

Hijack-GAN: Unintended-Use of Pretrained, Black-Box GANs

Hui-Po Wang¹ Ning Yu^{2,3} Mario Fritz¹

¹CISPA Helmholtz Center for Information Security, Germany

²Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

³University of Maryland, College Park

Abstract

While Generative Adversarial Networks (GANs) show increasing performance and the level of realism is becoming indistinguishable from natural images, this also comes with high demands on data and computation. We show that state-of-the-art GAN models – such as they are being publicly released by researchers and industry – can be used for a range of applications beyond unconditional image generation. We achieve this by an iterative scheme that also allows gaining control over the image generation process despite the highly non-linear latent spaces of the latest GAN models. We demonstrate that this opens up the possibility to re-use state-of-the-art, difficult to train, pre-trained GANs with a high level of control even if only black-box access is granted. Our work also raises concerns and awareness that the use cases of a published GAN model may well reach beyond the creators’ intention, which needs to be taken into account before a full public release. Code is available at <https://github.com/a514514772/hijackgan>.

1. Introduction

Generative Adversarial Nets (GANs) [14] have achieved remarkable success in many applications, such as image synthesis [20, 19, 21] and image translation [8, 10, 9, 33]. By learning a mapping between noise and images, the models are skilled to produce photo-realistic images from random noise. However, as the architectures become sophisticated [20, 19, 7], training modern GANs often requires massive data and computation resources. For example, it takes one high-quality face dataset, 8 V100 GPUs, and one week to train a single StyleGAN [1] model. In light of this trend, it is crucial to reuse existing pre-trained GANs, such as those being released by researchers or industry, for building green AI systems, in which the critical factor is to achieve other tasks beyond the original intention of GANs.

To reuse GANs for other tasks, prior works have shown that semantic manipulation can be realized by vector arith-

metic [23] or moving along constant attribute vectors [13, 28, 25, 12] in latent spaces. For instance, InterfaceGAN [25] demonstrates that facial attribute manipulation can be achieved by moving noise close to or away from the linear decision boundary of the desired attribute. Although these methods reveal the potential that pre-trained GANs could go beyond unconditional generation, they highly rely on the assumption of linear manifolds, thereby ignoring the nature of highly non-linear latent spaces (e.g., \mathcal{Z} -space of StyleGAN [25, 34]). This strong assumption could be harmful, especially for rare attributes, and lead to ineffective manipulation.

Inspired by this observation, we propose a novel framework, which gains high-level control over unconditional image generation by iteratively traversing the non-linear latent spaces. Specifically, we first train a proxy model that bypasses the gradients from one pretrained GAN and other fixed task models, and then dynamically decides the moving direction in each step, thus producing smoother and more effective attribute transition. Next, we propose an orthogonal constraint to solely edit the attribute of interest while retaining others in images. Despite only black-box access, we show that our method can achieve various unintended tasks, including manipulation over facial attributes, head poses, and landmarks.

As a result of our experiments, we find that our framework with pre-trained GANs not only facilitates other vision tasks but raises concerns regarding further usage. Even without access to model parameters, the models can still be applied to unintended tasks potentially for malicious purposes. The owners of GANs should be aware and cautious about the potential risks before releasing their models. Overall, our contributions are summarized below.

- We propose a framework which leverages off-the-shelf GANs to approach unintended vision tasks in a black-box setting.
- We propose a constraint that helps our framework solely edit the attribute of interest while retaining others in images.

- Extensive results show that our method can produce smoother and more effective manipulation while preserving non-target attributes better, as compared to prior work. We also shed light on the potential risks of unintended usage by gaining control over facial attributes, head poses, and landmarks.

2. Related Work

Generative Adversarial Nets (GANs). Since GAN [14] was proposed, it has advanced many applications such as image synthesis [1, 19, 21] and image translation [8, 10, 9, 33]. The rationale behind it is to map noise drawn from a simple distribution (e.g., Gaussian) to a real data distribution (e.g., images) by non-linear networks. Many works [20, 7, 19, 21, 30, 4] have been proposed to improve image quality while the demands on resources also significantly increase. In this work, we consider two state-of-the-art GANs, PGGAN [19] and StyleGAN [20], as our generator backbones. The former one takes a vector of noise as input and progressively upsamples features to generate images. The later one adopts a similar strategy but first embed the noise vector by neural networks and treat them as style representations. As observed by Shen *et al.* [25], the z -space of StyleGAN is expected to be more entangled than PGGAN. In general, GAN models for new tasks are often trained from scratch, making them inefficient and difficult to scale up. With our framework, we show that pre-trained GANs can well approach a range of applications beyond their original purposes.

Study on Latent Spaces of GANs. A considerable amount of works has attempted to understand the latent space of GANs. In particular, [23, 28] show that semantic manipulation is achievable by vector arithmetic on latent code; [20] show that style transfer can be achieved by mixing two latent represents. Some studies have attempted to identify semantically meaningful directions by self-supervised learning [29] or PCA on latent spaces [15]. They are able to perform transformations by moving noise toward the direction. However, since they approach it in an unsupervised fashion, they may be limited to generalize to new tasks beyond simple transformations.

Recent works further indicate that various semantic meanings may be implicitly encoded in the latent space of pre-trained GANs even though the concepts are not specified in the training set, ranging from memorability [13], transformation [18], to facial attributes [25, 12, 27]. For instance, InterfaceGAN [25] uses the normal vector of the SVM decision boundary to edit facial attributes by moving noise along the vector. Notably, most of the works mentioned above assume that the manifolds expand linearly in the latent space, ignoring that the underlying manifold could be extremely non-linear. In contrast to the prior

works, we propose a general framework that can be applied to various vision applications by traversing the non-linear manifolds in an iterative scheme.

Study on Leveraging Pre-trained GANs. Recently, several studies have investigated how to leverage existing GANs to achieve new tasks. One way is to treat pre-trained GANs as strong image priors, which can be achieved by either exploiting GANs as a component of their networks [22, 5, 6, 17, 3] or directly editing the latent code to find feasible solutions [1, 2, 11]. These methods often require access to model parameters or additional training data. In contrast, this work aims to study the potential usage of black-box pretrained GANs, advising the owners to be careful about the risk before releasing models.

3. Hijack-GAN

In this section, we first formalize the problem and then describe how to reuse models by our framework, followed by a proposed orthogonal constraint to improve disentanglement ability. Lastly, we discuss the practical considerations. The overall architecture is shown in Figure 1.

