High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer

Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany Runway ML

2022 / CVPR

발제자 : 정현우 (junghw333@gmail.com)

랩 : HUMANE Lab

2023-11-21

Diffusion Model

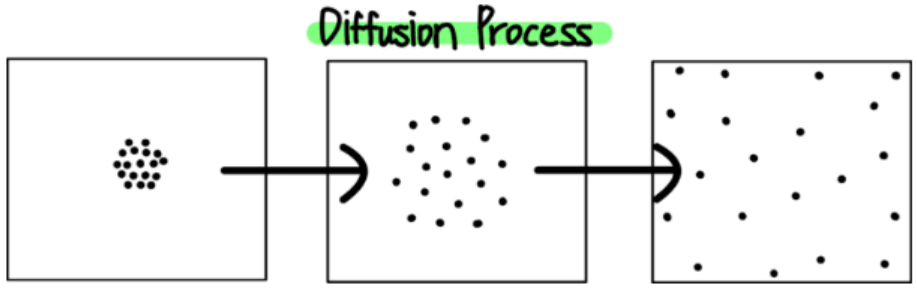


그림4. Forward Diffusion Process

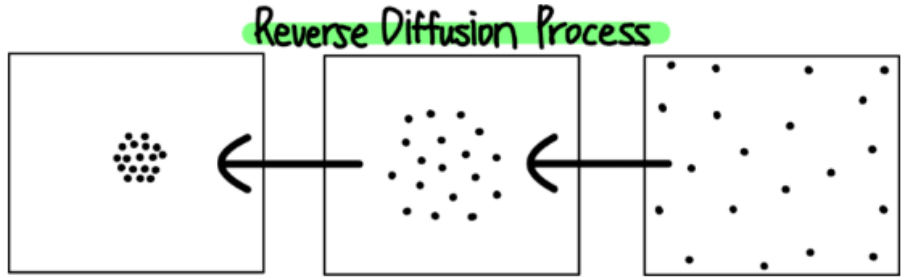


그림6. Reverse Diffusion Process

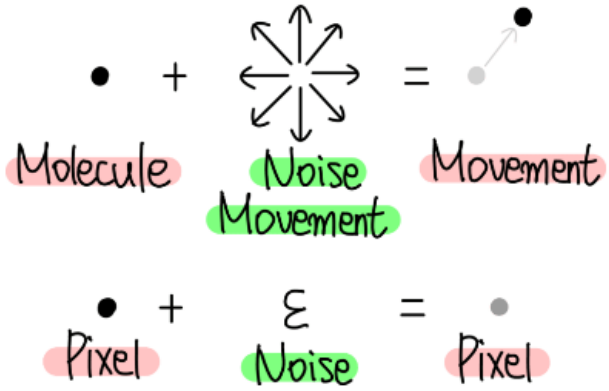


그림7. 이미지 픽셀에 Diffusion 적용하기

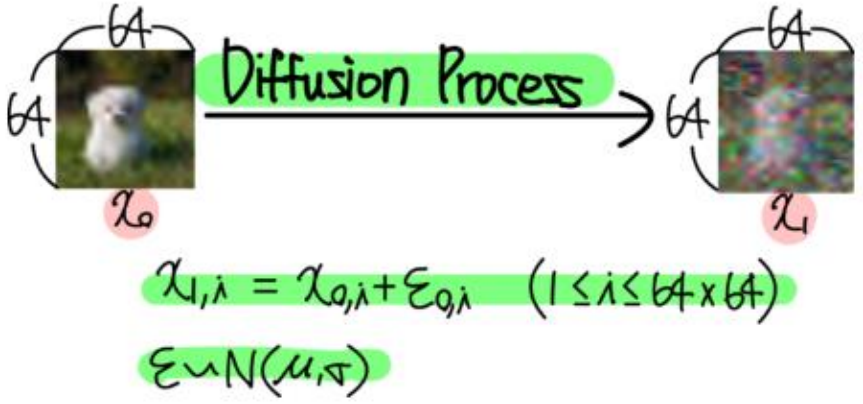
