Perceptual Generative Autoencoders

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

Abstract

Modern generative models are usually designed to match target distributions directly in the data space, where the intrinsic dimension of data can be much lower than the ambient dimension. We argue that this discrepancy may contribute to the difficulties in training generative models. We therefore propose to map both the generated and target distributions to a latent space using the encoder of a standard autoencoder, and train the generator (or decoder) to match the target distribution in the latent space. Specifically, we enforce the consistency in both the data space and the latent space with theoretically justified data and latent reconstruction losses. The resulting generative model, which we call a *perceptual generative autoencoder (PGA)*, is then trained with a maximum likelihood or variational autoencoder (VAE) objective. With maximum likelihood, PGAs generalize the idea of reversible generative models to unrestricted neural network architectures and arbitrary number of latent dimensions. When combined with VAEs, PGAs substantially improve over the baseline VAEs in terms of sample quality. Compared to other autoencoder-based generative models using simple priors, PGAs achieve state-of-the-art FID scores on CIFAR-10 and CelebA.

1. Introduction

Recent years have witnessed great interest in generative models, mainly due to the success of generative adversarial networks (GANs) (Goodfellow et al., 2014; Radford et al., 2016; Karras et al., 2018; Brock et al., 2019). Despite their prevalence, the adversarial nature of GANs can lead to a number of challenges, such as unstable training dynamics and mode collapse. Since the advent of GANs, substantial efforts have been devoted to addressing these challenges (Salimans et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018), while non-adversarial approaches that are free of these issues have also gained attention. Examples include variational autoencoders (VAEs) (Kingma & Welling, 2014), reversible generative models (Dinh et al., 2014; 2017; Kingma & Dhariwal, 2018), and Wasserstein autoencoders (WAEs) (Tolstikhin et al., 2018).

However, non-adversarial approaches often have significant limitations. For instance, VAEs tend to generate blurry samples, while reversible generative models require restricted neural network architectures or solving neural differential equations (Grathwohl et al., 2019). Furthermore, to use the change of variable formula, the latent space of a reversible model must have the same dimension as the data space, which is unreasonable considering that real-world, high-dimensional data (e.g., images) tends to lie on low-dimensional manifolds, and thus results in redundant latent dimensions and variability. Intriguingly, recent research (Arjovsky et al., 2017; Dai & Wipf, 2019) suggests that the discrepancy between the intrinsic and ambient dimensions of data also contributes to the difficulties in training GANs and VAEs.

In this work, we present a novel framework for training autoencoder-based generative models, with non-adversarial losses and unrestricted neural network architectures. Given a standard autoencoder and a target data distribution, instead of matching the target distribution in the data space, we map both the generated and target distributions to a latent space using an encoder, while also minimizing the divergence between the mapped distributions. We prove, under mild assumptions, that by minimizing a form of latent reconstruction error, matching the target distribution in the latent space implies matching it in the data space. We call this framework *perceptual generative autoencoder (PGA)*. We show that PGAs enable training generative autoencoders with maximum likelihood, without restrictions on architectures or latent dimensionalities. In addition, when combined with VAEs, PGAs can generate sharper samples than vanilla VAEs.[1]

We summarize our main contributions as follows:

[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 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Code is available at https://github.com/zj10/PGA.

- A training framework, PGA, for generative autoencoders is developed to match the target distribution in the latent space, which, we prove, ensures correct matching in data space.
- We combine PGA with the maximum likelihood objective, and remove the restrictions of reversible (flow-based) generative models on neural network architectures and latent dimensionalities.
- We combine PGA with the VAE objective, solving the VAE's issue of blurry samples without introducing any auxiliary models or sophisticated model architectures.

2. Related Work

Autoencoder-based generative models are trained by minimizing an data reconstruction loss with regularizations. As an early approach, denoising autoencoders (DAEs) (Vincent et al., 2008) are trained to recover the original input from an intentionally corrupted input. Then a generative model can be obtained by sampling from a Markov chain (Bengio et al., 2013). To sample from a decoder directly, most recent approaches resort to mapping a simple prior distribution to a data distribution using the decoder. For instance, variational autoencoders (VAEs) directly match data distributions by maximizing the evidence lower bound. In contrast, adversarial autoencoders (AAEs) (Makhzani et al., 2016) and Wasserstein autoencoders (WAEs) (Tolstikhin et al., 2018) work in the latent space to match the aggregated posterior with the prior, either by adversarial training or by minimizing their Wasserstein distance. Inspired by AAEs and WAEs, we develop a principled approach to matching data distributions in the latent space, aiming to improve the generative performance of AAEs and WAEs (Rubenstein et al., 2018), as well as that of VAEs (Rezende & Viola, 2018; Dai & Wipf, 2019). While previous work has explored the use of perceptual loss for a similar purpose (Hou et al., 2017), it relies on a VGG net pre-trained on ImageNet and provides no theoretical guarantees. In our work, the encoder of an autoencoder is jointly trained, such that matching the target distribution in the latent space guarantees the matching in the data space.

