Isometric Representation Learning for Disentangled Latent Space of Diffusion Models

Jaehoon Hahm[1*], Junho Lee[1*], Sunghyun Kim[1], Joonseok Lee[1,2]
[1]Seoul National University, [2]Google Research

ICML International Conference On Machine Learning

Motivation

Entangled latent space of Diffusion Models

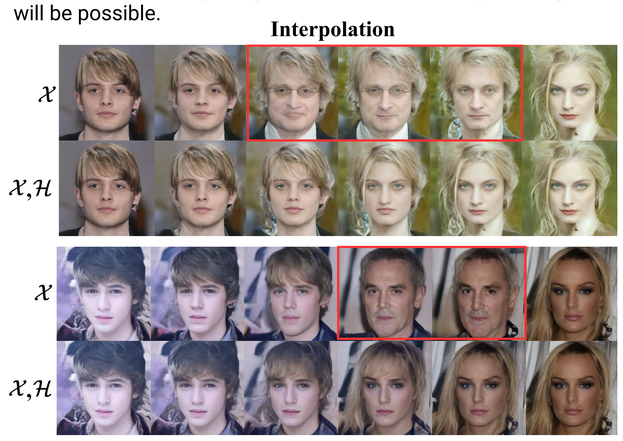
- Diffusion models lack understanding of its latent space compared to GANs and VAEs.
- Existence of abrupt changes during latent traversal of diffusion models.
- It implies that entanglement exists in the latent space of diffusion models.
- This often leads to suboptimal result in downstream tasks such as image interpolation, inversion, or editing.

StyleGAN2[1] StyleGAN3[2]

DDPM (GENIE[3])

Semantic Latent Space of Diffusion Models

- Kwon et al.[4] proposed the bottleneck feature space $\mathcal{H}$ as the semantic space of diffusion models
- Editing visual attributes of a given image is possible using $\mathcal{H}$
- If we can move along the geodesic of $\mathcal{H}$, disentangled image editing will be possible.



Our Contributions

- We achieve effectively disentangled latent space of diffusion models without compromising the quality of generated images.
- We propose Isometric Diffusion, a diffusion model that achieves a geometrically sound latent space by regularizing the mapping from the latent space to $\mathcal{H}$-space to be isometric.
- We verify the effectiveness of our proposed method through quantitative and qualitative evaluations on various applications.

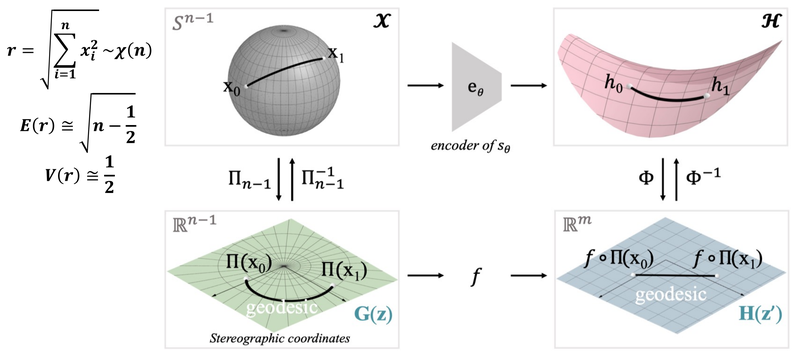
Method

Definition and Properties of Scaled Isometric Mapping

- The mapping function f is a scaled isometry[5] if and only if:

$$\mathbf{R}(z) = \mathbf{J}_f(z)^\top \mathbf{H}(f(z))\mathbf{J}_f(z)\mathbf{G}^{-1}(z) = c\boldsymbol{I},$$

- Properties of Isometry
  1) *Geodesic-preserving property:* equal sensitivity of each latent basis
  2) *Angle-preserving property:* preserving orthogonality, disentangled latent basis
- Due to the geodesic-preserving property, if the encoder e_θ becomes an scaled isometry, moving along the geodesics in $\mathcal{X}$ corresponds to moving along geodesics in $\mathcal{H}$