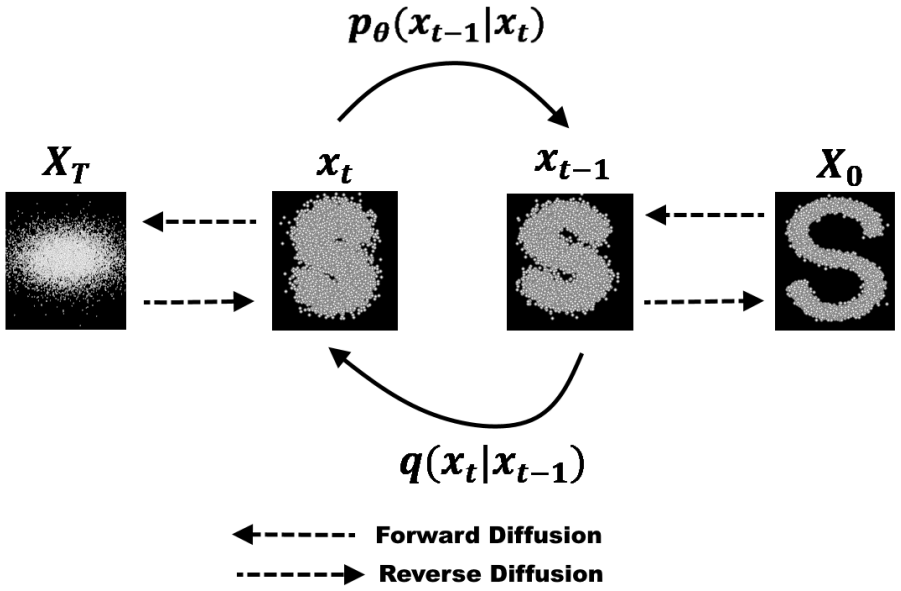


그림8. Image에 적용한 Diffusion

Diffusion Model



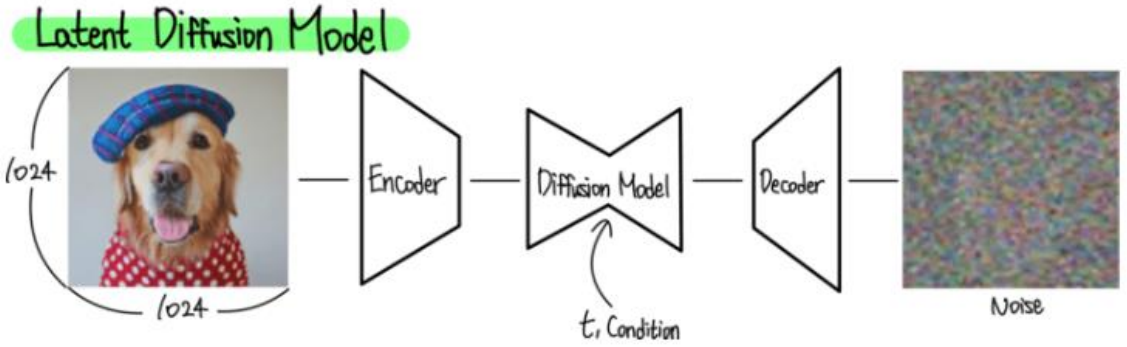
- Diffusion Model은 이미지에 노이즈를 입혀 노이즈가 낀 이미지로 만든다.
- 노이즈로부터 이미지를 복원하며 학습이 이루어진다.
- 이 때 입력으로 이미지를 넣어주게 된다.
- 이 때 복원하는 분포를 모르기 때문에 p(x)를 알아가는 것이 학습이다.

Latent Diffusion Model

- 기존 DM 모델의 문제점

- 픽셀값이 입력으로 들어가기 때문에 사람 눈이 보이지 않는 부분에 투자를 한다.
- 또한 그렇기에 계산 효율이 안 좋다. (우리가 인지할 수 없는 부분에 계산을 많이 투자한다.)

- => 때문에 이미지 그대로 사용하는 것이 아니라 잠재 벡터를 사용해서 이를 해결한다.



