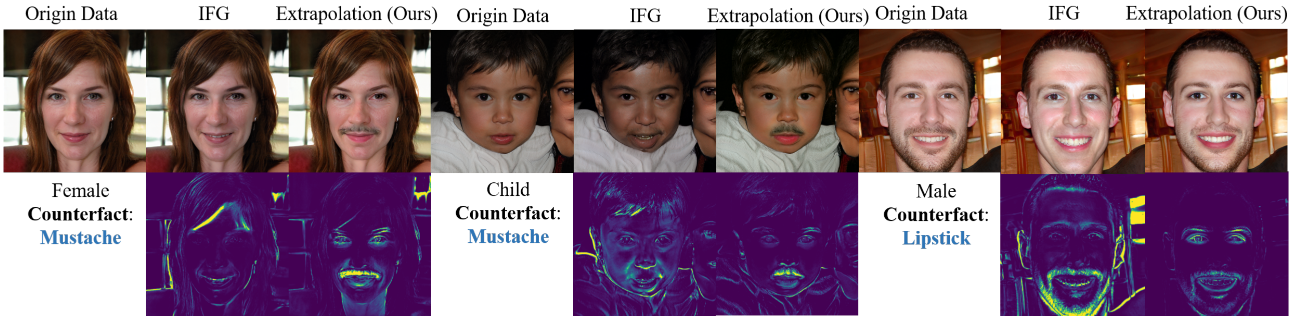


Figure 8. Counterfactual synthesis of latent image editing method (IFG) and our Dirac knowledge extrapolation method. Variations of image pixels are highlighted in the bottom row. Latent image editing method fails the counterfactual inference of women and children in mustache, removing mustache when extrapolating lipstick, while our method can easily handle these cases and induce changes more concentrated in the regions of interests.

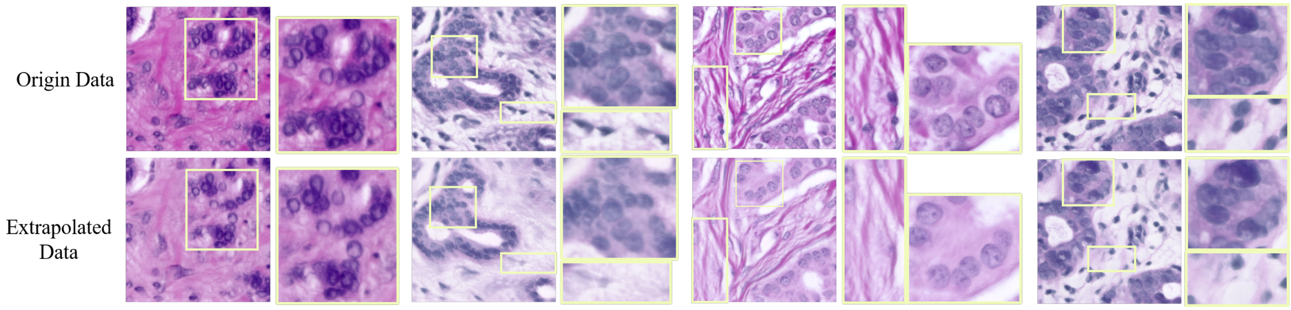


Figure 9. Few-shot Dirac knowledge extrapolation of tissue slices of breath cancer patients. The top row shows the original slices. The tissue slices are stained by H&E (Dapson & Horobin, 2009), containing bluish violet multi-core structures of the cancer nucleus and light red color of extracellular materials. Here we extrapolate each slice toward milder symptoms of cancer. The results are reported in the bottom row. In *most local regions* of the slices, the number of cancer nucleus decreases significantly, details are zoomed in with light green boxes. The whole training data to support our method is composed of merely *162* annotated slice images.

edge of interest. We infer the results of altering distributions of those facial attributes in Fig. 5, which shows counterfactual images such as women and children in 'mustache', men in 'lipstick', and children in 'gray hair'. As neural networks are disable to detect counterfactual results, we further conduct user study to confirm that whether the knowledge extrapolation globally succeeds. To do this, we randomly sample 3,000 images from the original generator distribution and the extrapolated distribution after extrapolating 'mustache' on FFHQ faces, respectively. We mix and randomly shuffle those images, and then ask testers three questions for each image: whether the person is 1) a child, 2) a female, and 3) mustached (Refer to Appendix for details of the user study). The result is reported in Fig. 6. We can find that PKD only significantly changes knowledge 'mustache' while maintaining the two *confounders* 'child' and 'female' nearly intact. Overall, the proposed PKD method can generate high-fidelity counterfactual results, although the exact causality relations are not pre-defined. Thus, counterfactual synthesis may not rely on the prior SCM to guide the generator. We demonstrate that PKD is also a competitive alternative for this task.

