Efficient Integration of Perceptual Variational Autoencoder into Dynamic Latent Scale GAN

Jeongik Cho

Overview

Dynamic Latent Scale GAN (DLSGAN) is an architecture-agnostic GAN inversion framework that rescales each latent dimension according to the variance of the encoder output. This dynamic scaling effectively regulates the entropy of the latent distribution, facilitating stable training and invertibility. DLSGAN is trained with the following objectives:

$$v \approx E_l(G(Z \circ s))^2$$

$$s = \sqrt{d_z} \cdot \frac{\sqrt{v}}{\|\sqrt{v}\|_2}$$

$$L_{enc} = \text{avg}\left(((Z - E_l(G(Z \circ s))) \circ s)^2\right)$$

$$L_d = L_{adv}^d + \lambda_{enc} L_{enc}, \quad L_g = L_{adv}^g + \lambda_{enc} L_{enc}$$

Here, E_l and G denote the latent encoder and generator, respectively, and Z is a sample from the prior distribution $\mathcal{N}(0, I)$. The variance vector v is estimated from the encoder outputs, and the scaling vector s adjusts the latent space to match the encoder's representational capacity. The operator $\circ$ denotes element-wise multiplication.

Proposed Method: PVDGAN

We propose PVDGAN, a perceptual variational DLSGAN framework that integrates a perceptual VAE loss into DLSGAN without requiring architectural changes. The key idea is to add scaled Gaussian noise to the latent encoder output so that the resulting vector remains distributed as $\mathcal{N}(0, I)$. This enables the same latent vector to be used for both adversarial training and perceptual reconstruction.

The perturbed latent vector is defined as:

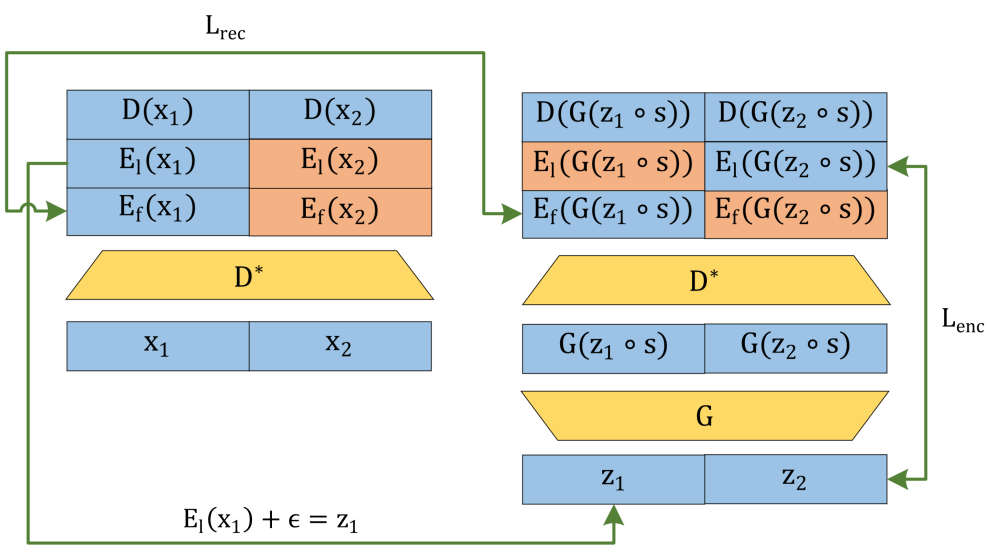
$$Z_x = E_l(x) + \mathcal{N}(0, I) \cdot \sqrt{1 - v}$$

Here, $E_l(x)$ is the latent encoder output, and v is the per-dimension variance vector. This formulation ensures that $Z_x \sim \mathcal{N}(0, I)$, maintaining compatibility with the generator's latent prior.

To incorporate perceptual supervision, we reuse an intermediate feature extractor $E_f(\cdot)$ from the discriminator and define the following reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x \sim p_{data}}\left[\|E_f(G(Z_x \circ s)) - E_f(x)\|_2^2\right]$$