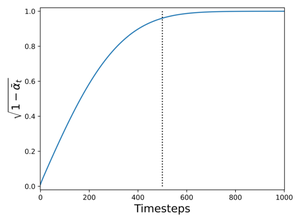


$$\Pi_{n-1}(\boldsymbol{x}) = \frac{1}{r - \boldsymbol{x}_n}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}), \quad \Pi_{n-1}^{-1}(\boldsymbol{z}) = \frac{r}{|\boldsymbol{z}|^2 + 1}(2\boldsymbol{z}_1, 2\boldsymbol{z}_2, \cdots, 2\boldsymbol{z}_{n-1}|\boldsymbol{z}|^2 - 1).$$

$$\mathbf{G}_{\text{stereographic}}(\boldsymbol{z}) = \frac{4r^4}{(|\boldsymbol{z}|^2 + r^2)^2}\boldsymbol{I}_{n-1}, \quad \leftarrow \text{Accounts for the geometry (distance, angle, volume, ...) of the given space}$$

Spherical Approximation of the Latent Space

- The radii of Gaussian noise vectors follow χ-distribution, so we can approximate the noise vectors reside on the hypersphere manifold $S^{n-1}(r)$
- We define the Riemannian metric on $S^{n-1}(r)$ by choosing stereographic coordinates as the local coordinates.
- For sufficiently large $t, \sqrt{1-\bar{\alpha}_t} \approx 1$, and the noise space comprised of $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0$ can be approximated as a sphere.



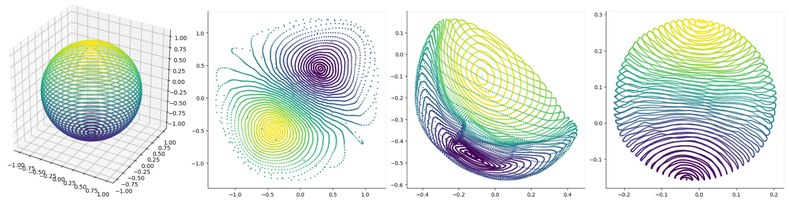
Isometric Regularizer for Diffusion Models

- Isometric Regularizer

$$\mathcal{L}_{\text{iso}}(f,t) = \frac{\mathbb{E}_{\boldsymbol{x}_t \sim P(\boldsymbol{x}_t)}[\text{Tr}(\mathbf{R}^2(\boldsymbol{z}_t))]}{\mathbb{E}_{\boldsymbol{x}_t \sim P(\boldsymbol{x}_t)}[\text{Tr}(\mathbf{R}(\boldsymbol{z}_t))]^2} = \frac{\mathbb{E}_{\boldsymbol{x}_t \sim P(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,\boldsymbol{I})}[\boldsymbol{v}^\top \mathbf{R}(\boldsymbol{z}_t)^\top \mathbf{R}(\boldsymbol{z}_t)\boldsymbol{v}]}{\mathbb{E}_{\boldsymbol{x}_t \sim P(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,\boldsymbol{I})}[\boldsymbol{v}^\top \mathbf{R}(\boldsymbol{z}_t)\boldsymbol{v}]^2}$$

$$\mathcal{L} = \mathcal{L}_{\text{dsm}} + \lambda_{\text{iso}}(\gamma, t)\mathcal{L}_{\text{iso}}(e_\theta, t) \qquad \lambda_{\text{iso}}(\gamma, t) = \lambda_{\text{iso}}\mathbf{1}_{t' > \gamma T}(t' = t)$$

- By using the isometric regularizer, we can encourage the encoder to become closer to an isometric mapping and obtain a geometrically disentangled latent space.