3.1. Problem Statement

We aim to study the possibility that pre-trained GANs can be used for unintended applications beyond unconditional generation. Specifically, we consider a victim generator $G : z \rightarrow I$ that maps noise $z \sim p(z)$ to realistic images I , and one or multiple victim task models $\mathcal{M} : I \rightarrow A$ that map images I to the attributes space A . We aim to find a trajectory \mathcal{T} in the latent space $p(z)$ such that, as traversing along the path, the desired task can be gradually achieved, which can be further expressed as follows:

$$\mathcal{L} \circ \mathcal{M} \circ G(z^{(i+1)}) \leq \mathcal{L} \circ \mathcal{M} \circ G(z^{(i)}), \forall z^{(i)} \in \mathcal{T}, \quad (1)$$

where \mathcal{L} is the loss function of unintended tasks. Note that the parameters and training data for the task are inaccessible, making it impossible to be directly solved by optimization methods since they require the gradients from the models.

3.2. Non-linear Traversal

To identify a meaningful trajectory in the black-box setting, we have to address the following two issues: (1) the gradients from the GAN and the task models are impassable. (2) we need to identify a trajectory that precisely describes the highly non-linear manifolds.

Since the gradients are impassable, we first train a proxy model \mathcal{P} to distill the information from the models. We assume that only z -space is available and synthesize data pairs $(z, \mathcal{M} \circ G(z))$ to train the proxy model such that it can map input noise to the attribute space; therefore, we

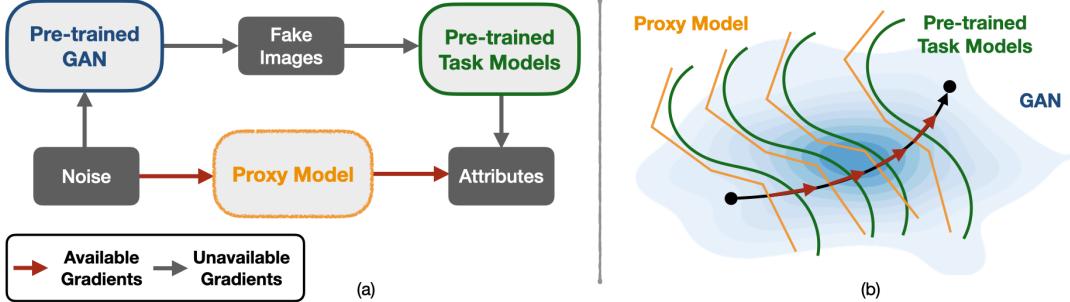


Figure 1. An overview of the proposed framework. The framework takes two steps to reuse GANs: (a) train a proxy model to distill information from pre-trained models, circumventing accessing the gradients of pre-trained models; (b) identify and iteratively traverse a non-linear trajectory under the guidance of gradients.

circumvent directly accessing the gradients while remaining informed about the relation between attributes and noise. Next, we compute the Jacobian matrix of the proxy model \mathcal{P} with respect to the input noise $z \in \mathbb{R}^n$.

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathcal{P}_1}{\partial z_1} & \dots & \frac{\partial \mathcal{P}_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{P}_m}{\partial z_1} & \dots & \frac{\partial \mathcal{P}_m}{\partial z_n} \end{bmatrix}, \quad (2)$$

where \mathcal{P}_j denotes j -th attributes predicted by the proxy. Each row vector of the matrix \mathbf{J} , termed \mathbf{J}_j for simplicity, can be interpreted as the direction in which the corresponding attribute changes most quickly; meanwhile, this vector also gives us a hint to identify a meaningful non-linear trajectory.

In light of this, we design an algorithm that iteratively updates the position of noise under the guidance of the vector \mathbf{J}_j :

$$z^{(i+1)} = z^{(i)} - \lambda \mathbf{J}_j^{(i)}, \quad (3)$$

where λ is a hyper-parameter deciding moving speed and $\mathbf{J}_j^{(i)}$ is associated with the attribute of interest at step i . Note that we normalize the $\mathbf{J}_j^{(i)}$ to better control the level of changes. By repeatedly computing Eq. 3 in each step, the target attribute in the generated image $G(z)$ would be gradually modified, granting us high-level control over image generation despite the black-box access.

3.3. Constraint for Disentanglement

In many cases, attributes may be entangled with each other, meaning that other non-target attributes would be changed if we solely follow the steepest direction in Eq. 3. To alleviate the problem, we additionally propose a constraint to encourage disentanglement. Since each row vector \mathbf{J}_j in Eq. 2 represents one direction that affects certain attributes most, we aim to find a vector which have the maximum inner product with the target direction \mathbf{J}_j while being

orthogonal to other non-target directions $\mathbf{J}_{k \neq j}$. We formulate the constraint as a linear program as follows.

$$\begin{aligned} & \underset{n}{\text{maximize}} \quad \mathbf{J}_j^T n \\ & \text{subject to} \quad An = 0, \end{aligned} \quad (4)$$

where n is the direction vector of interest, and each row of A consists of the attribute vector $\mathbf{J}_{k \neq j}$ on which we want to condition. By substituting n for $\mathbf{J}_j^{(i)}$ in Eq. 3, we can solely edit the attribute of interest while retaining other non-target attributes in the image. Note that since Eq. 4 is evaluated at each iteration, we are still capable of capturing the non-linear manifolds.

3.4. Implementation

We consider two state-of-the-art GANs, PGGAN [19] and StyleGAN [20], as the victim generators. The input of both models is 512-D noise drawn from a standard normal distribution. Following Shen *et al.* [25], we do not normalize the input noise for both models. The proxy models are implemented by a stack of fully-connected layers. Except for the last layer, each layer is followed by a ReLU activation function and a Dropout [26] layer with a rate of 0.2. We empirically find that a proxy model with 3 layers works sufficiently well for PGGAN, while it takes 8 layers to work on StyleGAN. It is also observed that the Dropout function plays a critical role to prevent over-fitting on StyleGAN. These observations imply that the z -space of StyleGAN may be highly entangled and non-linear.

For attribute manipulation, we find it important to train the proxy models with balanced datasets. We synthesize datasets for every attribute, each of which involves 100k positive and 100k negative samples. Since the annotations are unavailable, we adopt a ResNet-50 [16] classifier pre-trained on CelebAHQ [19] to annotate the generated images. We discard data with confidence lower than 0.9 to reduce the ambiguity. For other regression tasks, we generate random samples over the latent space and discard those with lower confidence. It is also observed that the proxy model

on StyleGAN benefits from more training data, while the effect is marginal on PGGAN.

4. Experiments

In this section, we provide analysis of our non-linear iterative scheme by applying it to attribute manipulation, head pose manipulation, and landmark editing, showing that our method benefits from the non-linearity and demonstrates the possibility of unintended usage.

4.1. Experiment Setup

Settings. We take two state-of-the-art GANs, StyleGAN [20] and PGGAN [19], as our backbone generators. Both models are pre-trained on CelebAHQ [19]. We adopt different task models according to the applications. We use ResNet-50 [16] pre-trained on CelebAHQ for attribute manipulation, HopeNet [24] pre-trained on 300W-LP [35] for head pose manipulation, and MTCNN [31] to crop faces and detect landmarks for landmark editing. Note that in all experiments, the model parameters and additional training data are unavailable.

