High-Resolution Image Synthesis with Latent Diffusion Models

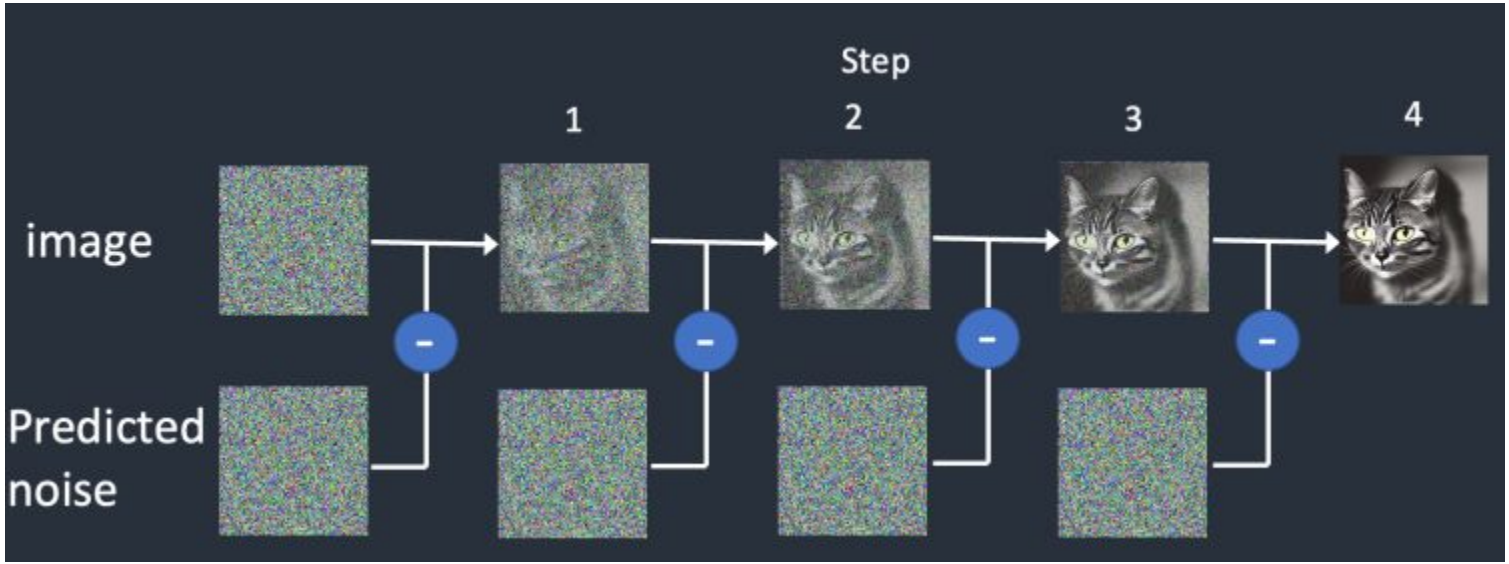
윤세환

목차

- 사전지식
  - diffusion model의 학습 과정
- introduction
- method
- experiments
- limitations
- 참고 링크

Diffusion Model의 학습 과정

- 랜덤하게 t 시점을 선택
- t 시점에 맞는 noise를 생성, 원본 데이터 x_0에 nosie를 더해 손상
- U-Net 네트워크로 하여금, 직전에 생성한 noise를 예측하도록 학습



Introduction

- 기존 diffusion 방식의 단점
  - 모델이 예측하는 결과값의 차원이 타 모델에 비해서 크다 (입력 이미지의 차원과 동일)
  - 훈련하는 과정에서, 매 t 스텝마다 noise를 예측해야 하기에, 훈련에 굉장히 많은 시간이 소요됨
  - 또한, 예측하는 경우에도 굉장히 많은 시간이 소요됨
    - A100 GPU 기준, 5만장을 예측하는데 5일이 소모됨.

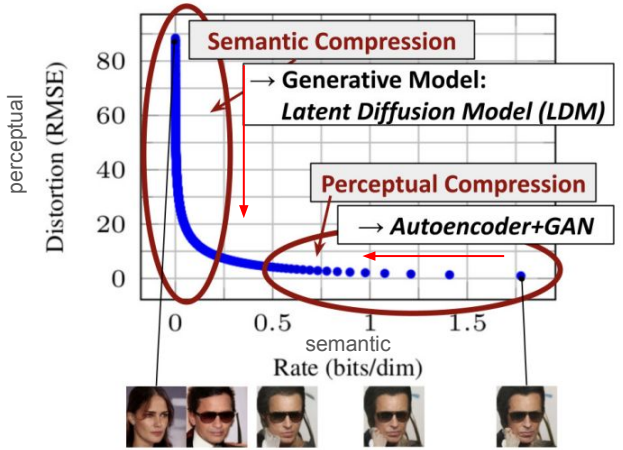
- 본 논문에서는 기존 Diffusion Model의 높은 연산량을 줄여 훈련 및 샘플링 과정에서의 시간과 소모되는 자원을 줄이고자 함

