Interpreting the Latent Space of GANs for Semantic Face Editing Supplementary Material

Yujun Shen[1], Jinjin Gu[2], Xiaoou Tang[1], Bolei Zhou[1]
[1]The Chinese University of Hong Kong [2]The Chinese University of Hong Kong, Shenzhen
{sy116, xtang, bzhou}@ie.cuhk.edu.hk, jinjingu@link.cuhk.edu.cn

1. Overview

This supplementary material contains the following information:

- We introduce the implementation details of the proposed InterFaceGAN in Sec.2.
- We provide the detailed proof of *Property 2* in the main paper in Sec.3.
- Please also refer to this video to see continuous attribute editing results.

2. Implementation Details

We choose five key facial attributes for analysis, including pose, smile (expression), age, gender, and eyeglasses. The corresponding positive directions are defined as turning right, laughing, getting old, changing to male, and wearing eyeglasses. Note that we can always plug in more attributes easily as long as the attribute detector is available.

To better predict these attributes from synthesized images, we train an auxiliary attribute prediction model using the annotations from the CelebA dataset [3] with ResNet-50 network [1]. This model is trained with multi-task losses to simultaneously predict smile, age, gender, eyeglasses, as well as the 5-point facial landmarks. Here, the facial landmarks will be used to compute yaw pose, which is also treated as a binary attribute (left or right) in further analysis. Besides the landmarks, all other attributes are learned as bi-classification problem with softmax cross-entropy loss, while landmarks are optimized with l_2 regression loss. As images produced by PGGAN and StyleGAN are with 1024×1024 resolution, we resize them to 224×224 before feeding them to the attribute model.

Given the pre-trained GAN model, we synthesize $500K$ images by randomly sampling the latent space. There are mainly two reasons in preparing such large-scale data: (i) to eliminate the randomness caused by sampling and make sure the distribution of the latent codes is as expected, and (ii) to get enough wearing-glasses samples, which are really rare in PGGAN model.

To find the semantic boundaries in the latent space, we use the pre-trained attribute prediction model to assign attribute scores for all $500K$ synthesized images. For each attribute, we sort the corresponding scores, and choose $10K$ samples with highest scores and $10K$ with lowest ones as candidates. The reason in doing so is that the prediction model is not absolutely accurate and may produce wrong prediction for ambiguous samples, *e.g.*, middle-aged person for age attribute. We then randomly choose 70% samples from the candidates as the training set to learn a linear SVM, resulting in a decision boundary. Recall that, normal directions of all boundaries are normalized to unit vectors. Remaining 30% are used for verifying how the linear classifier behaves. Here, for SVM training, the inputs are the $512d$ latent codes, while the binary labels are assigned by the auxiliary attribute prediction model.

3. Proof

In this part, we provide detailed proof of *Property 2* in the main paper. Recall this property as follow.

Property 2 *Given* $\mathbf{n} \in \mathbb{R}^d$ *with* $\mathbf{n}^T\mathbf{n} = 1$*, which defines a hyperplane, and a multivariate random variable* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$*, we have* $\mathrm{P}(|\mathbf{n}^T\mathbf{z}| \leq 2\alpha\sqrt{\frac{d}{d-2}}) \geq (1 - 3e^{-cd})(1 - \frac{2}{\alpha}e^{-\alpha^2/2})$ *for any* $\alpha \geq 1$ *and* $d \geq 4$*. Here* $\mathrm{P}(\cdot)$ *stands for probability and* c *is a fixed positive constant.*

Proof.

Without loss of generality, we fix $\mathbf{n}$ to be the first coordinate vector. Accordingly, it suffices to prove that $\mathrm{P}(|z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}}) \geq (1 - 3e^{-cd})(1 - \frac{2}{\alpha}e^{-\alpha^2/2})$, where z_1 denotes the first entry of $\mathbf{z}$.