Baselines. We consider two baselines here. First, following Denton *et al.* [12], we employ the same classifier as our framework to compute the gradient with respect to the initial point and take it as a constant attribute vector v . The attribute manipulation is realized as below,

$$z^{(i+1)} = z^{(i)} + \lambda v. \quad (5)$$

We refer to this baseline as *Linear*. The key difference is that the proposed framework recomputes gradients in every step. Similarly, the second baseline, InterfaceGAN [25], adopts the same linear strategy but derive the vector v from the normal vector of the decision boundary of a linear SVM, which can be viewed as a common direction that changes the attribute most. InterfaceGAN achieves conditional manipulation by pairwisely computing an orthogonal vector for the normal vectors. Note that, instead of using a common direction for all data, we derive moving directions for specific input.

4.2. Attribute Manipulation

We show that our framework can be applied to attribute manipulation. We consider unconditional and conditional settings, respectively. The former one is achieved by solely applying Eq. 3; The latter one adopts Eq. 4 to ensure the preservation of non-target attributes.

Unconditional Manipulation. In the unconditional setting, we expect to see an effective transition as compared to the baselines since our method dynamically decides the moving direction in each step and follows the steepest direction. Note that although the main objective here is to change the target attribute, other attributes might change arbitrarily.

Figure 2 compares our method to two baselines, *Linear* and InterfaceGAN, on 4 attributes. On PGGAN, all three methods successfully edit the attributes, while our method produces transition with much less distortion, e.g., smile and age. Also, we observe that our method and InterfaceGAN perform similarly on eyeglasses and gender. It may imply that for some attributes, the underlying manifold expands linearly as is derived by InterfaceGAN.

On the other hand, our method behaves clearly differently from the two linear baselines on StyleGAN, showing that the manifold may be highly non-linear. We verify the benefit of the iterative scheme by comparing our method to the baselines. First, *Linear* appears to fail to edit smiling even if it uses the same classifier as our framework. When comparing to InterfaceGAN, our method preserves more non-target contents than InterfaceGAN although the goal is not to preserve attributes. For example, age changes when editing eyeglasses, and hair color changes when editing smiles.

Effectiveness. To further verify the effectiveness of the proposed method, we compare our method to InterfaceGAN on logit changes over steps. We experiment on StyleGAN since it tends to be more non-linear as observed in Figure 2. Specifically, we choose 7 attributes and sample 2500 trajectories for each attribute, each of which consists of 40 steps with the step size equal to 0.2. Next, we compute the mean values over trajectories and report the target and non-target logit changes, respectively.

Figure 3 shows that our method can rapidly change the target attribute on most attributes, especially on rare attributes such as blond hair, pale skin, and narrow eyes. The result aligns with our hypothesis that not all attributes in the latent space distribute linearly; thus, our framework can benefit from the non-linearity. We additionally visualize the manipulation over the rare attributes in Figure 4, confirming the effectiveness of our method. In addition to effective manipulation, it is also seen that the non-target attributes of all methods are changed during the manipulation. It is thus crucial to limit such an undesired effect.

Conditional Manipulation. We present that our framework along with the proposed constraint in Eq. 4 can preserve the non-target attributes better. In this experiment, we consider 1 of 4 attributes to edit while conditioned on the other three. Figure 5 presents the results on PGGAN and StyleGAN, respectively. On PGGAN, our method strongly preserves the non-target attributes and effectively modify the target attribute. InterfaceGAN appears to be ineffective when it has to maintain non-target attributes. For example, the smiling is changed when editing eyeglasses. On StyleGAN, both methods succeed to edit the attributes while our method looks closer to the input. These results suggest that our method with the constraint can effectively edit the target attribute while retaining others.

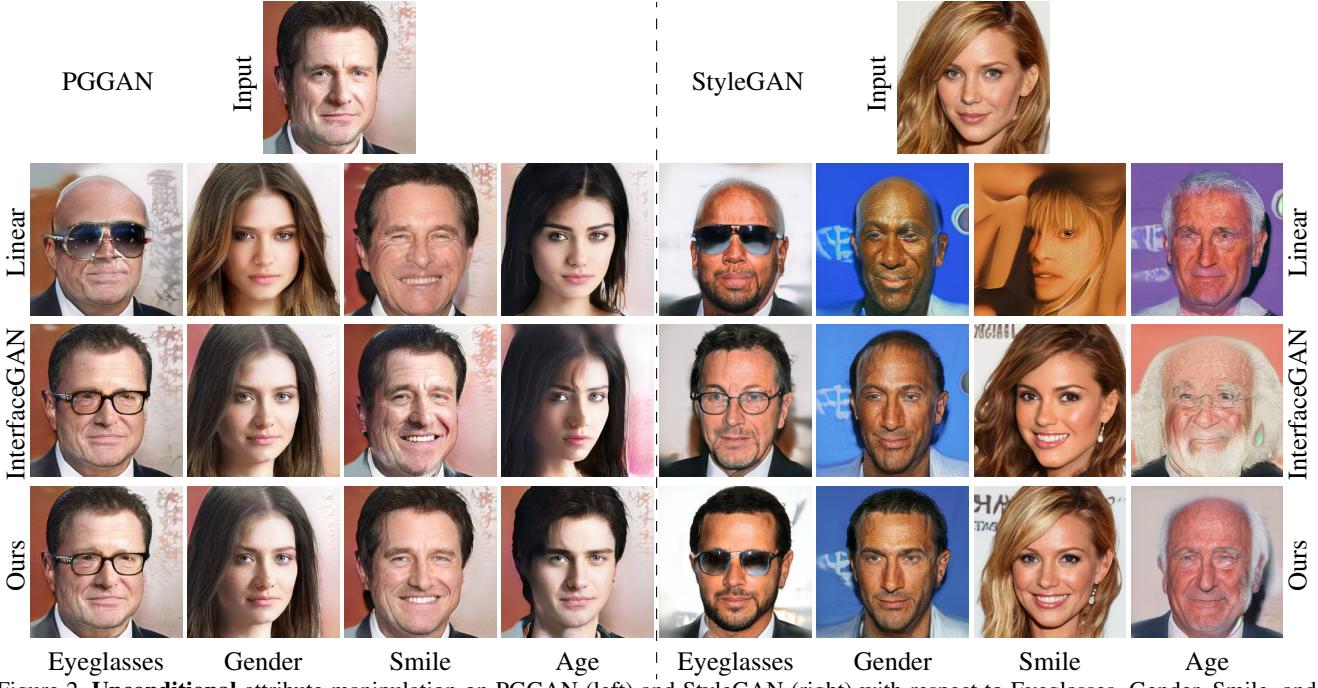


Figure 2. **Unconditional** attribute manipulation on PGGAN (left) and StyleGAN (right) with respect to Eyeglasses, Gender, Smile, and Age. We compare our method to *Linear* and InterfaceGAN.