In a different line of work, reversible generative models (Dinh et al., 2014; 2017) are developed to enable exact inference. Consequently, by the change of variables theorem, the likelihood of each data sample can be exactly computed and optimized. Recent work shows that they are capable of generating realistic images (Kingma & Dhariwal, 2018). However, to avoid expensive Jacobian determinant computations, reversible models can only be composed of restricted transformations, rather than general neural network architectures. While this restriction can be relaxed by utilizing recently developed neural ordinary differential equations (Chen et al., 2018; Grathwohl et al., 2019), they still rely on a shared dimensionality between the latent and data spaces, which remains an unnatural restriction. In this work, we use the proposed training framework to trade exact inference for unrestricted neural network architectures and arbitrary latent dimensionalities, generalizing maximum likelihood training to autoencoder-based models.

3. Methods

3.1. Perceptual Generative Model

Let $f_\phi : \mathbb{R}^D \to \mathbb{R}^H$ be the encoder parameterized by ϕ, and $g_\theta : \mathbb{R}^H \to \mathbb{R}^D$ be the decoder parameterized by θ. Our goal is to obtain a decoder-based generative model, which maps a simple prior distribution to a target data distribution, $\mathcal{D}$. Throughout this paper, we use $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the prior distribution. This section will introduce several related but different distributions, which are illustrated in Fig. 1a. A summary of notations is provided in the supplementary material.

For $\mathbf{z} \in \mathbb{R}^H$, the output of the decoder, $g_\theta(\mathbf{z})$, lies in a manifold that is at most H-dimensional. Therefore, if we train the autoencoder to minimize

$$L_r = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2\right], \tag{1}$$

where $\hat{\mathbf{x}} = g_\theta(f_\phi(\mathbf{x}))$, then $\hat{\mathbf{x}}$ can be seen as a projection of the input data, $\mathbf{x}$, onto the manifold of $g_\theta(\mathbf{z})$. Let $\hat{\mathcal{D}}$ denote the reconstructed data distribution, i.e., $\hat{\mathbf{x}} \sim \hat{\mathcal{D}}$. Given enough capacity of the encoder, $\hat{\mathcal{D}}$ is the best approximation to $\mathcal{D}$ (in terms of ℓ_2-distance), that we can obtain from the decoder, and thus can serve as a surrogate target distribution for training the decoder-based generative model.

Due to the difficulty in directly matching the generated distribution with the data-space target distribution, $\hat{\mathcal{D}}$, we reuse the encoder to map $\hat{\mathcal{D}}$ to a latent-space target distribution, $\hat{\mathcal{H}}$. We then transform the problem of matching $\hat{\mathcal{D}}$ in the data space into matching $\hat{\mathcal{H}}$ in the latent space. In other words, we aim to ensure that for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, if $f_\phi(g_\theta(\mathbf{z})) \sim \hat{\mathcal{H}}$, then $g_\theta(\mathbf{z}) \sim \hat{\mathcal{D}}$. In the following, we define $h = f_\phi \circ g_\theta$ for notational convenience.

To this end, we minimize the following latent reconstruction loss w.r.t. ϕ:

$$L_{lr,\mathcal{N}}^{\phi} = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|h(\mathbf{z}) - \mathbf{z}\|_2^2\right]. \tag{2}$$

Let $Z(\mathbf{x})$ be the set of all $\mathbf{z}$'s that are mapped to the same $\mathbf{x}$ by g_θ, we have the following theorem:

Theorem 1. *Assuming $\mathbb{E}[\mathbf{z}|\mathbf{x}] \in Z(\mathbf{x})$ for all $\mathbf{x}$ generated by g_θ, and sufficient capacity of f_ϕ; for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, if Eq. (2) is minimized and $h(\mathbf{z}) \sim \hat{\mathcal{H}}$, then $g_\theta(\mathbf{z}) \sim \hat{\mathcal{D}}$.*