Model Architecture

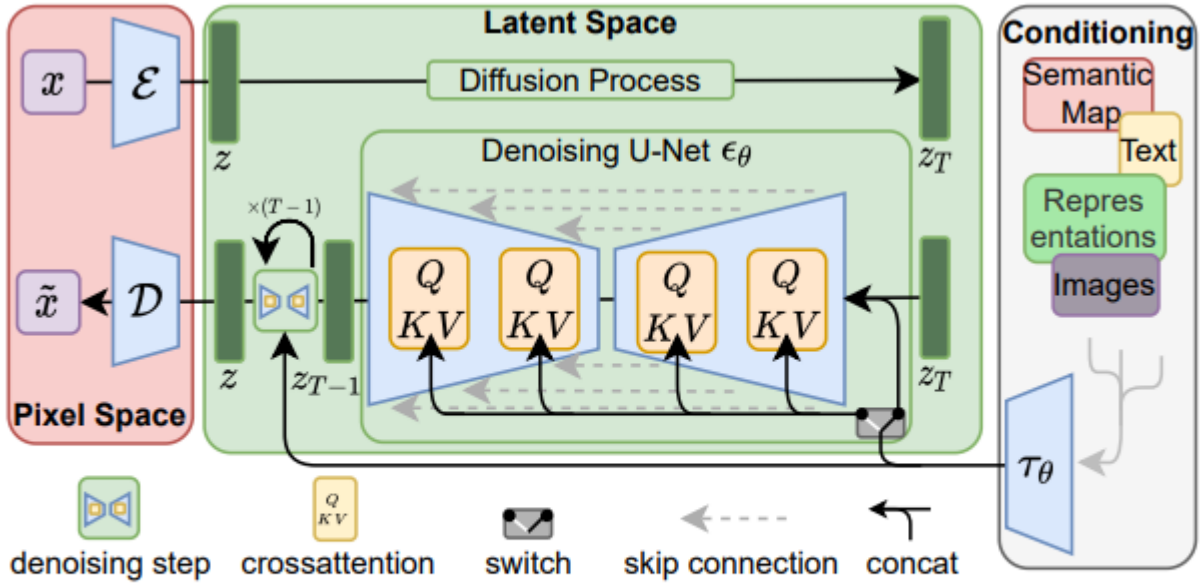


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

- 이미지 x가 들어가고 인코더를 사용해서 z를 만들어낸다.
- 이 때 z가 latent 벡터가 된다.
- 이 벡터에 노이즈를 추가하는 과정을 해서 zT를 만들어낸다.

Model Architecture

- # condition
- 이 구조는 다양한 형태의 조건을 입력 받을 수 있다.
- 컨디션은 각자에 맞는 적절한 인코더가 필요하다. 텍스트를 쓸거면 텍스트 인코더가 필요함.

Model Architecture

- # condition
- 조건으로 바운딩 박스 + 클래스 정보를 입력하면 자동으로 만들어준다.

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

Model Architecture

- # condition
- 조건은 빈 공간이 있는 이미지이다.

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

Model Architecture

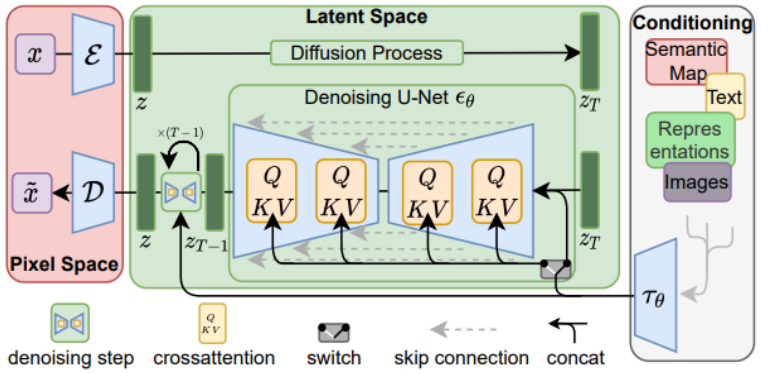


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

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

그림7. Query, Key, Value 수식

- 쿼리, 키, 벨류를 사용하는데, zt는 이미지 정보, y는 조건 정보이다.
- 이미지와 조건의 상관 관계를 고려하여 조건 정보에 가중치를 반영하는 것이다.
- 도메인별 인코더가 τθ이다.