Dirac Knowledge Extrapolation is very powerful. We report the results of Dirac Knowledge Extrapolation of PKD in Fig. 7, 8, and 9. The Dirac Knowledge Extrapolation focuses on extrapolating knowledge of a given image rather than the whole distribution. For ImageNet and FFHQ data domains, the pretrained models and posterior distributions are selected as in previous knowledge extrapolation, and the results are reported in Fig. 7 and 8, respectively. Here we are interested in comparing the knowledge extrapolation with the recent latent image editing methods (Shen et al., 2020; Patashnik et al., 2021; Tewari et al., 2020), which edit facial attributes by attribute vectors in latent spaces of pretrained GANs. Despite the success in editing usual attributes, those methods are generally weak when conducting counterfactual synthesis. A major reason is that their synthesis will always lie in the pretrained generator distribution, as the parameters of the GAN model are not selected. The pretrained generator distribution can hardly yield counterfactual results. Specifically, we report the comparison with a baseline latent image editing method called InterFaceGAN (IFG) (Shen et al., 2020) in Fig. 8 (More comparisons are given in Appendix). We further conduct Dirac knowledge extrapolation in the BreCaHAD data domain. This dataset consists of 162 slice images and each of them has annotations for cancer nuclear. We train a ResNet50 classifier to infer the posterior probability for cancer severity of a tissue slice image on the few-shot annotation images, and use the StyleGAN2-ADA model trained on this dataset as the pretrained generator. We infer the results of reducing cancer severity in Fig. 9. Despite the few-shot training data, the PKD method can successfully infer the expected results. In all cases, the Dirac Knowledge extrapolation demonstrates compelling performance.

PKD is efficient and flexible when applied to SOTA models. The fidelity of counterfactual synthesis of PKD significantly surpasses previous causal GAN works. The major reason is its flexibility to combine with state-of-the-art generative models. Previous methods need to adapt generator architectures to be compatible with their prior SCMs or causal graphs, which is a non-trivial task. To the best of our knowledge, this is the first work that is capable of conducting counterfactual synthesis of photo-realistic effect without changing the generator architecture. Moreover, we find that the PKD method is very efficient when applying to SOTA models, *i.e.*, StyleGAN2 and BigGAN. We find that appealing counterfactual synthesis in most cases can be attained within 10-20 PKD iterations, and the damage to the synthesis fidelity caused by PKD is negligible. In Tab. 1, we report numerical metrics (*i.e.*, Fréchet Inception Distance (FID), Inception Score (IS), and Path Length as in (Karras et al., 2019)) of synthesis quality after knowledge extrapolation, indicating that the influence to synthesis quality is little.

5. Conclusion

This paper studies the problem of knowledge extrapolation of GANs, where the original generated distribution of a pretrained GAN is altered under the guidance of a novel Principal Knowledge Descent method to obtain counterfactual synthesis. Different from traditional methods that conduct counterfactual synthesis based on prior SCMs, this paper proposes to leverage a simple assumption that the extrapolated distribution and the original distribution are indistinguishable except for the knowledge of interest. Thus, our work gets rid of the usual demands of traditional methods to change the generator architecture to obey prior causalities. As a result, the proposed method is much more convenient to apply to SOTA generator models, and can yield much more photo-realistic counterfactual results.

Acknowledgement

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grant U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

References

Aksac, A., Demetrick, D. J., Ozyer, T., and Alhajj, R. Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC research notes*, 12(1):1–3, 2019.

Anderson, D. and Burnham, K. Model selection and multimodel inference. *Second. NY: Springer-Verlag*, 63(2020): 10, 2004.

Arfken, G. B. and Weber, H. J. Mathematical methods for physicists, 1999.

Averitt, A. J., Vanitchanant, N., Ranganath, R., and Perotte, A. J. The counterfactual χ-GAN: Finding comparable cohorts in observational health data. *Journal of Biomedical Informatics*, 109:103515, 2020.

Beck, S. R., Robinson, E. J., Carroll, D. J., and Apperly, I. A. Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, 77(2): 413–426, 2006.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Dapson, R. and Horobin, R. Dyes from a twenty-first century perspective. *Biotechnic & Histochemistry*, 84(4): 135–137, 2009.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

Ehrlich, R. The human brain's algorithm for extrapolating motion, and its possible gender-dependence. *Neuroscience letters*, 374(1):38–42, 2005.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Hammersley, J. *Monte Carlo methods*. Springer Science & Business Media, 2013.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable GAN controls. *arXiv preprint arXiv:2004.02546*, 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.

Hildebrand, F. B. *Introduction to numerical analysis*. Courier Corporation, 1987.

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020b.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

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

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.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Nemirovsky, D., Thiebaut, N., Xu, Y., and Gupta, A. CounterGAN: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*, 2020.

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.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a.

Pearl, J. *Causality*. Cambridge university press, 2009b.

Penev, P. S. and Sirovich, L. The global dimensionality of face space. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pp. 264–270. IEEE, 2000.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Santosa, F. and Symes, W. W. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

Sauer, A. and Geiger, A. Counterfactual generative networks. In *International Conference on Learning Representations*, 2020.

Sekhon, J. S. The Neyman-Rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

Shen, Y., Yang, C., Tang, X., and Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Shore, J. and Johnson, R. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, 27 (4):472–482, 1981.

Tewari, A., Elgharib, M., Bernard, F., Seidel, H.-P., Pérez, P., Zollhöfer, M., and Theobalt, C. PIE: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

Thiagarajan, J., Narayanaswamy, V. S., Rajan, D., Liang, J., Chaudhari, A., and Spanias, A. Designing counterfactual generators using deep model inversion. *Advances in Neural Information Processing Systems*, 34, 2021.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Wu, Z., Lischinski, D., and Shechtman, E. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

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.