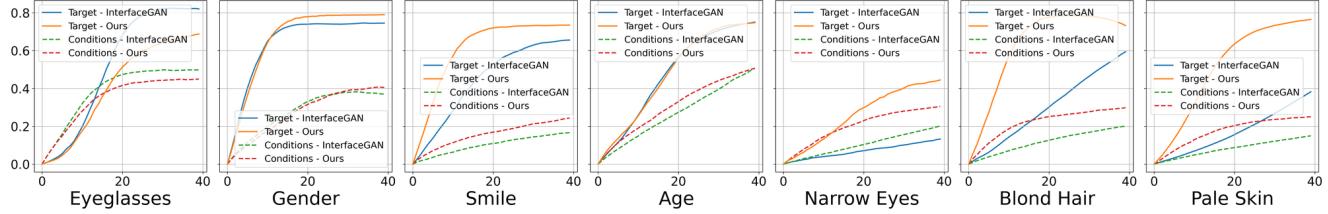


Figure 3. Logit changes over steps on StyleGAN. From left to right: Eyeglasses, Gender, Smile, Age, Narrow Eyes, Blond Hair, and Pale Skin. The solid lines represent the predictions of the target attributes while the dot lines represent the mean values over all the other non-target attribute predictions. Zoom in for better visualization.

Attribute Preservation. To inspect how well our method can preserve the non-target attributes, we follow the same sampling strategy as in Figure 3 and measure the ratio of target prediction changes to non-target prediction changes on StyleGAN. Higher values indicate that the subject can more effectively edit target attributes with lower perturbation to non-target attributes. We consider 7 attributes to edit. The former four attributes are conditioned on each other except themselves, and the latter three attributes are conditioned on all the former four. This design makes the latter three more challenging since they have to edit rare attributes while maintaining 4 attributes simultaneously.

Figure 6 compares our method to InterfaceGAN in both conditional (Eq. 4) and unconditional (Eq. 3) settings. It is immediately observed that our method with the proposed constraint achieves the highest ratio on almost all attributes, validating that the non-linear iterative scheme not only advances the attribute manipulation but also help the attribute

preservation. Interestingly, the ratio of our unconditional method on the last three attributes are higher than InterfaceGAN in the conditional setting. It could be attributed to inefficiency of linear methods. As observed in Figure 3, linear methods are less activate on rare attributes. Moreover, it uses the same direction for every step, which may be sub-optimal for the points far away from the initial point. Both two reasons together lead to the phenomenon.

4.3. Analysis

We conduct two analysis to provide more insights from the perspective of smoothness and function approximation.

Smoothness. We compare the trajectories generated by our method to those by linear methods such as InterfaceGAN in terms of smoothness. Previously, Karras *et al.* [20] propose a metric, Perceptual Path Length (PPL), to measure smoothness over the whole latent space. However, we only focus on smoothness of trajectories; thus we propose to use modi-

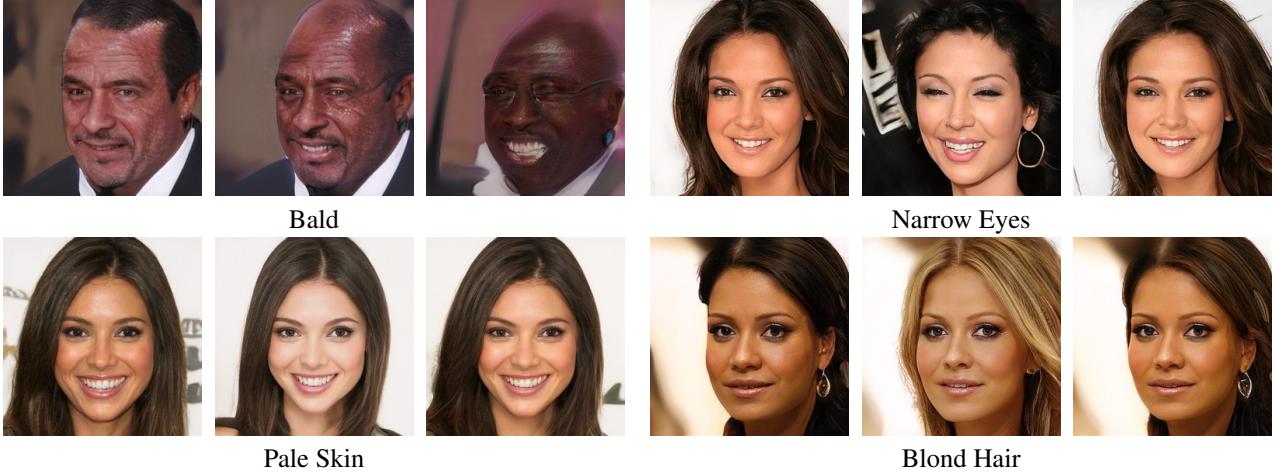


Figure 4. **Unconditional** rare attribute manipulation on StyleGAN with respect to Bald, Narrow Eyes, Pale Skin, and Blond Hair. Each set consists of input (left), our method (middle), and InterfaceGAN (right).

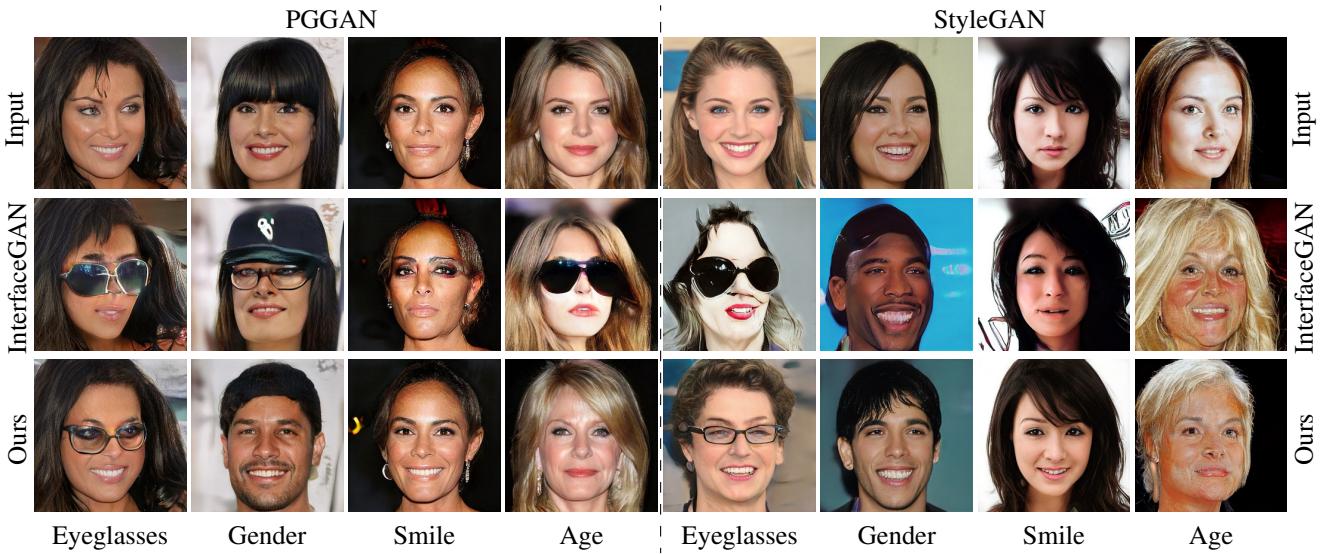


Figure 5. **Conditional** attribute manipulation on PGGAN (left) and StyleGAN (right) with respect to Eyeglasses, Gender, Smile, and Age. We compare our method to InterfaceGAN.