Introduction

- Semantic Compression
  - 의미(bits)를 유지하는 선에서, 다른 high-frequency detail들을 학습(RMSE)
  - 기존 Diffusion Model이 강점을 가지던 분야

- Perceptual Compression
  - 실제 모델이 데이터의 의미를 학습하는 단계

- 본 논문에서 제시하는 Latent Diffusion Model은 perceptual은 동등하되, 계산적으로 더 효율적인 space를 찾는 것을 목표로 함



bits/dim : NLL에서 밑이 2인 로그를 사용한 값을 픽셀의 총 개수로 나눈 것

Method

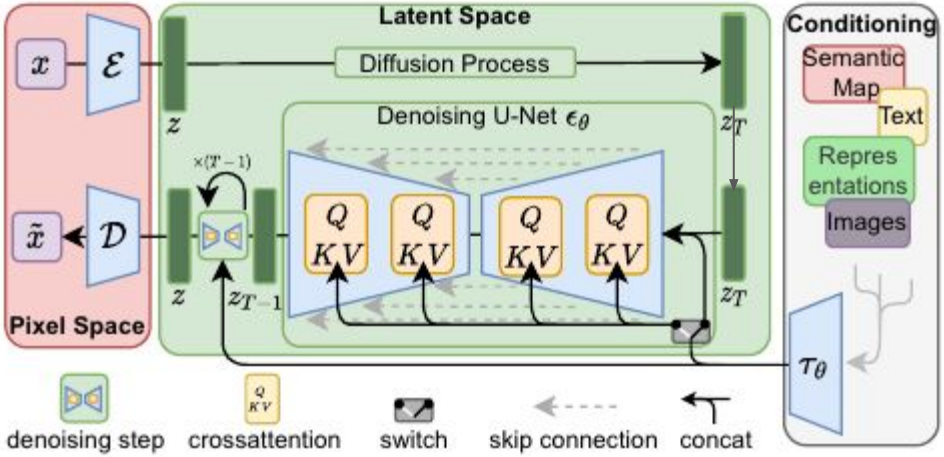


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \text{ with}$$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$

autoEncoder (왼쪽 빨강색 영역)은 pretrain된 모델

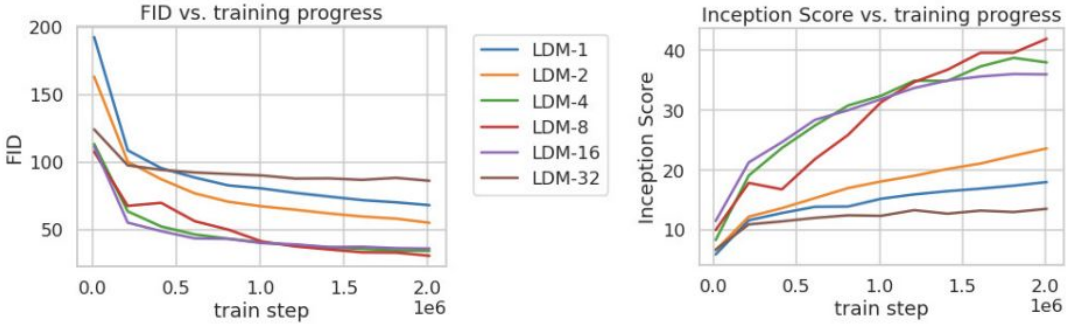
Method

- 일반적인 Diffusion Model (DM)의 경우, 원본 데이터 x에 대한 노이즈를 예측
- Latent Diffusion Model(LDM)의 경우, latent vector에 대한 노이즈를 예측

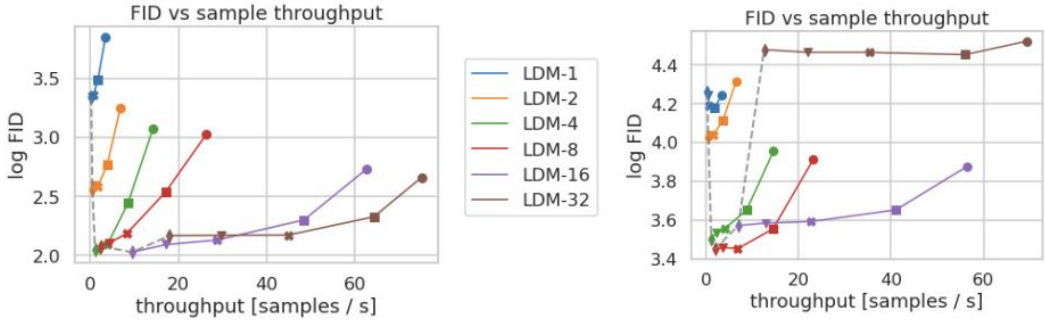
$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right],$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right].$$

Experiments



CelebA-HQ DataSet
(1024*1024)



ImageNet DataSet
(평균 469*387)