As shown in Fig.1, let H denote the set

$$\{\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_d) : ||\mathbf{z}||_2 \leq 2\sqrt{d}, |z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}}\},$$

where $||\cdot||_2$ stands for the l_2 norm. Obviously, we have

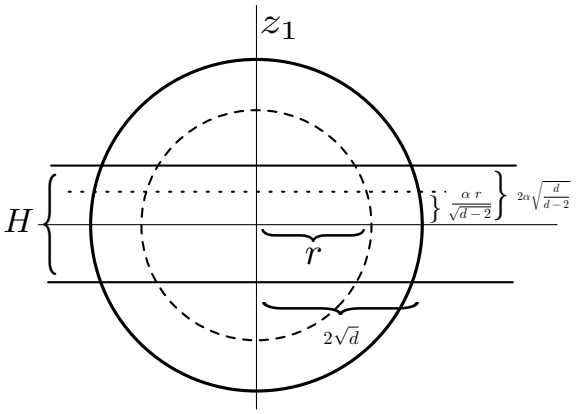

Figure 1: Illustration of *Property 2*, which shows that most of the probability mass of high-dimensional Gaussian distribution lies in the thin slab near the "equator".

$\mathrm{P}(H) \leq \mathrm{P}(|z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}})$. Now, we will show $\mathrm{P}(H) \geq (1-3e^{-cd})(1-\frac{2}{\alpha}e^{-\alpha^2/2})$

Considering the random variable $R = ||\mathbf{z}||_2$, with cumulative distribution function $F(R \leq r)$ and density function $f(r)$, we have

$$\begin{aligned}\mathrm{P}(H) &= \mathrm{P}(|z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}}|R \leq 2\sqrt{d})\mathrm{P}(R \leq 2\sqrt{d}) \\ &= \int_0^{2\sqrt{d}} \mathrm{P}(|z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}}|R=r)f(r)dr.\end{aligned}$$

According to *Theorem 1* below, when $r \leq 2\sqrt{d}$, we have

$$\begin{aligned}\mathrm{P}(H) &= \int_0^{2\sqrt{d}} \mathrm{P}(|z_1| \leq 2\alpha\sqrt{\frac{d}{d-2}}|R=r)f(r)dr \\ &= \int_0^{2\sqrt{d}} \mathrm{P}(|z_1| \leq \frac{2\sqrt{d}}{r}\frac{\alpha}{\sqrt{d-2}}|R=1)f(r)dr \\ &\geq \int_0^{2\sqrt{d}} \mathrm{P}(|z_1| \leq \frac{\alpha}{\sqrt{d-2}}|R=1)f(r)dr \\ &\geq \int_0^{2\sqrt{d}} (1-\frac{2}{\alpha}e^{-\alpha^2/2})f(r)dr \\ &= (1-\frac{2}{\alpha}e^{-\alpha^2/2})\int_0^{2\sqrt{d}} f(r)dr \\ &= (1-\frac{2}{\alpha}e^{-\alpha^2/2})\mathrm{P}(0 \leq R \leq 2\sqrt{d}).\end{aligned}$$

Then, according to *Theorem 2* below, by setting $\beta = \sqrt{d}$, we have

$$\begin{aligned}\mathrm{P}(H) &= (1-\frac{2}{\alpha}e^{-\alpha^2/2})\mathrm{P}(0 \leq R \leq 2\sqrt{d}) \\ &\geq (1-\frac{2}{\alpha}e^{-\alpha^2/2})(1-3e^{-cd}).\end{aligned}$$

Q.E.D.

Theorem 1 *Given a unit spherical* $\{\mathbf{z} \in \mathbb{R}^d : ||\mathbf{z}||_2 = 1\}$, *we have* $\mathrm{P}(|z_1| \leq \frac{\alpha}{\sqrt{d-2}}) \geq 1 - \frac{2}{\alpha}e^{-\alpha^2/2}$ *for any* $\alpha \geq 1$ *and* $d \geq 4$.

Proof.

By symmetry, we just prove the case where $z_1 \geq 0$. Also, we only consider about the case where $\frac{\alpha}{\sqrt{d-2}} \leq 1$.

