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

Hui-Po Wang[1] Ning Yu[2,3] Mario Fritz[1]

[1]CISPA Helmholtz Center for Information Security, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3]University of Maryland, College Park

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,

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

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.

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

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.

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

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 λ, initial point $z^{(0)}$, target attribute index j, proxy model $\mathcal{P}$, condition index set K
Output: Trajectory $\mathcal{T} = \{z^{(1)}, \dots, z^{(N)}\}$

Initialize $\mathcal{T} \leftarrow \emptyset$;
for *i* **in** $\{0, \dots, L-1\}$ **do:**
  Compute the Jacobian matrix $\mathbb{J}^{(i)}$ by Eg. A1 ;
  if $K \neq \emptyset$ **then**
    Construct matrix A by vectors $\mathbb{J}_{k \in K}^{(i)}$;
    Solve orthogonal vector n by Eq. A3 ;
  else
    $n \leftarrow \mathbb{J}_j^{(i)}$
  end
  $z^{(i+1)} \leftarrow z^{(i)} - \lambda n$ (Eq. A2);
  Add $z^{(i+1)}$ to $\mathcal{T}$;
end
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 observation. 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.

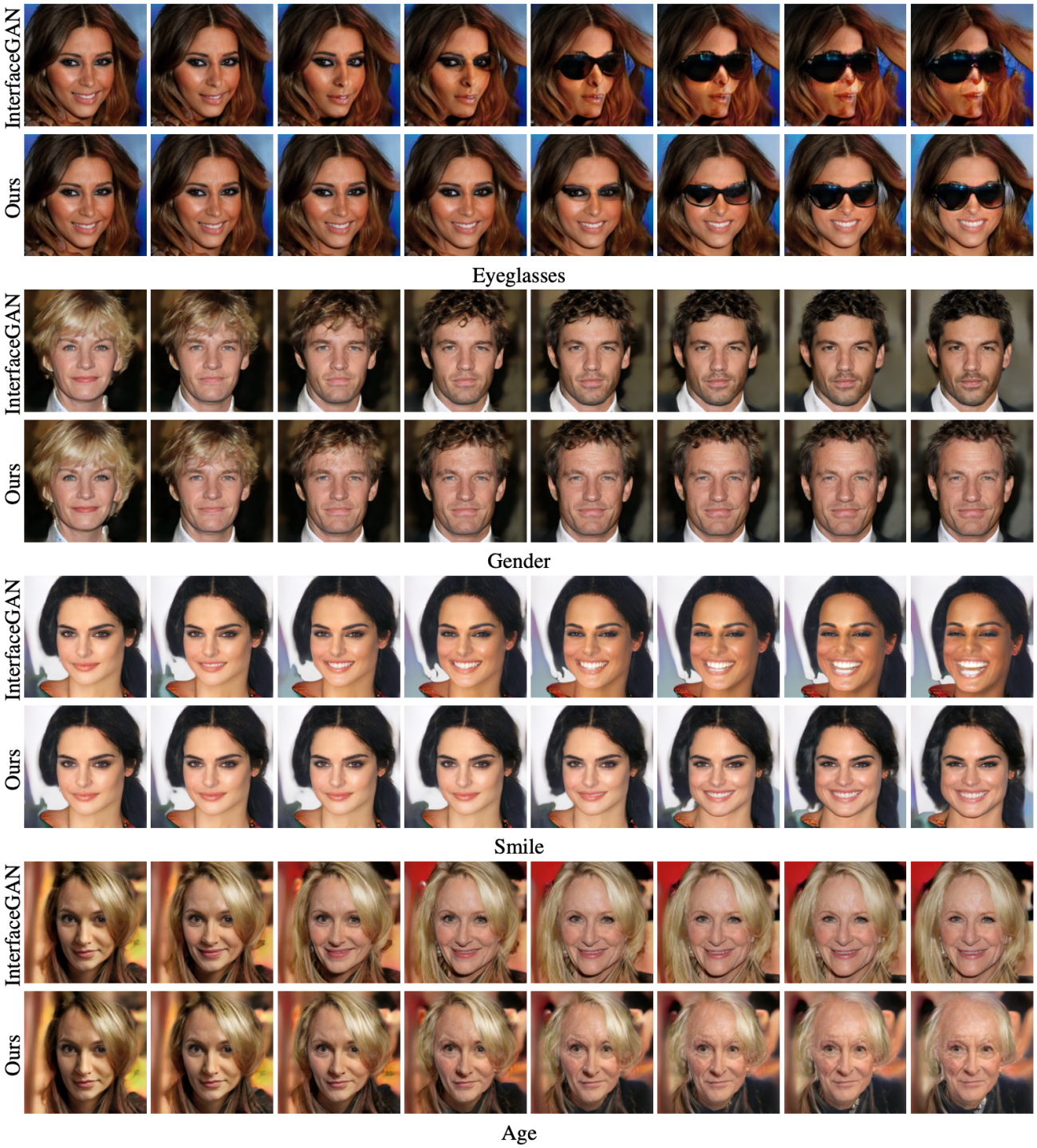

Figure A1. **(Conditional)** visualization of smoothness on PGGAN. All attributes should remain the same except the target one.