A. Proof to Theorems

A.1. Theorem 3.4

Proof. It is well known that the optimal discriminator is (Goodfellow et al., 2014)

$$\boldsymbol{D}_{\boldsymbol{\phi}^*} = \frac{\mathbb{P}_H}{\mathbb{P}_H + \mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}}}. \tag{17}$$

When $\boldsymbol{\theta} \in \mathcal{I}^l$, we have some probability distribution $\mathbb{P}_l$ for knowledge l such that

$$\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}}} = \frac{1 - \mathbb{P}_l}{\mathbb{P}_l}\mathbb{P}_H. \tag{18}$$

Thus we have

$$\boldsymbol{D}_{\boldsymbol{\phi}^*} = \frac{\mathbb{P}_H}{\mathbb{P}_H + \frac{1-\mathbb{P}_l}{\mathbb{P}_l}\mathbb{P}_H} = \mathbb{P}_l. \tag{19}$$

□

A.2. Theorem 3.7

Proof. Let $\boldsymbol{x} = \Delta\boldsymbol{\theta}$, $\boldsymbol{V} = \nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}_0}}, \mathbb{P}_{\bar{l}})$, N be the volume of parameters of the generator, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_N)^T$ be the Lagrangian multiplier vectors, and $\mathbf{1} = (1, ..., 1)^T$ be the vector of all ones. We write the Lagrangian dual function of Problem (10)

$$\begin{aligned} g(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \inf_{\boldsymbol{x}} \boldsymbol{V}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1 + \sum_{i=1}^{N} \boldsymbol{\beta}_i(\boldsymbol{x}_i - \epsilon) + \sum_{i=1}^{N} \boldsymbol{\gamma}_i(-\epsilon - \boldsymbol{x}_i) && (20) \\ &= \inf_{\boldsymbol{x}} (\boldsymbol{V} + \boldsymbol{\beta} - \boldsymbol{\gamma})^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1 - \epsilon(\boldsymbol{\beta} + \boldsymbol{\gamma})^T\mathbf{1}, && (21) \\ &\text{s.t. } \boldsymbol{\beta}_i \geq 0,\ \boldsymbol{\gamma}_i \geq 0,\ i = 1, \ldots, N. && (22) \end{aligned}$$

As $L(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is a convex function, its optimal value $\boldsymbol{x}^*$ is reached if and only if

$$0 \in \partial_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{V} + \boldsymbol{\beta} - \boldsymbol{\gamma} + \lambda\partial_{\boldsymbol{x}}\|\boldsymbol{x}^*\|_1, \tag{23}$$

where $\partial_{\boldsymbol{x}}$ denotes the sub-gradient operator of convex function with respect to $\boldsymbol{x}$. Considering

$$\partial_{\boldsymbol{x}}\|\boldsymbol{x}^*\|_1 = \{\boldsymbol{n} \in \mathbb{R}^N : \|\boldsymbol{n}\|_\infty \leq 1\}, \tag{24}$$

we have

$$0 \in \{\boldsymbol{V} + \boldsymbol{\beta} - \boldsymbol{\gamma} + \lambda\boldsymbol{n} : \|\boldsymbol{n}\|_\infty \leq 1\}. \tag{25}$$

Thus we get that

$$L(\boldsymbol{x}^*, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{cases} -\epsilon(\boldsymbol{\beta} + \boldsymbol{\gamma})^T\mathbf{1}, & \text{if } \|\boldsymbol{V} + \boldsymbol{\beta} - \boldsymbol{\gamma}\|_\infty \leq \lambda, \\ -\infty, & \text{otherwise.} \end{cases} \tag{26}$$

Then, the Lagrangian duality form of Problem (10) is

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} -\epsilon(\boldsymbol{\beta} + \boldsymbol{\gamma})^T\mathbf{1}, && (27) \\ &\text{s.t. } \boldsymbol{\beta}_i \geq 0,\ \boldsymbol{\gamma}_i \geq 0,\ |\boldsymbol{V}_i + \boldsymbol{\beta}_i - \boldsymbol{\gamma}_i| \leq \lambda. && (28) \end{aligned}$$

This problem can be easily solved by linear programming technique. A special solution $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ can be obtained by

$$\begin{cases} \boldsymbol{\beta}_i^* = \boldsymbol{\gamma}_i^* = 0, & \text{if } -\lambda < \boldsymbol{V}_i < \lambda, \\ \boldsymbol{\beta}_i^* = 0, \boldsymbol{\gamma}_i^* = \boldsymbol{V}_i - \lambda, & \text{if } \boldsymbol{V}_i \geq \lambda, \\ \boldsymbol{\beta}_i^* = -\lambda - \boldsymbol{V}_i, \boldsymbol{\gamma}_i^* = 0, & \text{if } \boldsymbol{V}_i \leq -\lambda, \end{cases} \tag{29}$$

for $i = 1, \ldots, N$. Recalling the Karush–Kuhn–Tucker conditions of convex optimization, we have that $\boldsymbol{x}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*$ satisfy the following conditions

$$\boldsymbol{\beta}_i^*(\boldsymbol{x}_i^* - \epsilon) = 0,\ \boldsymbol{\gamma}_i^*(-\epsilon - \boldsymbol{x}_i^*) = 0,\ i = 1, ..., N, \tag{30}$$

$$|\boldsymbol{V}_i + \boldsymbol{\beta}_i^* - \boldsymbol{\gamma}_i^*| < \lambda \Rightarrow \boldsymbol{x}_i^* = 0. \tag{31}$$