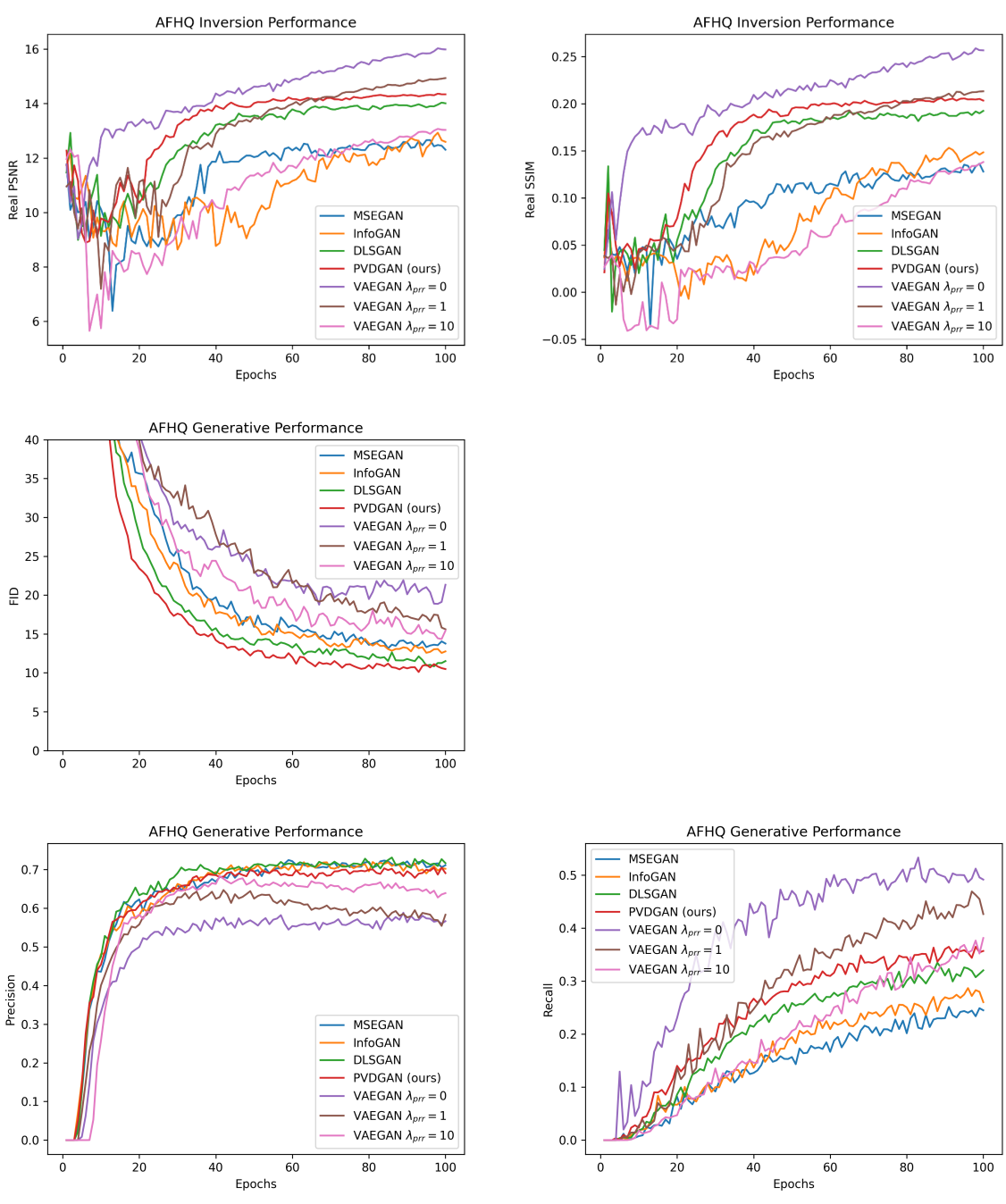
This loss aligns the generated output with perceptual features of the input and is used to train the encoder alongside the adversarial objectives.



The diagram above illustrates the loss flow of PVDGAN. Since the discriminator, latent encoder, and feature encoder share internal layers, the computational and memory cost is equivalent to that of a standard GAN.

Experimental Results

We evaluate PVDGAN against MSEGAN, InfoGAN, DLSGAN, and VAEGAN on the FFHQ and AFHQ datasets.



The graphs above compare inversion and generation performance, respectively. VAEGAN with a prior loss weight $\lambda_{prr} = 0$ achieves the best inversion performance, but behaves more like a perceptual autoencoder, resulting in reduced generative performance. As λ_{prr} increases, VAEGAN improves in generation quality but suffers a decline in inversion accuracy.

Excluding VAEGAN, the proposed PVDGAN consistently outperforms MSEGAN, InfoGAN, and DLSGAN in both inversion fidelity and generative diversity.

The figure above shows reconstruction results on the AFHQ dataset. Compared to other methods, PVDGAN preserves more fine-grained details in the reconstructed images, except for VAEGAN with a low λ_{prr}.

Conclusion

We demonstrate that integrating a perceptual VAE loss into DLSGAN improves both inversion accuracy and generative quality without increasing model complexity. By perturbing the encoder output with scaled noise and leveraging shared features from the discriminator, PVDGAN provides a unified, efficient training framework that enhances both reconstruction and sample diversity.

Full implementation is available at: `https://github.com/jeongik-jo/PVDGAN`