fied PPL (mPPL) that only consider noise sampled from the generated trajectories. Formally, we have

$$l_z = \mathbb{E}_{z^{(i)} \sim \mathcal{T}, \mathcal{T} \sim p(\mathcal{T})} [\frac{1}{\epsilon^2} d(G(\text{lerp}(z^{(i)}, z^{(i+1)}, t)), G(\text{lerp}(z^{(i)}, z^{(i+1)}, t + \epsilon)))] \quad (6)$$

where \mathcal{T} is a generated trajectory, $z^{(i)}$ is the noise of i -th step in the trajectory, G is the generator, lerp is linear interpolation function with factor t drawn from uniform distribution, ϵ is a small displacement, and $d(\cdot)$ is LPIPS metric [32]. For each attribute, we sample 1000 trajectories and each consists of 600 steps with step size 0.01, leading to 600k images in total. Following Karras *et al.* [20], we set ϵ to 1e-4 and report the mean value over all samples. Lower

values mean smoother trajectories.

Table 1 presents the evaluation results on both PGGAN and StyleGAN. In the unconditional setting, both methods produce comparably smooth trajectories. It also happens in the condition setting on PGGAN. It could be attributed to the relatively linear manifolds. Note that smoothness does not mean the effectiveness. One could generate images with no change to gain smoother transition. Lastly, our method surpasses the baseline on StyleGAN in the conditional setting by a large margin since it can strongly retain the non-target attributes, leading to smoother transition.

Function Approximation. We show that our iterative scheme can better traverse the highly non-linear latent space from the perspective of function approximation. We first re-

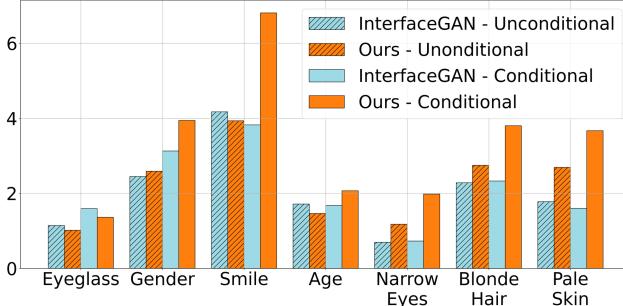


Figure 6. Attribute preservation ratio. We compare our method (orange) to InterfaceGAN (mist blue) in unconditional (non-stripe) and conditional (stripe) settings. The higher values, the better attributes are preserved during attribute manipulation.

	Cond.	Eyeglass	Gender	Smile	Age
PGGAN					
InterfaceGAN	N	60.69	65.00	54.49	61.16
	Ours	64.10	62.50	56.55	60.28
StyleGAN					
InterfaceGAN	N	99.15	101.65	96.90	96.62
	Ours	103.08	97.93	97.93	91.86
InterfaceGAN	Y	80.46	81.52	92.75	83.51
	Ours	57.23	71.67	78.18	55.41

Table 1. Modified Perceptual Path Length measured on PGGAN and StyleGAN. Lower is better.

	< 1	< 2	< 3	>= 3	Avg.
Eyeglasses					
InterfaceGAN	1.760	2.779	3.644	1.481	2.416
	1.675	2.401	2.469	1.557	2.026
Gender					
InterfaceGAN	5.702	4.045	1.798	0.891	3.109
Ours	4.469	3.694	1.790	0.812	2.692
Smile					
InterfaceGAN	1.764	1.783	1.693	0.961	1.550
Ours	3.191	2.391	1.611	0.921	2.028
Age					
InterfaceGAN	2.350	2.434	2.312	1.354	2.113
Ours	0.969	1.109	1.893	1.285	1.314

Table 2. Errors of first order Taylor approximation. Errors are reported according to the L2-norm between the initial point and the estimated points. f consists of the generator and the task model.

call the first order Taylor approximation,

$$f(x) = f(a) + f'(a)(x - a). \quad (7)$$

If a method models the underlying manifold better, it is expected to have lower error between the real logit $f(z^{(i+1)})$ and the estimated logit $f(z^{(i)}) + f'(z^{(i)})(z^{(i+1)} - z^{(i)})$. Note that the real gradient f' in Eq. 7 is unavailable for all baselines and our model. Instead, all methods use estimated gradients, namely J_j (Eq. 3) for our method and v (see Section 4.1) for InterfaceGAN.

Since InterfaceGAN uses normal vectors of SVM decision boundaries as guiding signals, we expect it to work well in a local region while becoming less accurate in the farther distance. To verify this assumption, we measure the errors and categorize them according to the L2 distance between the estimated and initial points. Table 2 compares our method to InterfaceGAN. It shows that our method approximates the underlying function more accurately on average. As for the smiling, since we discard the data with lower confidence (see Section 3.4), it may make the proxy model less robust at the region where many ambiguous samples happen. Lastly, due to the iterative scheme, our method achieves lower errors even when the estimated points moves far away from the initial points.

4.4. Head Pose Manipulation

To further study the possibility of unintended usage of pretrained GANs, we employ our framework to achieve head pose manipulation by leveraging pre-trained HopeNet [24] as the task model. In particular, we attempt to gain control over the yaw and pitch axes since the training dataset of the pre-trained GANs, CelebAHQ, is well-aligned by eyes; thus it is almost impossible to control the roll axis. It also suggests that the framework could be still limited by the training dataset of pretrained GANs even though we are capable of gaining control over the generation process. To generate fine-grained transition, we choose small step size 0.01 and produces 1000 steps in total.

Due to the proposed constraint in Eq. 4, our framework can solely edit one axis by conditioning on four attributes and the other axis. As shown in Figure 7, we smoothly interpolate the images for yaw and pitch axes. We also note that some artifacts or undesired changes happen when forcing the pose to go beyond certain degrees. With our framework, we can generate faces with arbitrary poses if they remain feasible in the training data of pre-trained GANs.