We then have

$$\begin{cases} \boldsymbol{x}_i^* - \epsilon = 0, & \text{if } \boldsymbol{\beta}_i^* \neq 0, \\ -\epsilon - \boldsymbol{x}_i^* = 0, & \text{if } \boldsymbol{\gamma}_i^* \neq 0, \\ \boldsymbol{x}_i^* = 0, & \text{if } \boldsymbol{\beta}_i^* = \boldsymbol{\gamma}_i^* = 0,\ |\boldsymbol{V}_i| < \lambda. \end{cases} \tag{32}$$

Combining Eq. (29), we then conclude the theorem

$$\begin{cases} \boldsymbol{x}_i = 0, & \text{if } |\boldsymbol{V}_i| \leq \lambda, \\ \boldsymbol{x}_i = \epsilon, & \text{if } \boldsymbol{V}_i < 0, 0 \leq \lambda < |\boldsymbol{V}_i|, \\ \boldsymbol{x}_i = -\epsilon, & \text{if } \boldsymbol{V}_i > 0, 0 \leq \lambda < |\boldsymbol{V}_i|. \end{cases} \tag{33}$$

□

A.3. Theorem 3.8

Proof. We first consider one step of Principal Knowledge Descent (PKD). Assume that the current step is k and $k < K$. Then we define

$$\Delta_k = H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^{k+1}}}, \mathbb{P}_{\bar{l}}) - H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}}) = \nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}^k + o(\epsilon), \tag{34}$$

$$\delta_k = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\left|f^r_{\boldsymbol{G}_{\boldsymbol{\theta}^{k+1}}} - f^r_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}\right|\right], \tag{35}$$

where $\Delta\boldsymbol{\theta}^k$ is given as in Theorem 3.7.

We first prove that $\Delta_k > 0$ for small ϵ. This is obvious since among the non-zero elements of $\Delta\boldsymbol{\theta}$, we have $\Delta\boldsymbol{\theta}_i^k = \epsilon$ if $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})_i \Delta\boldsymbol{\theta}_i^k > 0$, and $\Delta\boldsymbol{\theta}_i^k = -\epsilon$ if $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})_i \Delta\boldsymbol{\theta}_i^k < 0$. Thus the term $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}^k$ must be positive. In fact, we have

$$\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}^k \geq \lambda\epsilon, \tag{36}$$

as long as $\|\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})\|_\infty > \lambda$. While $o(\epsilon)$ is the high-order infinitesimal of ϵ, we have

$$\Delta_k = \nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})^T \Delta\boldsymbol{\theta}^k + o(\epsilon) \geq \lambda\epsilon + o(\epsilon) > 0. \tag{37}$$

Assume that there are M elements of $\Delta\boldsymbol{\theta}^k$ that are non-zero. Then we have

$$\Delta_k = \sum_{i=1}^{N} |\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})_i \Delta\boldsymbol{\theta}_i^k| + o(\epsilon) \geq M\lambda\epsilon + o(\epsilon), \tag{38}$$

as the non-zero elements of $\Delta\boldsymbol{\theta}^k$ corresponds to elements of $\nabla_{\boldsymbol{\theta}} H(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\theta}^k}}, \mathbb{P}_{\bar{l}})$ that have absolute values larger than λ.

Note that

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{\mathcal{X}}\|_\infty \leq \sum_{i=1}^{k} \|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i-1}\|_\infty \leq k\epsilon < K\epsilon, \tag{39}$$

where $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^{\mathcal{X}}$. Thus we have

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^k\|_\infty < \epsilon} \|\nabla_{\boldsymbol{\theta}} f^r_{\boldsymbol{G}_{\boldsymbol{\theta}}}\|_\infty \leq L, \tag{40}$$

as

$$\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^k\|_\infty < \epsilon\} \in \{\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathcal{X}}\|_\infty \leq K\epsilon\}. \tag{41}$$

Thus we also have

$$\delta_k \leq L\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[\|\Delta\boldsymbol{\theta}\|_1] = L\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[M\epsilon] = ML\epsilon. \tag{42}$$

Then we can conclude

$$\frac{\Delta_k}{\delta_k} \geq \frac{M\lambda\epsilon}{ML\epsilon} + o(1) = \frac{\lambda}{L} + o(1). \tag{43}$$

Note that

$$\delta \leq \delta_1 + ... + \delta_K, \tag{44}$$

and

$$\Delta = \Delta_1 + ... + \Delta_K. \tag{45}$$

We then have

$$\frac{\Delta}{\delta} = \frac{\sum_{i=1}^{K} \Delta_K}{\delta} \geq \frac{\sum_{i=1}^{K} \Delta_i}{\sum_{i=1}^{K} \delta_i} \geq \frac{\sum_{i=1}^{K} (\frac{\lambda}{L} + o(1))\delta_i}{\sum_{i=1}^{K} \delta_i} =\geq \frac{\lambda}{L} + o(1). \tag{46}$$

Thus we have

$$\frac{\Delta}{\delta} \geq \frac{\lambda}{L} + o(1). \tag{47}$$

□

B. Experiment Setting

Generative Model Choice For FFHQ data domain, we use the pretrained StyleGAN2 generator offered by Awesome Pretrained StyleGAN2 [2] with config-f and 512×512 resolution. For BreCaHAD data domain, we use the official pretrained StyleGAN2-ADA generator[3]. For ImageNet data domain, we use the BigGAN256-Deep model in the official TFhub repository [4].

