Supplementary Material for Perceptual Generative Autoencoders

Zijun Zhang[1] **Ruixiang Zhang**[2] **Zongpeng Li**[3] **Yoshua Bengio**[2] **Liam Paull**[2]

A. Notations

Table 1. Notations and definitions

f_ϕ/g_θ	encoder/decoder of an autoencoder
h	$h = f_\phi \circ g_\theta$
ϕ/θ	parameters of the encoder/decoder
D/H	dimensionality of the data/latent space
$\mathcal{D}$	distribution of data samples denoted by $\mathbf{x}$
$\mathcal{H}$	distribution of $f_\phi(\mathbf{x})$ for $\mathbf{x} \sim \mathcal{D}$
$\hat{\mathcal{D}}$	distribution of $\hat{\mathbf{x}} = g_\theta(f_\phi(\mathbf{x}))$ for $\mathbf{x} \sim \mathcal{D}$
$\hat{\mathcal{H}}$	distribution of $\hat{\mathbf{z}} = h(\mathbf{z})$ for $\mathbf{z} \sim \mathcal{H}$
$\widetilde{\mathcal{D}}$	distribution of $g_\theta(\mathbf{z})$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\widetilde{\mathcal{H}}$	distribution of $h(\mathbf{z})$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
L_r	standard reconstruction loss of the autoencoder
$L_{lr,\mathcal{N}}^{\phi}$	latent reconstruction loss of PGA for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, minimized w.r.t. ϕ
$L_{lr,\mathcal{H}}^{\phi}$	latent reconstruction loss of PGA for $\mathbf{z} \sim \mathcal{H}$, minimized w.r.t. ϕ
L_{nll}^{ϕ}	part of the negative log-likelihood loss of LPGA, minimized w.r.t. ϕ
L_{nll}^{θ}	part of the negative log-likelihood loss of LPGA, minimized w.r.t. θ
L_{vr}	VAE reconstruction loss of VPGA
L_{vkl}	VAE KL-divergence loss of VPGA
L_{vae}	$L_{vae} = L_{vr} + L_{vkl}$, VAE loss of VPGA

B. Proofs

B.1. Theorem 1

Proof sketch. We first show that any different $\mathbf{x}$'s generated by g_θ are mapped to different $\mathbf{z}$'s by f_ϕ. Let $\mathbf{x}_1 = g_\theta(\mathbf{z}_1)$, $\mathbf{x}_2 = g_\theta(\mathbf{z}_2)$, and $\mathbf{x}_1 \neq \mathbf{x}_2$. Since f_ϕ has sufficient capacity and Eq. (2) is minimized, we have $f_\phi(\mathbf{x}_1) = \mathbb{E}[\mathbf{z}_1|\mathbf{x}_1]$ and $f_\phi(\mathbf{x}_2) = \mathbb{E}[\mathbf{z}_2|\mathbf{x}_2]$. By assumption, $f_\phi(\mathbf{x}_1) \in Z(\mathbf{x}_1)$ and $f_\phi(\mathbf{x}_2) \in Z(\mathbf{x}_2)$. Therefore, since $Z(\mathbf{x}_1) \cap Z(\mathbf{x}_2) = \varnothing$, we have $f_\phi(\mathbf{x}_1) \neq f_\phi(\mathbf{x}_2)$.

For $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, denote the distributions of $g_\theta(\mathbf{z})$ and $h(\mathbf{z})$, respectively, by $\widetilde{\mathcal{D}}$ and $\widetilde{\mathcal{H}}$. We then consider the case where $\widetilde{\mathcal{D}}$ and $\hat{\mathcal{D}}$ are discrete distributions. If $g_\theta(\mathbf{z}) \nsim \hat{\mathcal{D}}$, then there exists an $\mathbf{x}$ that is generated by g_θ, such that $p_{\widetilde{\mathcal{H}}}(f_\phi(\mathbf{x})) = p_{\widetilde{\mathcal{D}}}(\mathbf{x}) \neq p_{\hat{\mathcal{D}}}(\mathbf{x}) = p_{\hat{\mathcal{H}}}(f_\phi(\mathbf{x}))$, contradicting that $h(\mathbf{z}) \sim \hat{\mathcal{H}}$. The result still holds when $\widetilde{\mathcal{D}}$ and $\hat{\mathcal{D}}$ approach continuous distributions, in which case $\widetilde{\mathcal{D}} = \hat{\mathcal{D}}$ almost everywhere. □