4.5. Landmark Editing

By utilizing pre-trained MTCNN [31], we take one step further to consider landmark editing, which is a relatively challenging application. The reasons are three-fold. First, a landmark is two coordinates, which are extremely local attributes as compared to, e.g., age or gender. Second, since

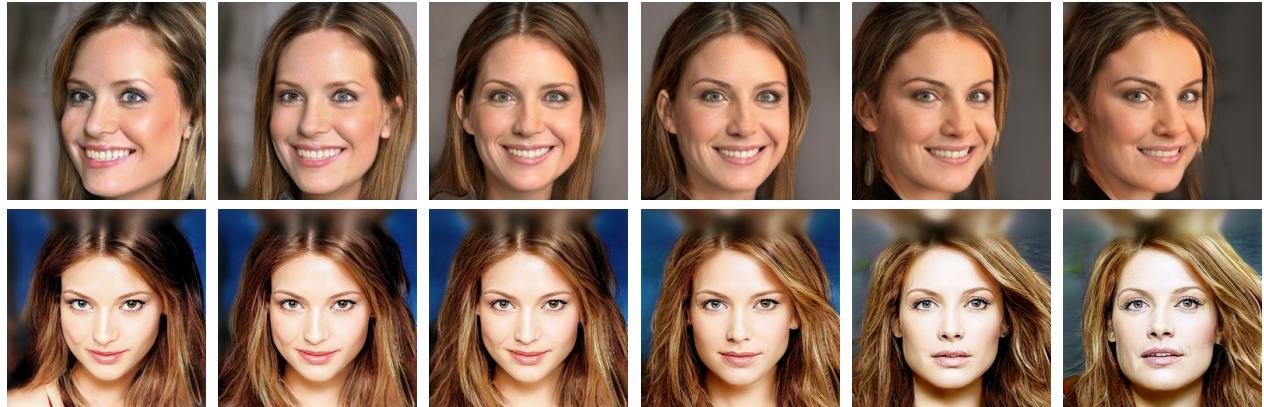


Figure 7. Pose manipulation on StyleGAN with respect to yaw (top) and pitch (bottom).

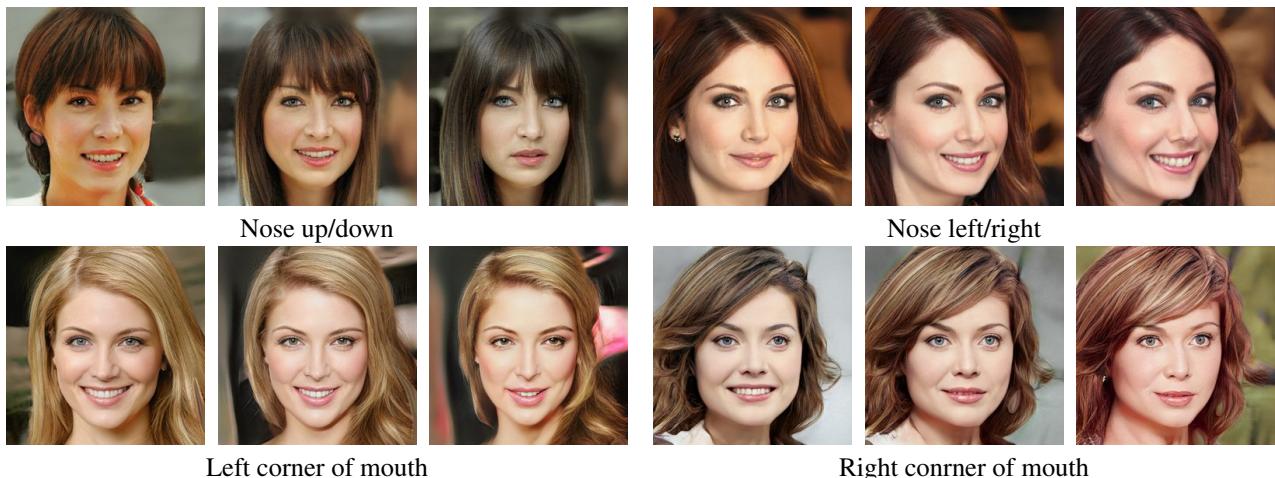


Figure 8. Landmark editting on StyleGAN with respect to Nose up/down, Nose left/right, Left corner of mouth, Right corner of mouth.

landmarks are coordinates, it becomes very sensitive for annotation. The performance may be degraded if the accuracy of task models are not satisfactory. Lastly, landmarks are often correlated with other features. Based on the results of previous experiments, we edit landmarks while conditioning on 4 attributes, poses, and a subset of landmarks. Note that since the training dataset is well aligned, the landmark of eyes are not editable.

Figure 8 demonstrates the results when editing the coordinates of noses and mouths. Surprisingly, we find some attributes are highly entangled. For instance, *mouth landmarks* is highly correlated with *smile*. To measure the correlation, we additionally compute the cosine similarity between direction vectors (Eq. 3). Higher similarity indicates higher entanglement. We analyze the correlation of *mouth landmarks vs. smile* (0.296) and *mouth landmarks vs. gender* (0.035). The former one is around 8.5 times larger than the latter. Additionally, *nose landmarks vs. yaw* are highly correlated by 0.5014. In this analysis, we do find measurable entanglement.

5. Conclusion

In this paper, we demonstrate that publicly released state-of-the-art GANs can be applied for a wide range of applications beyond the creators' intention. To achieve this, we propose a framework to gain high-level control over unconditional image generation. Extensive results show that the framework is advantageous over non-iterative linear baselines in many aspects and can be readily applied to various unintended tasks. While this allows to re-use high quality GANs that are becoming increasingly costly to train for new applications, we hope that this work also raises awareness regarding potential unintended usage, urging the creators and owners to be cautious about the possible risks of their models before a full release.

Acknowledgment

This work is partially funded by the Helmholtz Association within the project "Protecting Genetic Data with Synthetic Cohorts from Deep Generative Models (PRO-GENE-GEN)" (ZT-I-PF-5-23).