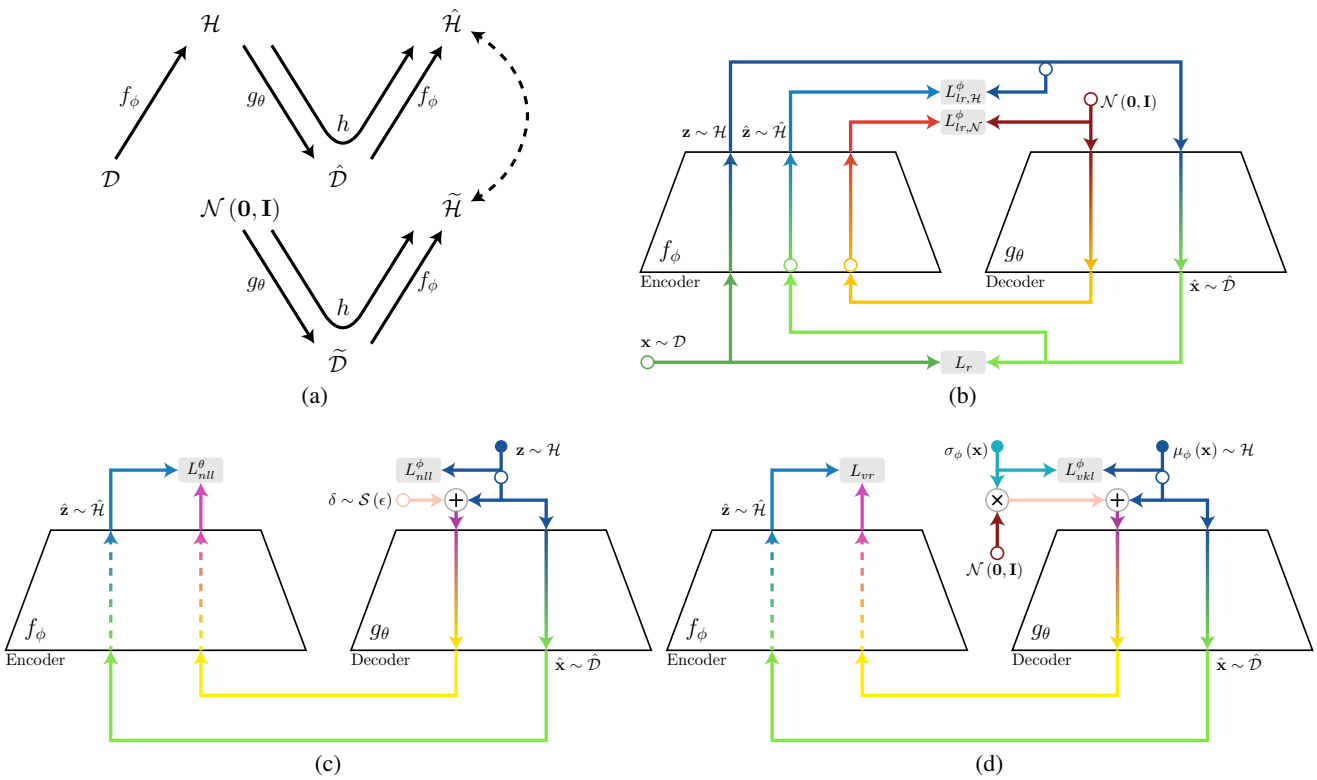


Figure 1. Illustration of the training process of PGAs. (a) shows the distributions involved in training PGAs, where the dashed arrow points to the two latent-space distributions to be matched. The overall loss function consists of (b) the basic PGA losses, and either (c) the LPGA-specific losses or (d) the VPGA-specific losses. Circles indicate where the gradient is truncated, and dashed lines indicate where the gradient is ignored when updating parameters.

We defer the proof to the supplementary material. Note that Theorem 1 requires that different $\mathbf{x}$'s generated by g_θ (from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{H}$) are mapped to different $\mathbf{z}$'s by f_ϕ. In theory, minimizing Eq. (2) would suffice, since $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is supported on the whole $\mathbb{R}^H$. However, there can be $\mathbf{z}$'s with low probabilities in $\mathcal{N}(\mathbf{0}, \mathbf{I})$, but with high probabilities in $\mathcal{H}$ that are not well covered by Eq. (2). Therefore, it is sometimes helpful to minimize another latent reconstruction loss on $\mathcal{H}$:

$$L^{\phi}_{lr,\mathcal{H}} = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\mathcal{H}}\left[\|h(\mathbf{z}) - \mathbf{z}\|_2^2\right]. \tag{3}$$

In practice, we observe that $L^{\phi}_{lr,\mathcal{H}}$ is often small without explicit minimization, which we attribute to its consistency with the minimization of L_r. Moreover, minimizing the latent reconstruction losses w.r.t. θ is not required by Theorem 1, and it degrades the performance empirically. In addition, the use of ℓ_2-norm in the reconstruction losses is not a necessity, and the framework can be easily extended to other norm definitions.

By Theorem 1, the problem of training the generative model reduces to training h to map $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\hat{\mathcal{H}}$, which we refer to as the perceptual generative model. The basic loss function of PGAs is given by

$$L_{pga} = L_r + \alpha L^{\phi}_{lr,\mathcal{N}} + \beta L^{\phi}_{lr,\mathcal{H}}, \tag{4}$$

where α and β are hyperparameters to be tuned. Eq. (4) is also illustrated in Fig. 1b.

In the subsequent subsections, we present a maximum likelihood approach, as well as a VAE-based approach to train the perceptual generative model. To build intuition before delving into the details, we note that both of these two approaches work by attracting the latent representations of data samples to the origin, while expanding the volume occupied by each sample in the latent space. These two tendencies together push $\mathcal{H}$ closer to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, such that $\tilde{\mathcal{H}}$ matches $\hat{\mathcal{H}}$. This observation further leads to a unified view of the two approaches.

3.2. A Maximum Likelihood Approach

We first assume the invertibility of h. For $\hat{\mathbf{x}} \sim \hat{\mathcal{D}}$, let $\hat{\mathbf{z}} = f_\phi(\hat{\mathbf{x}}) = h(\mathbf{z}) \sim \hat{\mathcal{H}}$. We can train h directly with maximum likelihood using the change of variables formula

as

$$\mathbb{E}_{\hat{\mathbf{z}}\sim\hat{\mathcal{H}}}\left[\log p\left(\hat{\mathbf{z}}\right)\right]=\mathbb{E}_{\mathbf{z}\sim\mathcal{H}}\left[\log p\left(\mathbf{z}\right)-\log\left|\det\left(\frac{\partial h\left(\mathbf{z}\right)}{\partial\mathbf{z}}\right)\right|\right], \tag{5}$$

where $p\left(\mathbf{z}\right)$ is the prior distribution, $\mathcal{N}\left(\mathbf{0},\mathbf{I}\right)$. Since the actual generative model to be trained is the decoder (parameterized by θ), we would like to maximize Eq. (5) only w.r.t. θ. However, directly optimizing the first term in Eq. (5) requires computing $\mathbf{z}=h^{-1}\left(\hat{\mathbf{z}}\right)$, which is usually unknown. Nevertheless, for $\hat{\mathbf{z}}\sim\hat{\mathcal{H}}$, we have $h^{-1}\left(\hat{\mathbf{z}}\right)=f_{\phi}\left(\mathbf{x}\right)$ and $\mathbf{x}\sim\mathcal{D}$, and thus we can minimize the following loss function w.r.t. ϕ instead:

$$L_{nll}^{\phi}=-\mathbb{E}_{\mathbf{z}\sim\mathcal{H}}\left[\log p\left(\mathbf{z}\right)\right]=\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\|f_{\phi}\left(\mathbf{x}\right)\right\|_{2}^{2}\right]. \tag{6}$$