B.2. Proposition 1

Proof. Let $\mathbf{J}(\mathbf{z}) = \partial h(\mathbf{z}) / \partial \mathbf{z}$, $\mathbf{P} = \begin{bmatrix} \delta_1 & \delta_2 & \cdots & \delta_H \end{bmatrix}$, and $\hat{\mathbf{P}} = \mathbf{J}(\mathbf{z}) \mathbf{P} = \begin{bmatrix} \hat{\delta}_1 & \hat{\delta}_2 & \cdots & \hat{\delta}_H \end{bmatrix}$, where $\Delta = \{\delta_1, \delta_2, \ldots, \delta_H\}$ is an orthogonal set of H-dimensional vectors. Since $\det\left(\hat{\mathbf{P}}\right) = \det(\mathbf{J}(\mathbf{z})) \det(\mathbf{P})$, we have

$$\log|\det(\mathbf{J}(\mathbf{z}))| = \log\left|\det\left(\hat{\mathbf{P}}\right)\right| - \log|\det(\mathbf{P})|. \tag{1}$$

By the geometric interpretation of determinants, the volume of the parallelotope spanned by Δ is

$$\mathrm{Vol}(\Delta) = |\det(\mathbf{P})| = \prod_{i \in [H]} \|\delta_i\|_2, \tag{2}$$

where $[H] = \{1, 2, \ldots, H\}$. While $\hat{\Delta} = \left\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_H\right\}$ is not necessarily an orthogonal set, an upper bound on $\mathrm{Vol}\left(\hat{\Delta}\right)$ can be derived in a similar fashion. Let $\hat{\Delta}_k = \left\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_k\right\}$, and a_k be the included angle between $\hat{\delta}_k$ and the plane spanned by $\hat{\Delta}_{k-1}$. We have

$$\begin{aligned} \mathrm{Vol}\left(\hat{\Delta}_2\right) &= \left\|\hat{\delta}_1\right\|_2 \left\|\hat{\delta}_2\right\|_2 \sin a_2, \\ \text{and } \mathrm{Vol}\left(\hat{\Delta}_k\right) &= \mathrm{Vol}\left(\hat{\Delta}_{k-1}\right) \left\|\hat{\delta}_k\right\|_2 \sin a_k. \end{aligned} \tag{3}$$

Given fixed $\left\|\hat{\delta}_k\right\|_2, \forall k \in [H]$, $\mathrm{Vol}\left(\hat{\Delta}_2\right)$ is maximized when $a_2 = \pi/2$, i.e., $\hat{\delta}_1$ and $\hat{\delta}_2$ are orthogonal; and $\mathrm{Vol}\left(\hat{\Delta}_k\right)$ is maximized when $\mathrm{Vol}\left(\hat{\Delta}_{k-1}\right)$ is maximized and $a_k = \pi/2$. By induction on k, we can conclude that $\mathrm{Vol}\left(\hat{\Delta}\right)$ is maximized when $\hat{\Delta} = \hat{\Delta}_H$ is an orthogonal set,

[1]University of Calgary, Canada [2]MILA, Université de Montréal, Canada [3]Wuhan University, China. Correspondence to: Zijun Zhang <zijun.zhang@ucalgary.ca>, Ruixiang Zhang <ruixiang.zhang@umontreal.ca>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

and therefore

$$\mathrm{Vol}\left(\hat{\Delta}\right)=\left|\det\left(\hat{\mathbf{P}}\right)\right|\leq\prod_{i\in[H]}\left\|\hat{\delta}_i\right\|_2. \tag{4}$$