Posterior Estimation Model Choice For FFHQ data domain, we use the official pretrained ResNet50 classifiers provided by StyleGAN2 authors [5] as the posterior distribution $\mathbb{P}_l$. For BreCaHAD, we train a ResNet50 regressive model on the annotation subset of BreCaHAD dataset. The annotations mark all cancer nucleus in the tissue slice images. We train the ResNet50 regressive model to predict the number of cancer nucleus in each tissue slice image. We halt the training at the error rate of 10% in the training set to avoid overfitting, as the training dataset is small. The final predict scores are further normalized to $[0, 1]$ to yield posterior estimation to the severity of cancer. For ImageNet data domain, we use the official ResNet50 classifier provided by TensorFlow [6]. The ResNet50 classifier outputs a 1,000 dimensional vector, each of which predicts the posterior of a given image category. We choose the dimension corresponding to our knowledge of interest as the final posterior estimation $\mathbb{P}_l$.

Hyper-parameter Setting For 'Lipstick' extrapolation of FFHQ data domain, we set $K = 4$; for 'Gray Hair' extrapolation of FFHQ data domain, we set $K = 7$. For all the other experiments, we set $K = 10$. For FFHQ domain and BreCaHAD domain, we set $\epsilon = 1e - 3$; for ImageNet domain, we set $\epsilon = 1e - 5$. The choice of λ is set according to Fig. 4, where we choose a λ that is slightly smaller than λ_{max} for each experiment. For FFHQ domain, we set $\lambda = 1.8e - 4$; for BreCaHAD domain, we set $\lambda = 4.8e - 4$; for ImageNet domain, we set $\lambda = 1.3e - 4$. For all Dirac Knowledge Extrapolations, we set $\xi = 0.01$.

Training of InterFaceGAN We obtain the semantic boundary vectors as directed by the InterFaceGAN paper. We use the ResNet50 classifiers provided by the StyleGAN2 authors to annotate 50,000 random samples of the StyleGAN2 generator, and then train a Support Vector Machine (SVM) to predict the binary annotations predicted by the ResNet50 classifiers. The normalized support vectors of those SVMs are chosen to serve as the semantic boundaries for latent image editing.

[2] https://github.com/justinpinkney/awesome-pretrained-stylegan2

[3] https://github.com/NVlabs/stylegan2-ada/

[4] https://tfhub.dev/deepmind/biggan-deep-256/1

[5] https://github.com/NVlabs/stylegan2

[6] https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50/ResNet50

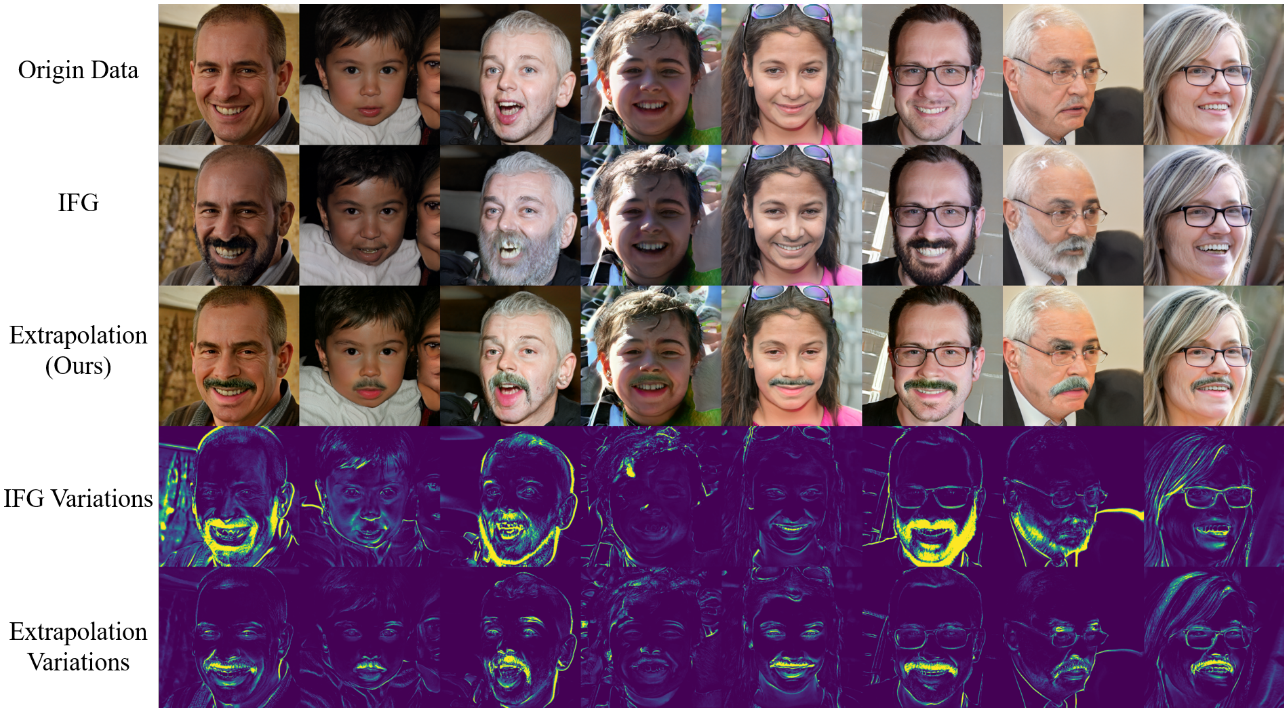


Figure 10. Dirac knowledge extrapolation of mustache on randomly sampled latent codes generating FFHQ faces. IFG denotes InterFaceGAN. IFG can successfully add mustache for regular cases, *e.g.*, mature male, but fails the counterfactual cases.