Let U denote the set $\{\mathbf{z} \in \mathbb{R}^d : ||\mathbf{z}||_2 = 1, z_1 \geq \frac{\alpha}{\sqrt{d-2}}\}$, and K denote the set $\{\mathbf{z} \in \mathbb{R}^d : ||\mathbf{z}||_2 = 1, z_1 \geq 0\}$. It suffices to prove that the surface of U area and the surface of K area in Fig.2 satisfy

$$\frac{surf(U)}{surf(K)} \leq \frac{2}{\alpha}e^{-\alpha^2/2},$$

where $surf(\cdot)$ stands for the surface area of a high dimensional geometry. Let $A(d)$ denote the surface area of a d-dimensional unit-radius ball. Then, we have

$$\begin{aligned}surf(U) &= \int_{\frac{\alpha}{\sqrt{d-2}}}^1 (1-z_1^2)^{\frac{d-2}{2}}A(d-1)dz_1 \\ &\leq \int_{\frac{\alpha}{\sqrt{d-2}}}^1 e^{-\frac{d-2}{2}z_1^2}A(d-1)dz_1 \\ &\leq \int_{\frac{\alpha}{\sqrt{d-2}}}^1 \frac{z_1\sqrt{d-2}}{\alpha}e^{-\frac{d-2}{2}z_1^2}A(d-1)dz_1 \\ &\leq \int_{\frac{\alpha}{\sqrt{d-2}}}^\infty \frac{z_1\sqrt{d-2}}{\alpha}e^{-\frac{d-2}{2}z_1^2}A(d-1)dz_1 \\ &= \frac{A(d-1)}{\alpha\sqrt{d-2}}e^{-\alpha^2/2}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}surf(K) &= \int_0^1 (1-z_1^2)^{\frac{d-2}{2}}A(d-1)dz_1 \\ &\geq \int_0^{\frac{1}{\sqrt{d-2}}} (1-z_1^2)^{\frac{d-2}{2}}A(d-1)dz_1 \\ &\geq \frac{1}{\sqrt{d-2}}(1-\frac{1}{d-2})^{\frac{d-2}{2}}A(d-1).\end{aligned}$$

Considering the fact that $(1-x)^a \geq 1-ax$ for any $a \geq 1$ and $0 \leq x \leq 1$, we have

$$\begin{aligned}surf(K) &\geq \frac{1}{\sqrt{d-2}}(1-\frac{1}{d-2})^{\frac{d-2}{2}}A(d-1) \\ &\geq \frac{1}{\sqrt{d-2}}(1-\frac{1}{d-2}\frac{d-2}{2})A(d-1) \\ &= \frac{A(d-1)}{2\sqrt{d-2}}.\end{aligned}$$

Accordingly,

$$\frac{surf(U)}{surf(K)} \leq \frac{\frac{A(d-1)}{\alpha\sqrt{d-2}}e^{-\alpha^2/2}}{\frac{A(d-1)}{2\sqrt{d-2}}} = \frac{2}{\alpha}e^{-\alpha^2/2}.$$

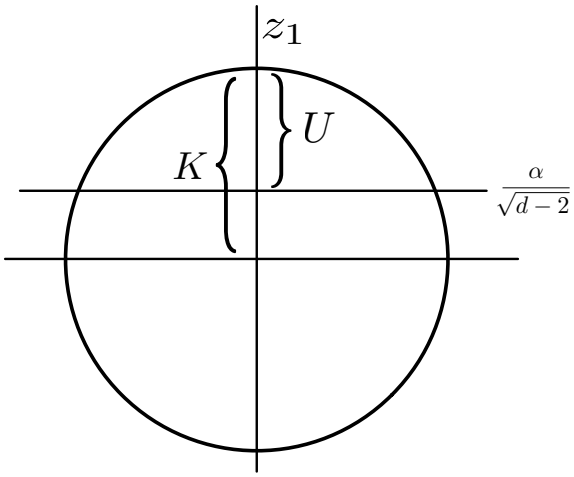

Figure 2: Diagram for *Theorem 1*.

Q.E.D.

Theorem 2 (Gaussian Annulus Theorem [2]) *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq ||\mathbf{z}||_2 \leq \sqrt{d} + \beta$, where c is a fixed positive constant.*

That is to say, given $\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_d)$, $\beta \leq \sqrt{d}$, and a constant $c > 0$, we have

$$\mathrm{P}(\sqrt{d} - \beta \leq ||\mathbf{z}||_2 \leq \sqrt{d} + \beta) \geq (1 - 3e^{-c\beta^2}).$$

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] John Hopcroft and Ravi Kannan. *Foundations of Data Science*. 2014. 3

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1