To avoid computing the Jacobian in the second term of Eq. (5), which is slow for unrestricted architectures, we approximate the Jacobian determinant and derive a loss function to be minimized w.r.t. θ:

$$\begin{aligned}L_{nll}^{\theta}&=\frac{H}{2}\mathbb{E}_{\mathbf{z}\sim\mathcal{H},\delta\sim\mathcal{S}(\epsilon)}\left[\log\frac{\left\|h\left(\mathbf{z}+\delta\right)-h\left(\mathbf{z}\right)\right\|_{2}^{2}}{\left\|\delta\right\|_{2}^{2}}\right]\\&\approx\mathbb{E}_{\mathbf{z}\sim\mathcal{H}}\left[\log\left|\det\left(\frac{\partial h\left(\mathbf{z}\right)}{\partial\mathbf{z}}\right)\right|\right],\end{aligned} \tag{7}$$

where $\mathcal{S}\left(\epsilon\right)$ can be either $\mathcal{N}\left(\mathbf{0},\epsilon^{2}\mathbf{I}\right)$, or a uniform distribution on a small $(H-1)$-sphere of radius ϵ centered at the origin. The latter choice is expected to introduce slightly less variance. Note that if we also minimize Eq. (7) w.r.t. ϕ, the encoder will be trained to ignore the difference between $g_{\theta}\left(\mathbf{z}+\delta\right)$ and $g_{\theta}\left(\mathbf{z}\right)$, in which case Theorem 1 no longer holds.

Eqs. (6) and (7) are illustrated in Fig. 1c. We show below that the approximation in Eq. (7) gives an upper bound when $\epsilon\rightarrow 0$.

Proposition 1. *For $\epsilon\rightarrow 0$,*

$$\log\left|\det\left(\frac{\partial h\left(\mathbf{z}\right)}{\partial\mathbf{z}}\right)\right|\leq\frac{H}{2}\mathbb{E}_{\delta\sim\mathcal{S}(\epsilon)}\left[\log\frac{\left\|h\left(\mathbf{z}+\delta\right)-h\left(\mathbf{z}\right)\right\|_{2}^{2}}{\left\|\delta\right\|_{2}^{2}}\right]. \tag{8}$$

The inequality is tight if h is a multiple of the identity function around $\mathbf{z}$.

We defer the proof to the supplementary material. We note that while the approximation in Eq. (7) is derived from the change of variables formula, there is no direct usage of the latter. As a result, the invertibility of h is not required by the resulting method. Indeed, when h is invertible at some point $\mathbf{z}$, the latent reconstruction loss ensures that h is close to the identity function around $\mathbf{z}$, and hence the tightness of the upper bound in Eq. (8). Otherwise, when h is not invertible at some $\mathbf{z}$, the logarithm of the Jacobian determinant at $\mathbf{z}$ becomes infinite, in which case Eq. (5) cannot be optimized. Nevertheless, since $\left\|h\left(\mathbf{z}+\delta\right)-h\left(\mathbf{z}\right)\right\|_{2}^{2}$ is unlikely to be zero if the model is properly initialized, the approximation in Eq. (7) remains finite, and thus can be optimized regardless.

To summarize, we train the autoencoder to obtain a generative model by minimizing the following loss function:

$$L_{lpga}=L_{pga}+\gamma\left(L_{nll}^{\phi}+L_{nll}^{\theta}\right). \tag{9}$$

We refer to this approach as maximum likelihood PGA (LPGA).

3.3. A VAE-based Approach

The original VAE is trained by maximizing the evidence lower bound on $\log p\left(\mathbf{x}\right)$ as

$$\begin{aligned}&\log p\left(\mathbf{x}\right)\\&\geq\log p\left(\mathbf{x}\right)-\mathbb{KL}(q\left(\mathbf{z}'|\mathbf{x}\right)\ ||\ p\left(\mathbf{z}'|\mathbf{x}\right))\\&=\mathbb{E}_{\mathbf{z}'\sim q(\mathbf{z}'|\mathbf{x})}\left[\log p\left(\mathbf{x}|\mathbf{z}'\right)\right]-\mathbb{KL}(q\left(\mathbf{z}'|\mathbf{x}\right)\ ||\ p\left(\mathbf{z}'\right)),\end{aligned} \tag{10}$$

where $p\left(\mathbf{x}|\mathbf{z}'\right)$ is modeled with the decoder, and $q\left(\mathbf{z}'|\mathbf{x}\right)$ is modeled with the encoder. Note that $\mathbf{z}'$ denotes the stochastic version of $\mathbf{z}$, whereas $\mathbf{z}$ remains deterministic for the basic PGA losses in Eqs. (2) and (3). In our case, we would like to modify Eq. (10) in a way that helps maximize $\log p\left(\hat{\mathbf{z}}\right)$, where $\hat{\mathbf{z}}=h\left(\mathbf{z}\right)$. Therefore, we replace $p\left(\mathbf{x}|\mathbf{z}'\right)$ on the r.h.s. of Eq. (10) with $p\left(\hat{\mathbf{z}}|\mathbf{z}'\right)$, and derive a lower bound on $\log p\left(\hat{\mathbf{z}}\right)$ as