Combining Eq. (1) with Eqs. (2) and (4), we obtain

$$\log\left|\det\left(\mathbf{J}\left(\mathbf{z}\right)\right)\right|\leq\sum_{i\in[H]}\left(\log\left\|\hat{\delta}_i\right\|_2-\log\left\|\delta_i\right\|_2\right). \tag{5}$$

We proceed by randomizing Δ. Let $\Delta_k=\{\delta_1,\delta_2,\dots,\delta_k\}$. We inductively construct an orthogonal set, $\Delta=\Delta_H$. In step 1, δ_1 is sampled from $\mathcal{S}\left(\epsilon\right)$, a uniform distribution on a $(H-1)$-sphere of radius ϵ, $S\left(\epsilon\right)$, centered at the origin of an H-dimensional space. In step k, δ_k is sampled from $\mathcal{S}\left(\epsilon;\Delta_{k-1}\right)$, a uniform distribution on an $(H-k)$-sphere, $S\left(\epsilon;\Delta_{k-1}\right)$, in the orthogonal complement of the space spanned by Δ_{k-1}. Step k is repeated until H mutually orthogonal vectors are obtained.

Obviously, when $k=H-1$, for all $j>k$ and $j\leq H$, $p\left(\delta_j|\Delta_k\right)=p\left(\delta_j|\Delta_{H-1}\right)=\mathcal{S}\left(\delta_j|\epsilon;\Delta_{H-1}\right)=\mathcal{S}\left(\delta_j|\epsilon;\Delta_k\right)$. When $1\leq k<H$, assuming for all $j>k$ and $j\leq H$, $p\left(\delta_j|\Delta_k\right)=\mathcal{S}\left(\delta_j|\epsilon;\Delta_k\right)$, we get

$$p\left(\delta_j|\Delta_{k-1}\right)=\int\limits_{S\left(\epsilon;\Delta_{k-1}\cup\{\delta_j\}\right)}p\left(\delta_k|\Delta_{k-1}\right)p\left(\delta_j|\Delta_k\right)d\delta_k, \tag{6}$$

where $S\left(\epsilon;\Delta_{k-1}\cup\{\delta_j\}\right)$ is in the orthogonal complement of the space spanned by $\Delta_{k-1}\cup\{\delta_j\}$. Since $p\left(\delta_k|\Delta_{k-1}\right)$ is a constant on $S\left(\delta_k|\epsilon;\Delta_{k-1}\right)$, and $S\left(\epsilon;\Delta_{k-1}\cup\{\delta_j\}\right)\subset S\left(\epsilon;\Delta_{k-1}\right)$, $p\left(\delta_k|\Delta_{k-1}\right)$ is also a constant on $S\left(\epsilon;\Delta_{k-1}\cup\{\delta_j\}\right)$. In addition, $\delta_k\in S\left(\epsilon;\Delta_{k-1}\cup\{\delta_j\}\right)$ implies that $\delta_j\in S\left(\epsilon;\Delta_k\right)$, on which $p\left(\delta_j|\Delta_k\right)$ is also a constant. Then it follows from Eq. (6) that, for all $\delta_j\in S\left(\epsilon;\Delta_{k-1}\right)$, $p\left(\delta_j|\Delta_{k-1}\right)$ is a constant. Therefore, for all $j>k-1$ and $j\leq H$, $p\left(\delta_j|\Delta_{k-1}\right)=\mathcal{S}\left(\delta_j|\epsilon;\Delta_{k-1}\right)$. By backward induction on k, we conclude that the marginal probability density of δ_k, for all $k\in[H]$, is $p\left(\delta_k\right)=\mathcal{S}\left(\delta_k|\epsilon\right)$.