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Int. Conf. Comput. Vis.*, pages 4432–4441, 2019. [1](#), [2](#)
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8296–8305, 2020. [2](#)
- [3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv e-prints*, pages arXiv–2008, 2020. [2](#)
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. [2](#)
- [5] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *Eur. Conf. Comput. Vis.*, 2020. [2](#)
- [6] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4):1–11, 2019. [2](#)
- [7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Int. Conf. Learn. Represent.*, 2018. [1](#), [2](#)
- [8] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1900–1909, 2019. [1](#), [2](#)
- [9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8789–8797, 2018. [1](#), [2](#)
- [10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8188–8197, 2020. [1](#), [2](#)
- [11] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5771–5780, 2020. [2](#)
- [12] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. [1](#), [2](#), [4](#)
- [13] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Int. Conf. Comput. Vis.*, pages 5744–5753, 2019. [1](#), [2](#)
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–2680, 2014. [1](#), [2](#)
- [15] Erik Häkkinen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. [2](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. [3](#), [4](#)
- [17] Shady Abu Hussein, Tom Tiler, and Raja Giryes. Image-adaptive gan based reconstruction. In *AAAI*, volume 34, pages 3121–3129, 2020. [2](#)
- [18] Ali Jahanian, Lucy Chai, and Phillip Isola. On the “steerability” of generative adversarial networks. In *Int. Conf. Learn. Represent.*, 2019. [2](#)
- [19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018. [1](#), [2](#), [3](#), [4](#)
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
- [21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. [1](#), [2](#)
- [22] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *Eur. Conf. Comput. Vis.*, 2020. [2](#)
- [23] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. [1](#), [2](#)
- [24] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. [4](#), [7](#)
- [25] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9243–9252, 2020. [1](#), [2](#), [3](#), [4](#)
- [26] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. [3](#)
- [27] Ayush Tewari, Mohamed Elgarib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6142–6151, 2020. [2](#)
- [28] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7064–7073, 2017. [1](#), [2](#)
- [29] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, 2020. [2](#)

- [30] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. [2](#)
- [31] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. [4](#), [7](#)
- [32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. [6](#)
- [33] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. In *AAAI*, pages 13058–13065, 2020. [1](#), [2](#)
- [34] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Eur. Conf. Comput. Vis.*, 2020. [1](#)
- [35] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 146–155, 2016. [4](#)

Appendix

A. Quality of Generated Images

We follow the evaluation protocol of StyleGAN and report average FID scores. As shown in Table A1, we observe that both methods degrade the image quality of StyleGAN. As compared to InterfaceGAN, our method performs slightly better in the unconditional setting and significantly better in the conditional setting. The experiment confirms the effectiveness of our non-linear method on the quality of generated images. Note that following InterfaceGAN, we do not normalize latent codes and find the vanilla setting slightly worse than the one reported by StyleGAN (5.04).

	Unconditional	Conditional
Interfacegan	26.70	21.46
Ours	23.79	14.90
Vanilla GAN	11.14 (5.04)	

Table A1. FID evaluation on StyleGAN. Lower is better.

B. Detailed Iterative Algorithm

We detail the proposed iterative framework in Algorithm 1. We first recall our proposed framework. The Jacobian matrix of the proxy model is computed as follows,

$$\mathbb{J} = \begin{bmatrix} \frac{\partial \mathcal{P}_1}{\partial z_1} & \dots & \frac{\partial \mathcal{P}_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{P}_m}{\partial z_1} & \dots & \frac{\partial \mathcal{P}_m}{\partial z_n} \end{bmatrix}, \quad (\text{A1})$$

where \mathcal{P}_j denotes j-th attributes predicted by the proxy. We iteratively explore the latent space by dynamically discovering the most representative direction, which can be shown as below.

$$z^{(i+1)} = z^{(i)} - \lambda \mathbb{J}_j^{(i)}, \quad (\text{A2})$$

where λ is a hyper-parameter deciding moving speed and $\mathbb{J}_j^{(i)}$ is associated with the attribute of interest at step i. To encourage disentanglement, we further propose an orthogonal constraint.

$$\begin{aligned} & \underset{n}{\text{maximize}} \quad \mathbb{J}_j^T n \\ & \text{subject to} \quad An = 0, \end{aligned} \quad (\text{A3})$$

where n is the direction vector of interest, and each row of A consists of the attribute vector $\mathbb{J}_{k \neq j}$ on which we want to condition. Note that we take decreasing logit values as an example for simplicity. One could achieve the opposite direction by gradient ascent.

Algorithm 1: HijackGAN

```

Input: Trajectory length  $L$ , step size  $\lambda$ , initial point
 $z^{(0)}$ , target attribute index  $j$ , proxy model  $\mathcal{P}$ ,
condition index set  $K$ 
Output: Trajectory  $\mathcal{T} = \{z^{(1)}, \dots, z^{(N)}\}$ 
1 Initialize  $\mathcal{T} \leftarrow \emptyset$ ;
2 for  $i$  in  $\{0, \dots, L - 1\}$  do:
3   Compute the Jacobian matrix  $\mathbb{J}^{(i)}$  by Eg. A1 ;
4   if  $K \neq \emptyset$  then
5     Construct matrix  $A$  by vectors  $\mathbb{J}_{k \in K}^{(i)}$  ;
6     Solve orthogonal vector  $n$  by Eq. A3 ;
7   else
8      $| n \leftarrow \mathbb{J}_j^{(i)}$ 
9   end
10   $| z^{(i+1)} \leftarrow z^{(i)} - \lambda n$  (Eq. A2);
11  Add  $z^{(i+1)}$  to  $\mathcal{T}$  ;
12 end
13 return  $\mathcal{T}$ 

```

C. Visualization of Smoothness

Although we have verified the smoothness of our method in terms of modified Perceptual Path Length in Table 1 of the main paper, we additionally provide qualitative results on PGGAN (Figure A1) and StyleGAN (Figure A2), respectively. All experiments are conducted in the conditional setting, meaning that we solely edit one attribute, and others should remain the same. In particular, we take 40 steps with step size 0.2 and show the images from every 5 steps. Note that, the closer to the right-hand side, the farther from the initial point.

From Figure A1 and Figure A2, we make two observations. First, our method preserves attributes better on both models. For example, our method can preserve smiling when editing eyeglasses, and preserve the smiling when editing age on both models. Second, on StyleGAN, our method can produce smoother transitions. For example, gender and eyeglasses.