$$\begin{aligned}&\log p\left(\hat{\mathbf{z}}\right)\\&\geq\log p\left(\hat{\mathbf{z}}\right)-\mathbb{KL}(q\left(\mathbf{z}'|\mathbf{x}\right)\ ||\ p\left(\mathbf{z}'|\hat{\mathbf{z}}\right))\\&=\mathbb{E}_{\mathbf{z}'\sim q(\mathbf{z}'|\mathbf{x})}\left[\log p\left(\hat{\mathbf{z}}|\mathbf{z}'\right)\right]-\mathbb{KL}(q\left(\mathbf{z}'|\mathbf{x}\right)\ ||\ p\left(\mathbf{z}'\right)).\end{aligned} \tag{11}$$

Similar to the original VAE, we make the assumption that $q\left(\mathbf{z}'|\mathbf{x}\right)$ and $p\left(\hat{\mathbf{z}}|\mathbf{z}'\right)$ are Gaussian; i.e., $q\left(\mathbf{z}'|\mathbf{x}\right)=\mathcal{N}\left(\mathbf{z}'\ \middle|\ \mu_{\phi}\left(\mathbf{x}\right),\mathrm{diag}\left(\sigma_{\phi}^{2}\left(\mathbf{x}\right)\right)\right)$, and $p\left(\hat{\mathbf{z}}|\mathbf{z}'\right)=\mathcal{N}\left(\hat{\mathbf{z}}|\mu_{\theta,\phi}\left(\mathbf{z}'\right),\sigma^{2}\mathbf{I}\right)$. Here, $\mu_{\phi}\left(\cdot\right)=f_{\phi}\left(\cdot\right)$, $\mu_{\theta,\phi}\left(\cdot\right)=h\left(\cdot\right)$, and $\sigma>0$ is a tunable scalar. Note that if σ is fixed, the first term on the r.h.s. of Eq. (11) has a trivial maximum, where $\mathbf{z}$, $\hat{\mathbf{z}}$, and $\mu_{\theta,\phi}\left(\mathbf{z}'\right)$ are all close to zero. To circumvent this, we set σ proportional to the ℓ_{2}-norm of $\mathbf{z}$.

The VAE variant is trained by minimizing

$$\begin{aligned}&L_{vae}=L_{vr}+L_{vkl}^{\phi}\\&=-\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{z}'\sim q(\mathbf{z}'|\mathbf{x})}\left[\log p\left(\hat{\mathbf{z}}|\mathbf{z}'\right)\right]-\mathbb{KL}(q\left(\mathbf{z}'|\mathbf{x}\right)\ ||\ p\left(\mathbf{z}'\right))\right],\end{aligned} \tag{12}$$

where L_{vr} and L_{vkl}^{ϕ} correspond, respectively, to the reconstruction and KL divergence losses of VAE, as illustrated in Fig. 1d. In L_{vr}, while the gradient through $\sigma_{\phi}^{2}\left(\mathbf{x}\right)$ remains unchanged, we ignore the gradient passed directly from L_{vr} to the encoder, due to a similar reason discussed for Eq. (7). Accordingly, the overall loss function is given by

$$L_{vpga}=L_{pga}+\eta L_{vae}. \tag{13}$$

We refer to this approach as variational PGA (VPGA).

3.4. A High-level View of the PGA Framework

We summarize what each loss term achieves, and explain from a high-level how they work together.

Data reconstruction loss (Eq. (1)): For Theorem 1 to hold, we need to use the reconstructed data distribution ($\hat{\mathcal{D}}$), instead of the original data distribution ($\mathcal{D}$), as the target distribution. Therefore, minimizing the data reconstruction loss ensures that the target distribution is close to the data distribution.

Latent reconstruction loss (Eqs. (2) and (3)): The encoder (f_ϕ) is reused to map data-space distributions to the latent space. As shown by Theorem 1, minimizing the latent reconstruction loss (w.r.t. the parameters of the encoder) ensures that if the generated distribution and the target distribution can be mapped to the same distribution ($\hat{\mathcal{H}}$) in the latent space by the encoder, then the generated distribution and the target distribution are the same.

Maximum likelihood loss (Eqs. (6) and (7)) or **VAE loss** (Eq. (12)): The decoder (g_θ) and encoder (f_ϕ) together can be considered as a perceptual generative model ($f_\phi \circ g_\theta$), which is trained to map $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to the latent-space target distribution ($\hat{\mathcal{H}}$) by minimizing either the maximum likelihood loss or the VAE loss.

The first loss allows to use the reconstructed data distribution as the target distribution. The second loss transforms the problem of matching the target distribution in the data space into matching it in the latent space. The latter problem is then solved by the third loss. Therefore, the three losses together ensure that the generated distribution is close to the data distribution.