Style Space Analysis (Wu et al., 2021) We also compare the PKD method in Dirac distribution with another latent image editing method Style Space Analysis (SSA) (Wu et al., 2021), the results are reported in Fig. 14. We use the official code provided by Style Space Analysis authors [7].

User Study We conduct user studies on the FFHQ data domain for 'mustache' extrapolation by recruiting 50 volunteers. We randomly sample 3,000 latent codes from the prior distribution $\mathcal{N}(\mathcal{O}, \mathcal{I})$, and feed them to the pretrained StyleGAN2 generator to produce 3,000 synthesized facial images. The same operation is performed to the extrapolated distribution to yield another independent 3,000 synthesized facial images after knowledge extrapolation. Then we mix the StyleGAN2 synthesis facial images with the extrapolated synthesis facial images to yield a 6,000 testing set, and then randomly shuffle the testing set. All 50 volunteers are asked to answer the following questions for each image of the testing set: 1) Does the person in the image have mustache; 2) Is the person in the image a child without regard to mustache; 3) what is the gender of the person in the image without regard to mustache. For each question, the answer with the highest votes will be the winner.

Hardware Setting To train StyleGAN2 on 512×512 resolution FFHQ dataset and InterFaceGAN semantic boundaries, we use 8 NVIDIA V100 GPUs. To train the ResNet50 regressive model for BreCaHAD data domain, we use 1 NVIDIA GTX1080Ti GPU. For all the remaining experiments of knowledge extrapolation, we use 1 NVIDIA V100 GPU.

[7] https://github.com/betterze/StyleSpace

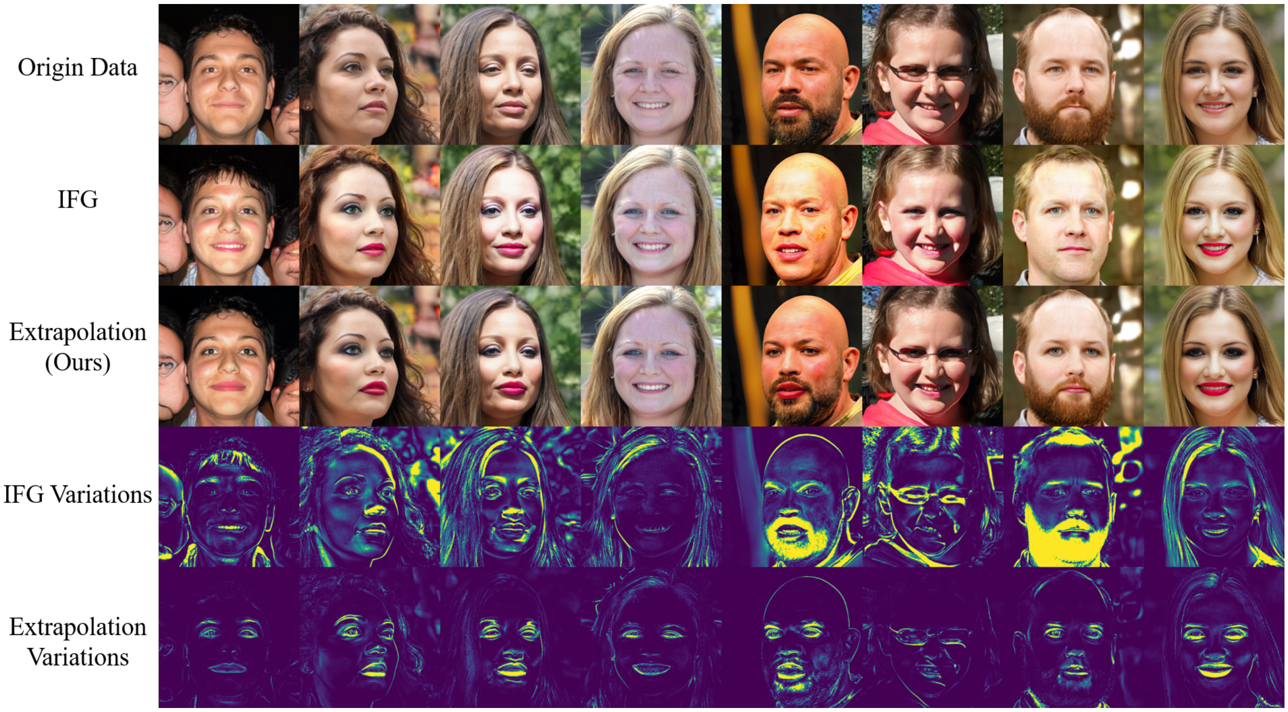


Figure 11. Dirac knowledge extrapolation of lipstick on randomly sampled latent codes generating FFHQ faces. IFG denotes InterFaceGAN. IFG can successfully add lipstick for regular cases, *e.g.*, female, but fails the counterfactual cases, *e.g.*, removing mustache or changing gender of the male cases.