Since Eq. (5) holds for any randomly (as defined above) sampled Δ, we have

$$\begin{aligned}\log\left|\det\left(\mathbf{J}\left(\mathbf{z}\right)\right)\right|&\leq\mathbb{E}_{\Delta}\left[\sum_{i\in[H]}\left(\log\left\|\hat{\delta}_i\right\|_2-\log\left\|\delta_i\right\|_2\right)\right]\\&=H\mathbb{E}_{\delta\sim\mathcal{S}(\epsilon)}\left[\log\left\|\hat{\delta}\right\|_2-\log\left\|\delta\right\|_2\right].\end{aligned} \tag{7}$$

If h is a multiple of the identity function around $\mathbf{z}$, then $\mathbf{J}\left(\mathbf{z}\right)=C\mathbf{I}$, where $C\in\mathbb{R}$ is a constant. In this case, $\hat{\Delta}$ becomes an orthogonal set as Δ, and therefore the inequalities in Eqs. (4), (5), and (7) become tight. Furthermore, it is straightforward to extend the above result to the case $\delta\sim\mathcal{N}\left(\mathbf{0},\epsilon^2\mathbf{I}\right)$, considering that $\mathcal{N}\left(\mathbf{0},\epsilon^2\mathbf{I}\right)$ is a mixture of $\mathcal{S}\left(\epsilon\right)$ with different ϵ's.

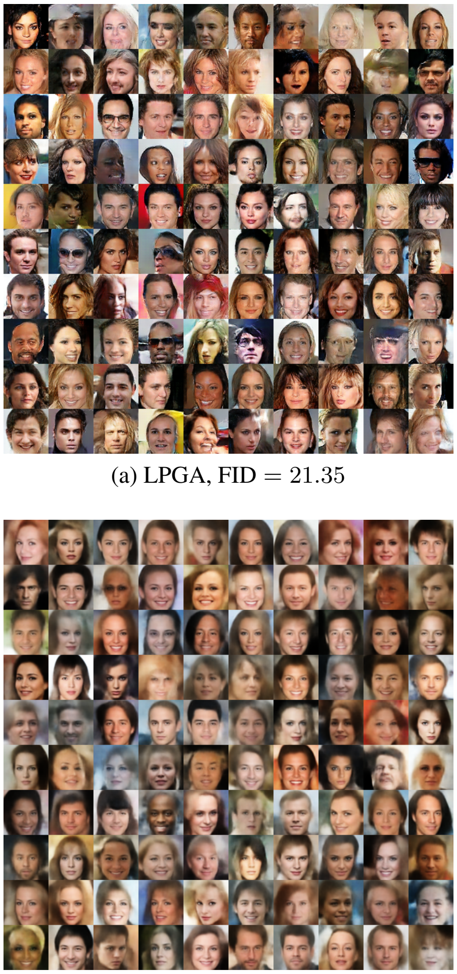
The Taylor expansion of h around $\mathbf{z}$ gives

$$h\left(\mathbf{z}+\delta\right)=h\left(\mathbf{z}\right)+\mathbf{J}\left(\mathbf{z}\right)\delta+\mathcal{O}\left(\delta^2\right). \tag{8}$$

Therefore, for $\delta\to\mathbf{0}$ or $\epsilon\to0$, we have $\hat{\delta}=\mathbf{J}\left(\mathbf{z}\right)\delta=h\left(\mathbf{z}+\delta\right)-h\left(\mathbf{z}\right)$. The result follows. □