3.5. A Unified Approach

While the loss functions of maximum likelihood and VAE seem completely different in their original forms, they share remarkable similarities when considered in the PGA framework (see Figs. 1c and 1d). Intuitively, observe that

$$L^{\phi}_{vkl} = L^{\phi}_{nll} + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_{i\in[H]} \left[\sigma^2_{\phi,i}(\mathbf{x}) - \log\left(\sigma^2_{\phi,i}(\mathbf{x})\right)\right], \tag{14}$$

which means both L^{ϕ}_{nll} and L^{ϕ}_{vkl} tend to attract the latent representations of data samples to the origin. In addition, by minimizing $\log|\det(\partial h(\mathbf{z})/\partial \mathbf{z})|$, L^{θ}_{nll} expands the volume occupied by each sample in the latent space, which can be also achieved by L_{vr} with the second term of Eq. (14).

More concretely, we observe that both L^{θ}_{nll} and L_{vr} are minimizing the difference between $h(\mathbf{z})$ and $h(\mathbf{z}+\delta')$, where δ' is some additive zero-mean noise. However, they differ in that the variance of δ' is fixed for L^{θ}_{nll}, but is trainable for L_{vr}; and the distance between $h(\mathbf{z})$ and $h(\mathbf{z}+\delta')$ are defined in two different ways. In fact, L_{vr} is a squared ℓ_2-distance derived from the Gaussian assumption on $\hat{\mathbf{z}}$, whereas L^{θ}_{nll} can be derived similarly by assuming that $d^H = \|\hat{\mathbf{z}} - h(\mathbf{z}+\delta)\|_2^H$ follows a reciprocal distribution as

$$p\left(d^H; a, b\right) = \frac{1}{d^H\left(\log(b) - \log(a)\right)}, \tag{15}$$

where $a \le d^H \le b$, and $a > 0$. The exact values of a and b are irrelevant, as they only appear in an additive constant when we take the logarithm of $p\left(d^H; a, b\right)$.

Since there is no obvious reason for assuming Gaussian $\hat{\mathbf{z}}$, we can instead assume $\hat{\mathbf{z}}$ to follow the distribution defined in Eq. (15), and multiply H by a tunable scalar, γ', similar to σ. Furthermore, we can replace δ in Eq. (7) with $\delta' \sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right)$, as it is defined for VPGA with a subtle difference that here $\sigma^2_\phi(\mathbf{x})$ is constrained to be greater than ϵ^2. As a result, LPGA and VPGA are unified into a single approach, which has a combined loss function as

$$L_{lvpga} = L_{pga} + \gamma' L_{vr} + \gamma L^{\phi}_{nll} + \eta L^{\phi}_{vkl}. \tag{16}$$

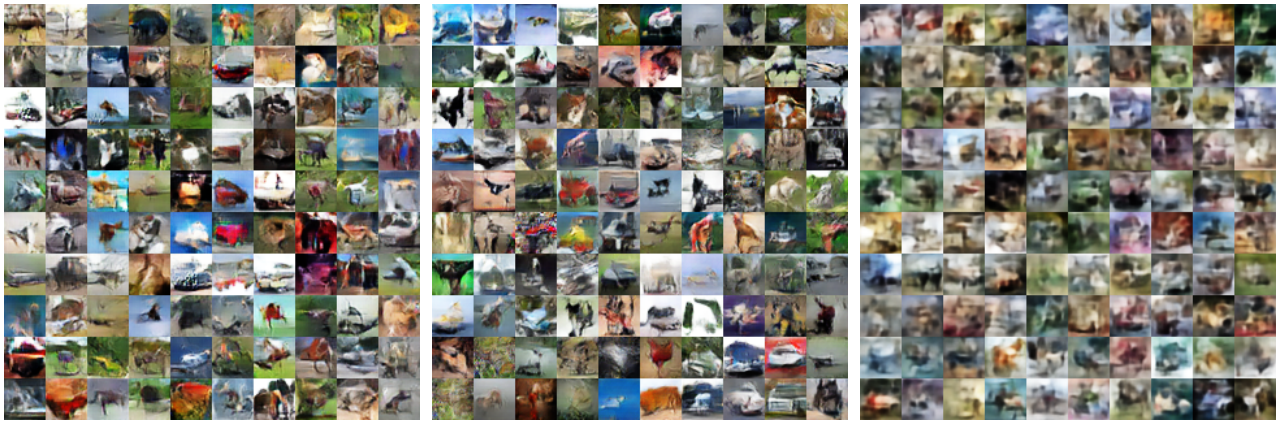
When $\gamma' = \gamma$ and $\eta = 0$, Eq. (16) is equivalent to Eq. (9), considering that $\sigma^2_\phi(\mathbf{x})$ will be optimized to approach ϵ^2. Similarly, when $\gamma = 0$, Eq. (16) is equivalent to Eq. (13). Interestingly, it also becomes possible to have a mix of LPGA and VPGA by setting all three hyperparameters to positive values. This approach mainly serves to demonstrate the connection between LPGA and VPGA, and is less practical due to the extra hyperparameters. We refer to this approach as LVPGA.

4. Experiments

In this section, we evaluate the performance of LPGA and VPGA on three image datasets, MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009), and CelebA (Liu et al., 2015). For CelebA, we employ the discriminator and generator architecture of DCGAN (Radford et al., 2016) for the encoder and decoder of PGA. We half the number of filters (i.e., 64 filters for the first convolutional layer) for faster experiments, while more filters are observed to improve performance. Due to smaller input sizes, we reduce the number of convolutional layers accordingly for MNIST and CIFAR-10, and add a fully-connected layer of 1024 units for MNIST, as done in Chen et al. (2016). SGD with a momentum of 0.9 is used to train all models. Other hyperparameters are tuned heuristically, and could be improved by a more extensive grid search. For fair comparison, σ is tuned for both VAE and VPGA. All experiments are performed on a single GPU.