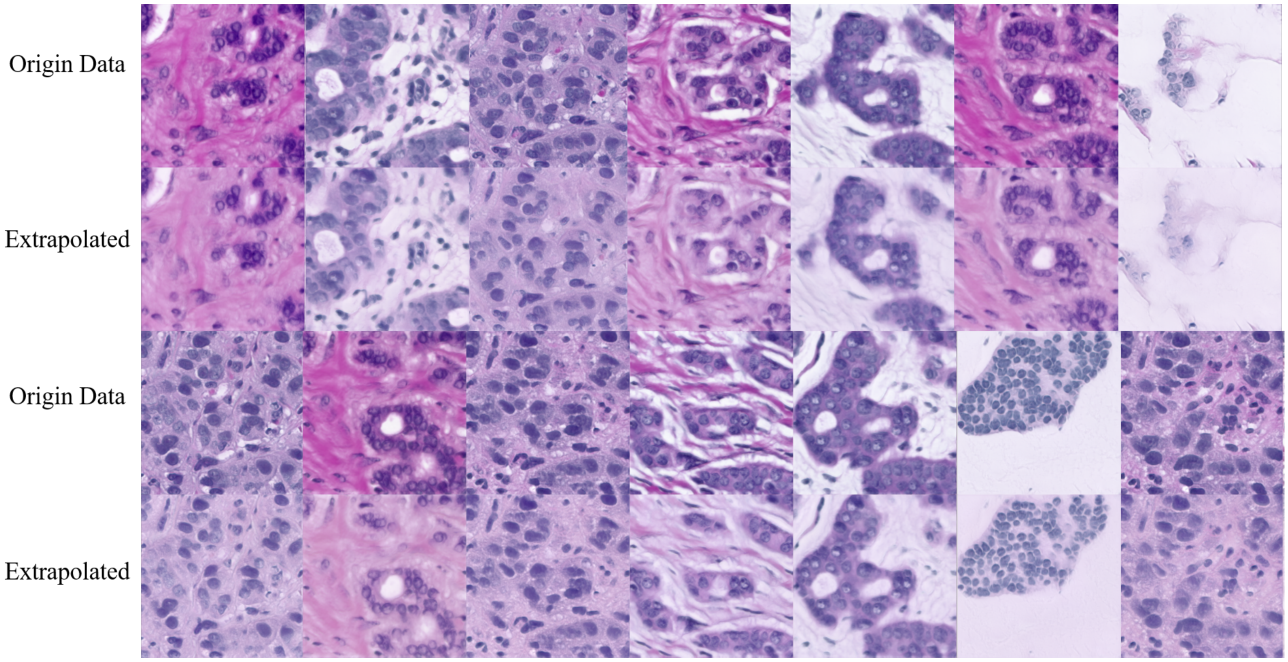
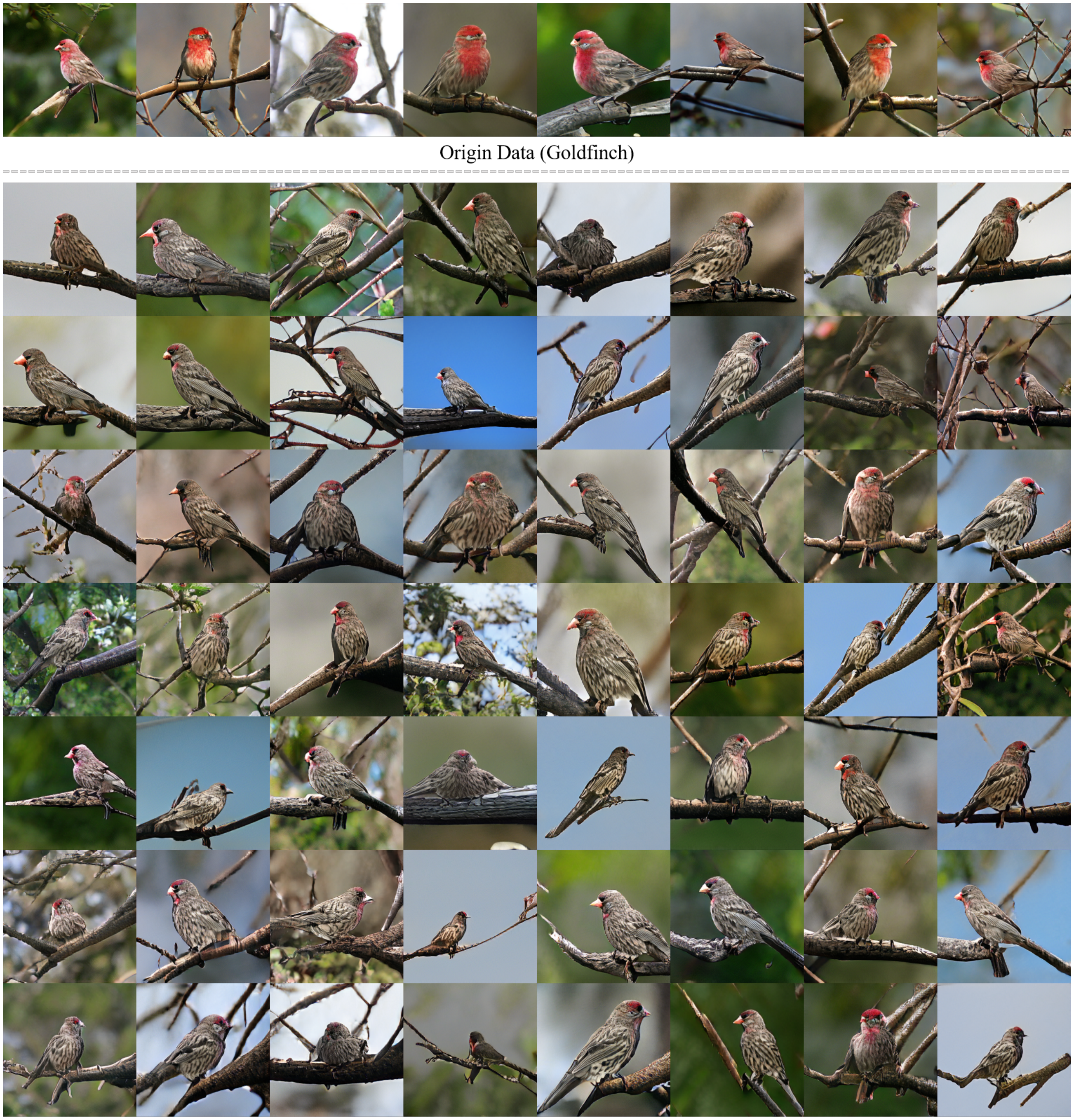