Experiments

CelebA-HQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-
VQGAN+T. [23] (k=400)	10.2	-	-
PGGAN [39]	8.0	-	-
LSGM [93]	7.22	-	-
UDM [43]	7.16	-	-
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49

FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	9.57	-	-
U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
UDM [43]	5.54	-	-
StyleGAN [41]	4.16	0.71	0.46
ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	3.86	-	-
ProjectedGAN [76]	**1.59**	0.61	0.44
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**

LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	5.51	-	-
DDPM [30]	4.9	-	-
UDM [43]	4.57	-	-
StyleGAN [41]	2.35	0.59	0.48
ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.52**	0.61	0.34
LDM-4 (ours, 200-s)	2.95	**0.66**	0.48

Experiments

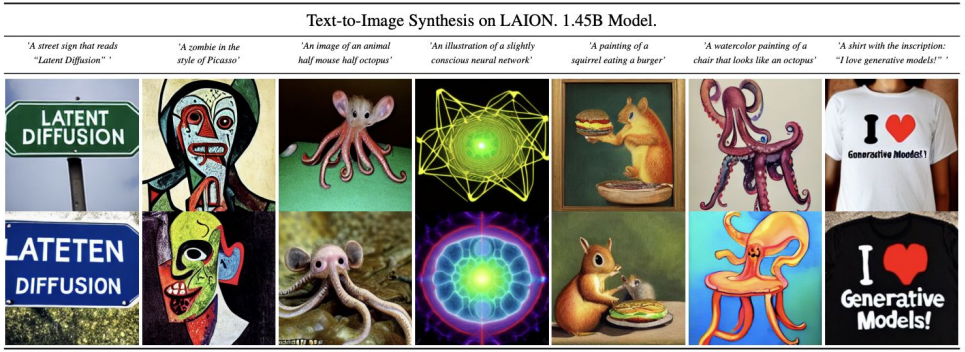


Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

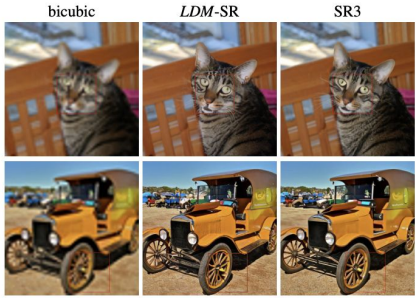


Figure 10. ImageNet 64→256 super-resolution on ImageNet-Val. *LDM-SR* has advantages at rendering realistic textures but SR3 can synthesize more coherent fine structures. See appendix for additional samples and cropouts. SR3 results from [72].

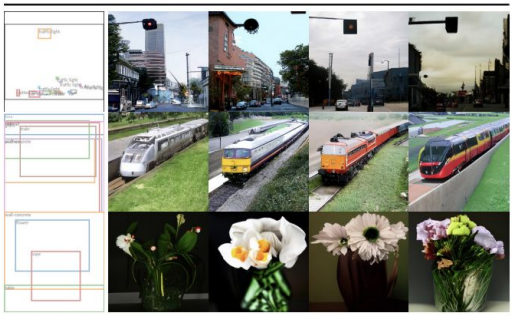


Figure 8. Layout-to-image synthesis with an *LDM* on COCO [4], see Sec. 4.3.1. Quantitative evaluation in the supplement D.3.

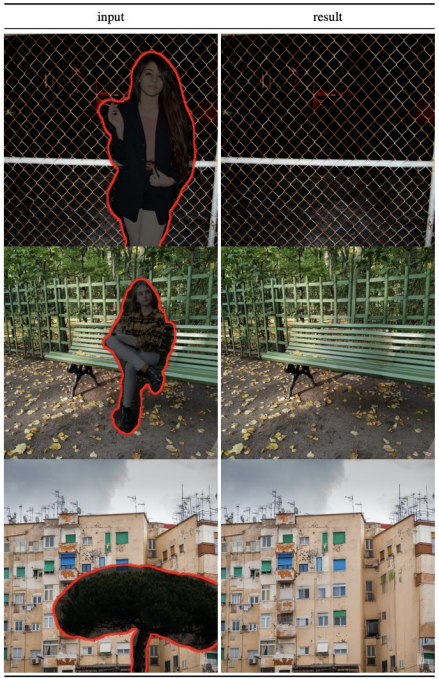


Figure 11. Qualitative results on object removal with our *big, w/ ft* inpainting model. For more results, see Fig. 22.

Limitations

- 기존 픽셀 기반의 방식인 Diffusion Model에 비해 계산 요구량을 크게 감소시킬 수 있지만, 여전히 DM 모델의 방식상 샘플링에 있어서는 GAN 보다 느림
- 픽셀 단위가 아닌, latent space상에서 작업을 수행하기 때문에, 더 높은 정밀도가 필요한 고화질의 경우 기존 Diffusion Model에 비해 품질 손실이 발생할 수 있음

참고 링크

- 원본 논문 : https://arxiv.org/abs/2112.10752
- Stable diffusion에 대한 기본적인 이론 : https://www.internetmap.kr/entry/Basic-Theory-of-Stable-Diffusion#training