As shown in Fig. 2, the visual quality of the PGA-generated samples is significantly improved over that of VAEs. In

(a) MNIST by LPGA (b) MNIST by VPGA (c) MNIST by VAE



(d) CIFAR-10 by LPGA (e) CIFAR-10 by VPGA (f) CIFAR-10 by VAE

(g) CelebA by LPGA (h) CelebA by VPGA (i) CelebA by VAE

Figure 2. Random samples generated by LPGA, VPGA, and VAE. Note how LPGA and VPGA images are less blurry than those from the VAE.

particular, PGAs generate much sharper samples on CIFAR-10 and CelebA compared to vanilla VAEs. The results of LVPGA much resemble that of either LPGA or VPGA, depending on the hyperparameter settings. In addition, we use the Fréchet Inception Distance (FID) (Heusel et al., 2017) to evaluate the proposed methods, as well as VAE. For each model and each dataset, we take 5,000 generated samples to compute the FID score. The results (with standard errors of

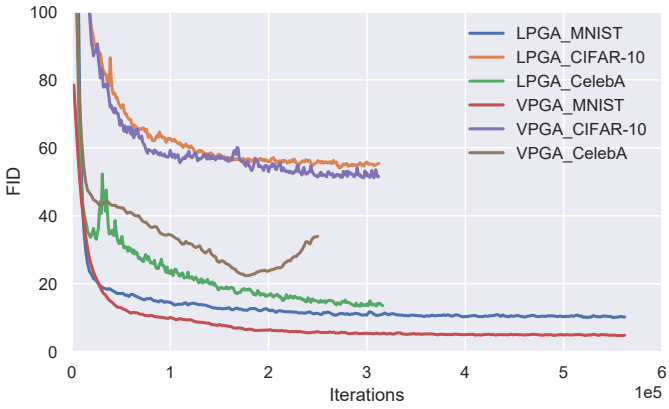
Table 1. FID scores of autoencoder-based generative models. The first block shows the results from Ghosh et al. (2019), where CV-VAE stands for constant-variance VAE, and RAE stands for regularized autoencoder. The second block shows our results of LPGA, VPGA, LVPGA, and VAE.

Model	MNIST	CIFAR-10	CelebA
VAE	19.21	106.37	48.12
CV-VAE	33.79	94.75	48.87
WAE	20.42	117.44	53.67
RAE-L2	22.22	80.80	51.13
RAE-SN	19.67	84.25	44.74
VAE	13.83 ± 0.06	115.74 ± 0.63	43.60 ± 0.33
LPGA	10.34 ± 0.15	55.87 ± 0.25	14.53 ± 0.52
VPGA	$\mathbf{4.97} \pm 0.07$	$\mathbf{51.51} \pm 1.16$	24.73 ± 1.25
LVPGA	6.32 ± 0.16	52.94 ± 0.89	$\mathbf{13.80} \pm 0.20$

3 or more runs) are summarized in Table. 1. Compared to other autoencoder-based non-adversarial approaches (Tolstikhin et al., 2018; Kolouri et al., 2019; Ghosh et al., 2019), where similar but larger architectures are used, we obtain substantially better FID scores on all three datasets. Note that the results from Ghosh et al. (2019) shown in Table. 1 are obtained using slightly different architectures and evaluation protocols. Nevertheless, their results of VAE align well with ours, suggesting a good comparability of the results. Interestingly, as a unified approach, LVPGA can indeed combine the best performances of LPGA and VPGA on different datasets. For CelebA, we show further results on 140x140 crops and latent space interpolations in the supplementary material. While PGA has largely bridged the performance gap between generative autoencoders and GANs, there is still a noticeable gap between them especially on CIFAR-10. For instance, the FIDs of WGAN-GP and SN-GAN on CIFAR-10 using a similar architecture are respectively 40.2 and 25.5 (Miyato et al., 2018), as compared to 51.5 of VPGA.

Empirically, different PGA variants share the same optimal values of α and β (Eq. (4)) when trained on the same dataset. For LPGA, γ (Eq. (9)) tends to vary in a small range for different datasets (e.g., $1.5\text{e}{-2}$ for MNIST and CIFAR-10, and $1\text{e}{-2}$ for CelebA). For VPGA, η (Eq. (13)) can vary widely (e.g., $2\text{e}{-2}$ for MNIST, $3\text{e}{-2}$ for CIFAR-10, and $2\text{e}{-3}$ for CelebA), and thus is slightly more difficult to tune. The training process of PGAs is stable in general, given the non-adversarial losses. As shown in Fig. 3a, the total losses change little after the initial rapid drops. This is due to the fact that the encoder and decoder are optimized towards different objectives, as can be observed from Eqs. (4), (9), and (12). In contrast, the corresponding FIDs, shown in Fig. 3b, tend to decrease monotonically during training. However, when trained on CelebA, there is a significant performance gap between LPGA and VPGA, and the FID of the latter