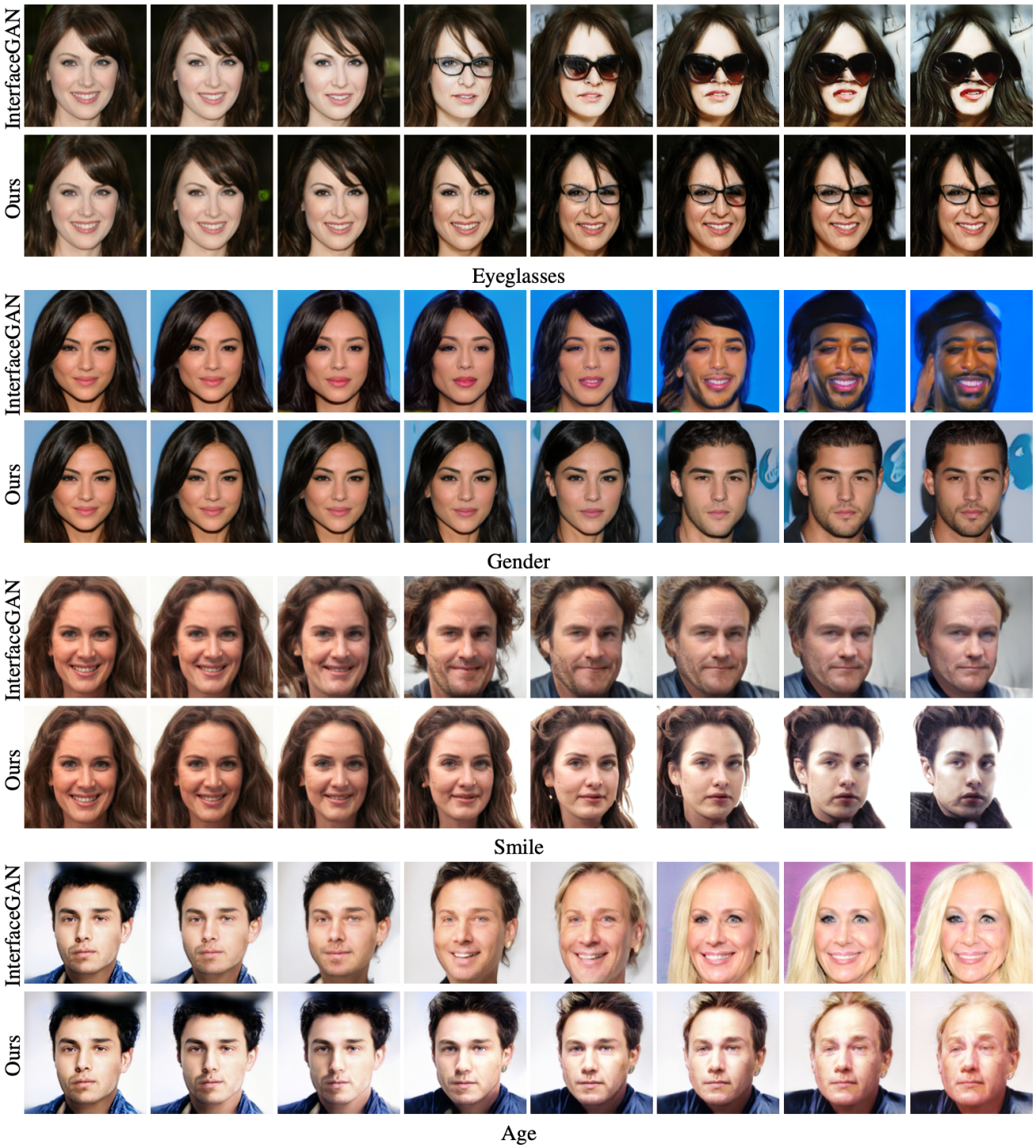

Figure A2. **(Conditional)** visualization of smoothness on StyleGAN. All attributes should remain the same except the target one.