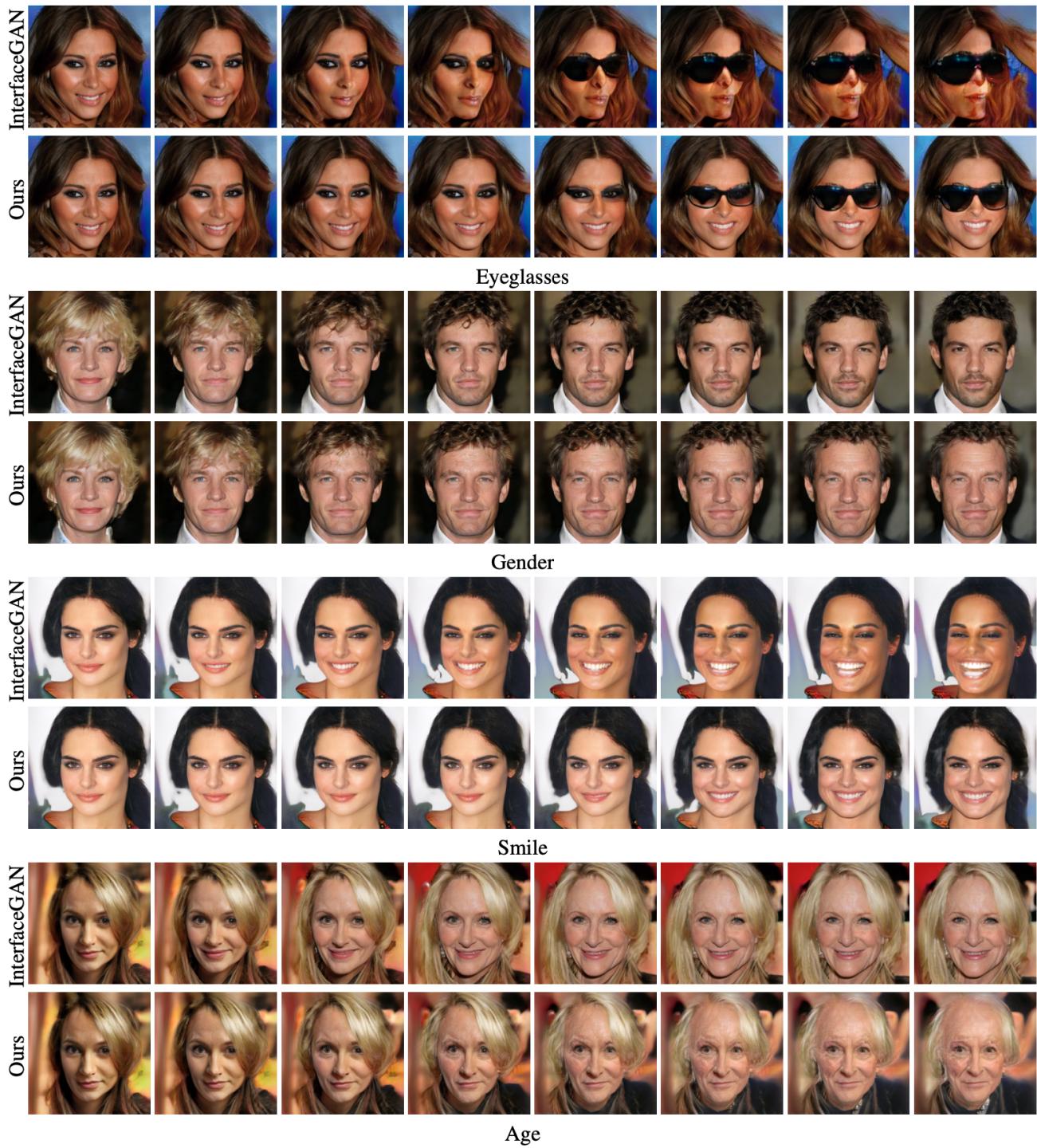


Figure A1. **(Conditional)** visualization of smoothness on PGGAN. All attributes should remain the same except the target one.

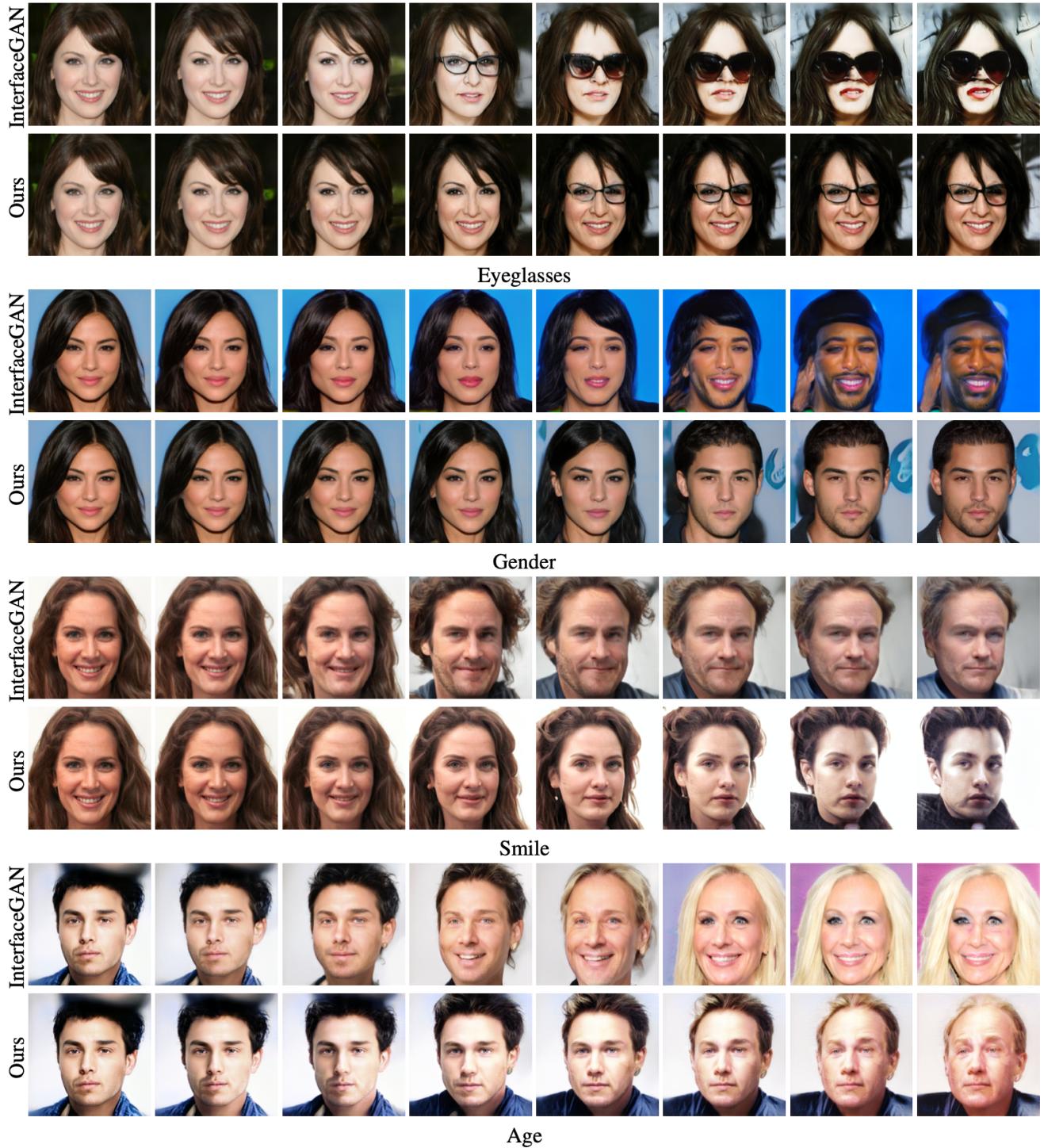


Figure A2. **(Conditional)** visualization of smoothness on StyleGAN. All attributes should remain the same except the target one.