(a) S^2 manifold (b) Reconstruction only (c) $\mathcal{L}_{\text{iso}}$ with $\mathbf{G} = \boldsymbol{I}$ (d) $\mathcal{L}_{\text{iso}}$ with full Eq. (6)

Results

Quantitative Analysis

Dataset	Model	FID-10k↓ Base	FID-10k↓ Ours	PPL-50k↓ Base	PPL-50k↓ Ours	mRTL↓ Base	mRTL↓ Ours	MCN ↓ Base	MCN ↓ Ours	VoR ↓ Base	VoR ↓ Ours	LS ↓ Base	LS ↓ Ours
CIFAR-10	DDPM	**10.27**	12.50	105	**76**	2.03	**1.92**	155	**107**	**0.50**	0.57	-	-
LSUN-Church	DDPM	**10.56**	13.01	2028	**1587**	3.71	**3.21**	375	**217**	1.92	**1.37**	-	-
LSUN-Bedrooms	DDPM	**9.49**	11.95	4515	**3809**	3.38	**3.21**	320	**186**	1.69	**1.12**	-	-
CelebA-HQ	DDPM	**15.89**	16.18	648	**455**	2.67	**2.50**	497	**180**	1.42	**0.85**	1.91	**1.51**
CelebA-HQ	LDM	**10.79**	11.46	439	**397**	2.89	**2.73**	322	**198**	1.04	**0.54**	2.38	**2.15**

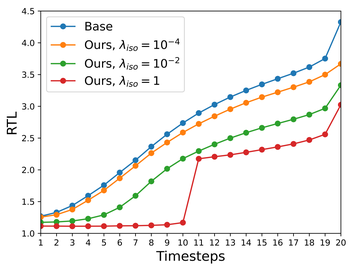
- **FID** – Generation quality
- **mRTL, MCN, VoR** – Proximity to isometry
- **PPL, LS** – Smoothness of latent space, entanglement metric

- Image inversion and reconstruction

Regularizer	PPL-50k	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
-	401	0.00862	0.597	20.6	0.517
$\mathcal{L}_{\text{pl}}$ (Path length reg.)	368	0.00667	0.614	21.7	0.521
$\mathcal{L}_{\text{iso}}$ (Ours)	**340**	**0.00599**	**0.674**	**22.2**	**0.436**

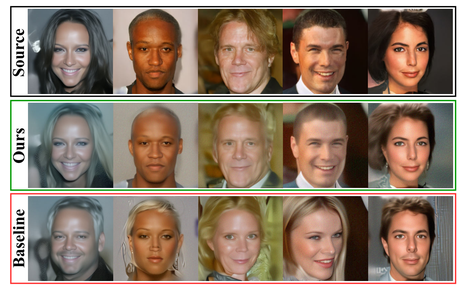
- Ablation studies

Regularizer	G	FID-10k↓	PPL-50k↓
-	-	15.89	648
$\mathcal{L}_{\text{pl}}$ (Path length reg.)	$\mathbf{I}$	20.04	552
$\mathcal{L}_{\text{iso}}$	$\mathbf{I}$	16.60	619
$\mathcal{L}_{\text{iso}}$ (Ours)	$\mathbf{G}_\mathbf{s}$	16.18	**455**

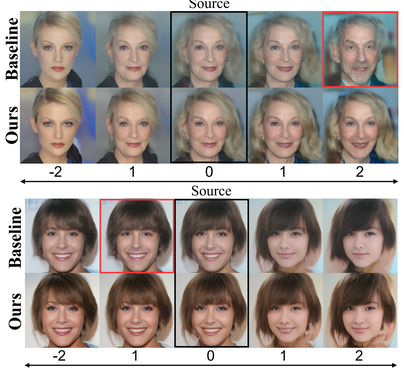


Qualitative Analysis

- Image inversion and reconstruction



- Linearity



- Interpolation (CelebA-HQ, LSUN-Church, LSUN-Bedrooms)

[1] Karras et al., "Analyzing and improving the image quality of stylegan.", 2020 [2] Karras et al., "Alias-free generative adversarial networks", 2021 [3] Dockhorn et al., "GENIE: Higher-order Denoising Diffusion Solvers", 2022 [4] Kwon et al., "Diffusion Models already have a Semantic Latent Space", 2023 [5] Lee et al., "Regularized Autoencoders for Isometric Representation Learning", 2022