C. More Results on CelebA

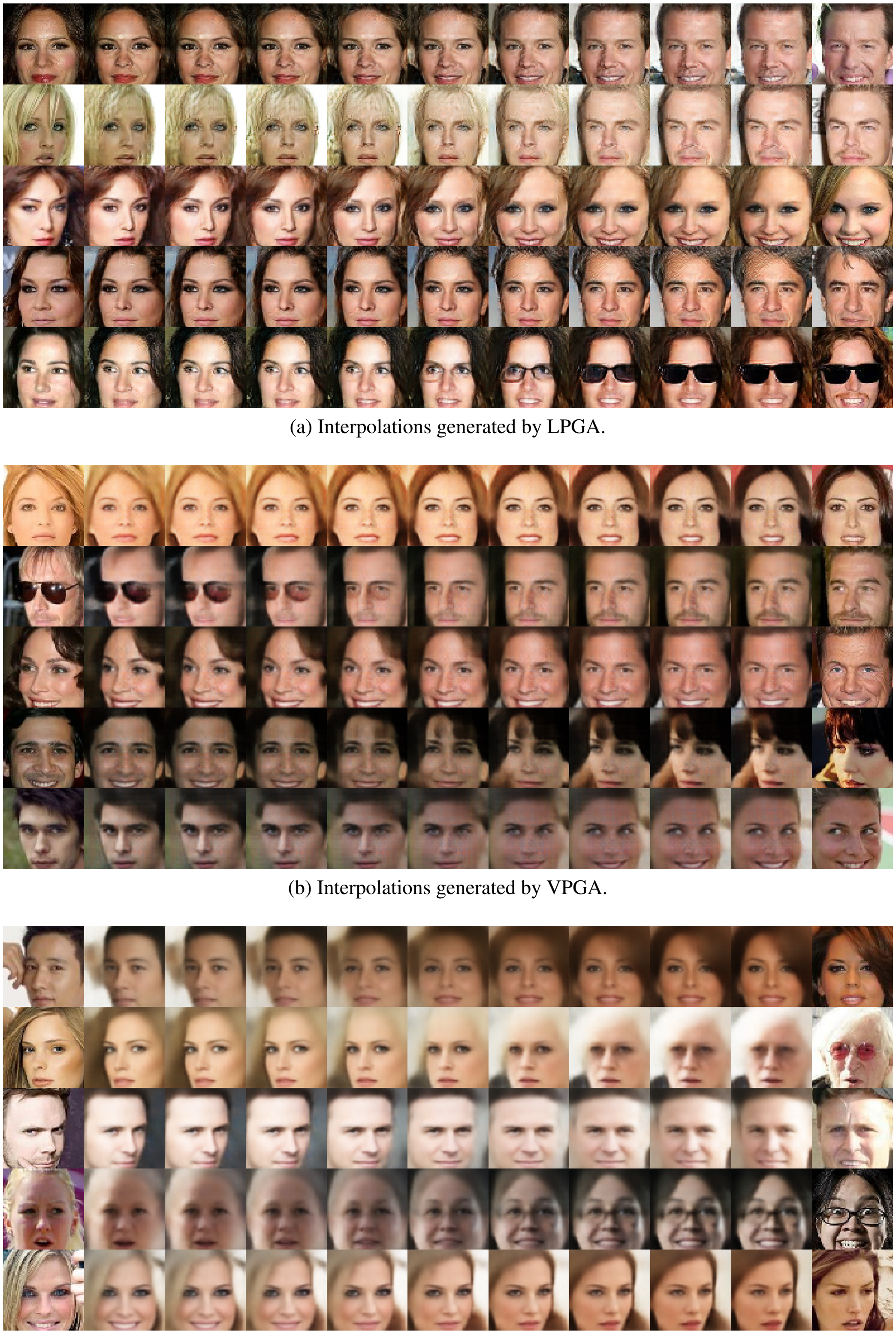
In Fig. 1, we compare the generated samples and FID scores of LPGA and VAE on 140x140 crops. In this experiment, we use the full DCGAN architecture (i.e., 128 filters for the first convolutional layer) for both LPGA and VAE. Other hyperparameter settings remain the same as for 108x108 crops. In Fig. 2, we show latent space interpolations of CelebA samples.



(a) LPGA, FID $=21.35$

(b) VAE, FID $=54.25$

Figure 1. Random CelebA (140x140 crops) samples generated by LPGA and VAE.



(a) Interpolations generated by LPGA.

(b) Interpolations generated by VPGA.

(c) Interpolations generated by VAE.

Figure 2. Latent space interpolations on CelebA.