(a) Total loss



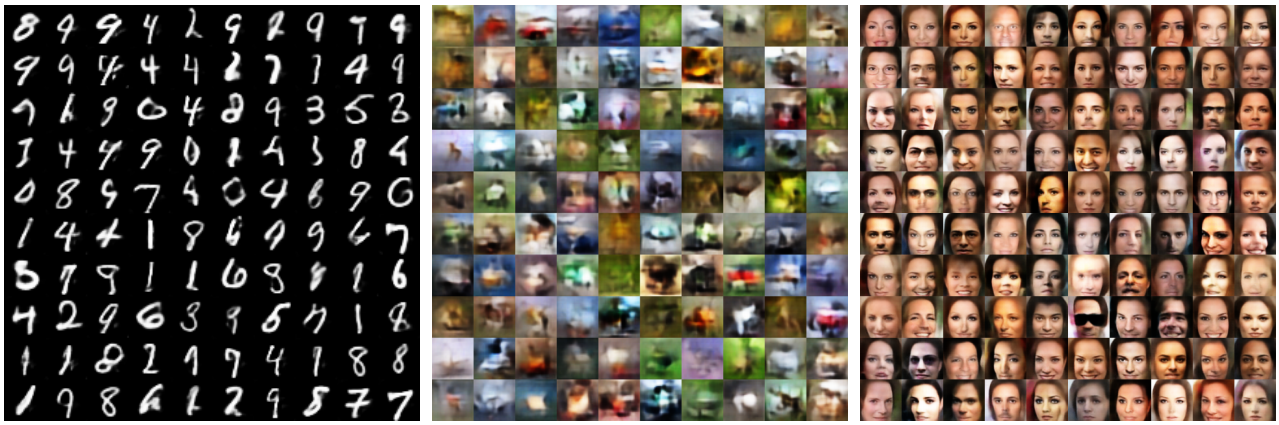
(b) FID

Figure 3. Training curves of LPGA and VPGA.

starts to increase after a certain point of training. We suspect this phenomenon is related to the limited expressiveness of the variational posterior, which is not an issue for LPGA.

It is worth noting that stability issues can occur when batch normalization (Ioffe & Szegedy, 2015) is introduced, since both the encoder and decoder are fed with multiple batches drawn from different distributions. At convergence, different input distributions to the decoder (e.g., $\mathcal{H}$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$) are expected to result in similar distributions of the internal representations, which, intriguingly, can be imposed to some degree by batch normalization. Therefore, it is observed that when batch normalization does not cause stability issues, it can substantially accelerate convergence and lead to slightly better generative performance. Furthermore, we observe that LPGA tends to be more stable than VPGA in the presence of batch normalization.

Finally, we conduct an ablation study. While the loss functions of LPGA and VPGA both consist of multiple components, they are all theoretically motivated and indispensable. Specifically, the data reconstruction loss minimizes



(a) MNIST, FID = 16.99 (b) CIFAR-10, FID = 114.19 (c) CelebA, FID = 48.02

Figure 4. Random samples generated by LPGA without the latent reconstruction losses ($\alpha = \beta = 0$). Compared to the samples in Fig. 2, we observe a degradation.

the discrepancy between the input data and its reconstruction. Since the reconstructed data distribution serves as the surrogate target distribution, removing the data reconstruction loss will result in a random target. Moreover, removing the maximum likelihood loss of LPGA or the VAE loss of VPGA will leave the perceptual generative model untrained. In both cases, no valid generative model can be obtained. Nevertheless, it is interesting to see how the latent reconstruction loss contributes to the generative performance. Therefore, we retrain the LPGAs without the latent reconstruction loss and report the results in Fig. 4. Compared to Fig. 2a, 2d, 2g, and the results in Table 1, the performance significantly degrades both visually and quantitatively, confirming the importance of the latent reconstruction loss.

5. Conclusion

We proposed a framework, PGA, for training autoencoder-based generative models, with non-adversarial losses and unrestricted neural network architectures. By matching target distributions in the latent space, PGAs trained with maximum likelihood generalize the idea of reversible generative models to unrestricted neural network architectures and arbitrary latent dimensionalities. In addition, it improves the performance of VAE when combined together. Under the PGA framework, we further show that maximum likelihood and VAE can be unified into a single approach.

In principle, the PGA framework can be combined with any method that can train the perceptual generative model. While we have only considered non-adversarial approaches, an interesting future work would be to combine it with an adversarial discriminator trained on latent representations. Moreover, the compatibility issue with batch normalization deserves further investigation.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Bengio, Y., Yao, L., Alain, G., and Vincent, P. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pp. 899–907, 2013.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6571–6583, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Dai, B. and Wipf, D. Diagnosing and enhancing vae models. In *International Conference on Learning Representations*, 2019.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. In *International Conference on Learning Representations Workshop*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

Ghosh, P., Sajjadi, M. S. M., Vergari, A., Black, M., and Schölkopf, B. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Grathwohl, W., Chen, R. T., Betterncourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Hou, X., Shen, L., Sun, K., and Qiu, G. Deep feature consistent variational autoencoder. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1133–1141. IEEE, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

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.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10236–10245, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

LeCun, Y., Cortes, C., and Burges, C. J. C. The mnist handwritten digit database, 1998.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. Adversarial autoencoders. In *International Conference on Learning Representations Workshop*, 2016.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Rezende, D. J. and Viola, F. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Rubenstein, P. K., Schoelkopf, B., and Tolstikhin, I. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008.