Figure 12. Dirac knowledge extrapolation of tissue slices of cancer patients. The cancer severity is reduced in the extrapolated cases.



Origin Data (Goldfinch)

Extrapolated Data **(Counterfact: Vulture Fur)**

Figure 13. Uncurated lists of BigGAN generations after extrapolating knowledge dimensions.

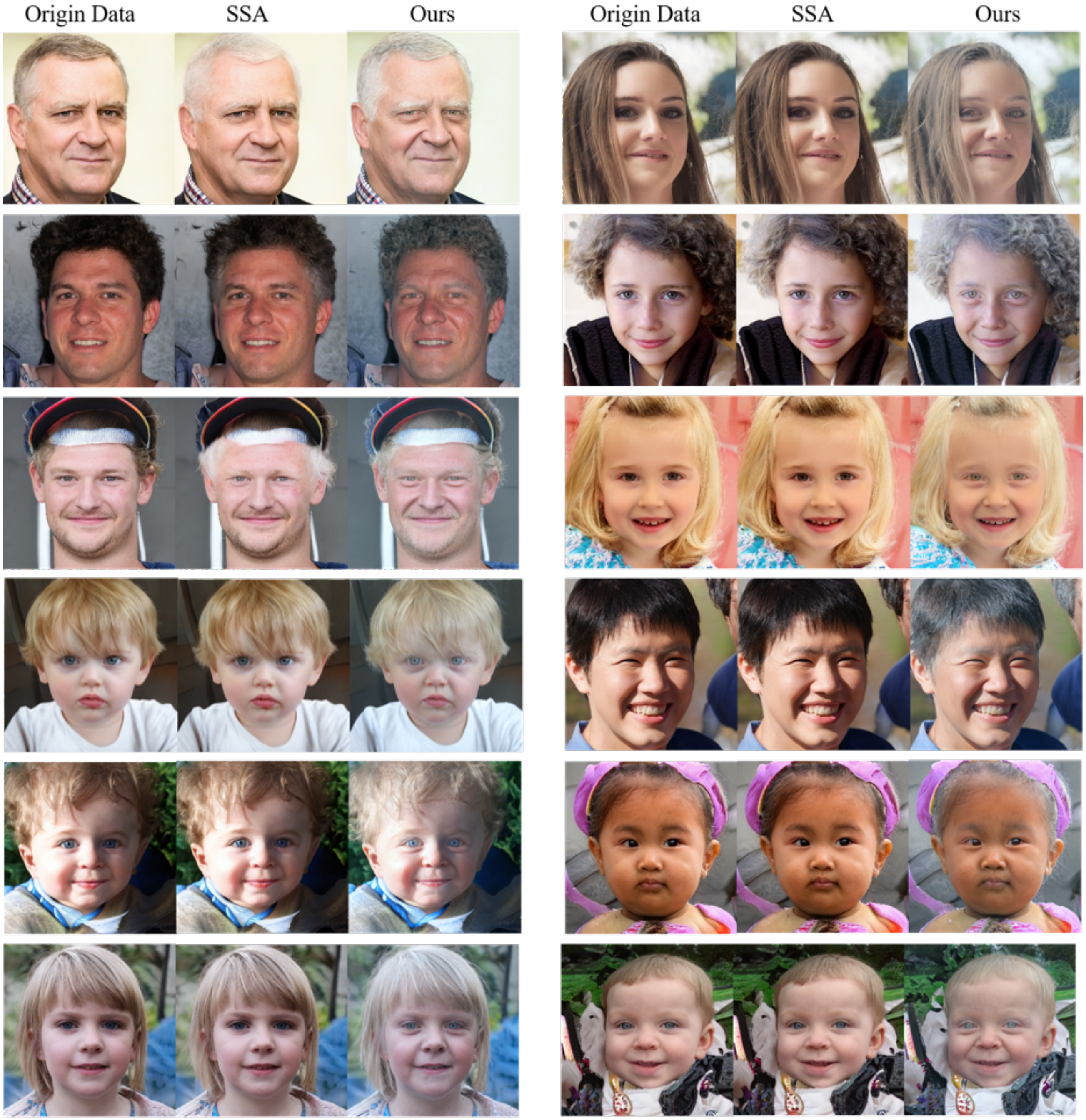


Figure 14. Extrapolation of knowledge ‘graph hair’. We compare our method with the latent image editing method Style Space Analysis (SSA). We set the editing strength of SSA to yield slightly stronger degree of hair grayness than ours in regular cases, *e.g.*, middle-age male, as shown in the first three rows of the left side. We then investigate the performance of both methods in counterfactual cases—young women or children in gray hair. The results show that SSA can hardly handle the counterfactual cases, while our method still works well.