Model Architecture

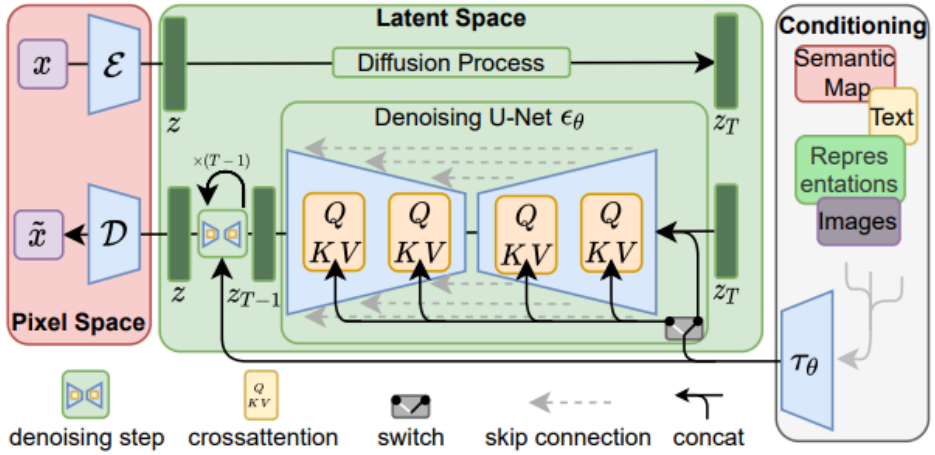


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

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

그림6. Cross Attention 수식

- 이미지와 조건 사이의 정보의 상관 관계를 고려하기 위해서 가장 많이 사용되는 방법은 크로스 어텐션이다.
- 이는 동일한 메커니즘을 사용해서 두 정보의 상관 관계를 고려하는 방법이다. self는 하나의 정보에서 중요한 것 찾기.

Model Architecture

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

그림4. Diffusion Model Loss Function

- Diffusion model이 원복 이미지와 복원 이미지와 시점 t를 입력으로 넣은 Loss 식을 계산 했다.

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

그림8. Stable Diffusion Model Loss Function

- Stable Diffusion Model은 잠재 벡터와 시점 t 그리고 조건을 넣어 Loss 식을 계산한다.

Table 해석

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	N_{params}	
CogView† [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE† [109]	26.94	<u>26.02</u>	75M	
GLIDE* [59]	<u>12.24</u>	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	**11.84**	-	4B	c.f.g for AR models [98] $s = 5$
LDM-KL-8	23.31	$20.03_{\pm 0.33}$	1.45B	250 DDIM steps
*LDM-KL-8-G**	12.63	$\mathbf{30.29}_{\pm 0.42}$	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

Table 2. Evaluation of text-conditional image synthesis on the 256×256-sized MS-COCO [51] dataset: with 250 DDIM [84] steps our model is on par with the most recent diffusion [59] and autoregressive [26] methods despite using significantly less parameters. †/*:Numbers from [109]/ [26]

- IS : 얼마나 예측을 잘하는지, 얼마나 명확한 확률 분포를 만드는지, 높을수록 좋다
- FID : 생성된 이미지와 이미지 집합 사이의 거리, 낮을수록 좋다.
- 테이블 설명을 보면 파라미터 대비 성능이 좋다는 것을 강조한다.
- 조건은 텍스트이다.

Table 해석

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
BigGan-deep [3]	6.95	$203.6_{\pm 2.6}$	**0.87**	0.28	340M	-
ADM [15]	10.94	100.98	0.69	**0.63**	554M	250 DDIM steps
ADM-G [15]	4.59	186.7	0.82	0.52	608M	250 DDIM steps
LDM-4 (ours)	10.56	$103.49_{\pm 1.24}$	0.71	0.62	400M	250 DDIM steps
LDM-4-G (ours)	**3.60**	$\mathbf{247.67}_{\pm 5.59}$	**0.87**	0.48	400M	250 steps, c.f.g [32], $s = 1.5$

Table 3. Comparison of a class-conditional ImageNet *LDM* with recent state-of-the-art methods for class-conditional image generation on ImageNet [12]. A more detailed comparison with additional baselines can be found in D.4, Tab. 10 and F. *c.f.g.* denotes classifier-free guidance with a scale s as proposed in [32].

- IS : 얼마나 예측을 잘하는지, 얼마나 명확한 확률 분포를 만드는지, 높을수록 좋다
- FID : 생성된 이미지와 이미지 집합 사이의 거리, 낮을수록 좋다.
- 이 조건에서는 높은 성능을 보이고 있다. (G는 classifier scale을 추가한 것인데, 무엇을 의미하는 것인지 명확히 모르겠음.)
- 조건은 바운딩 박스 + 클래스이다.
- 그 외 inpating, super resolution 등의 작업에서도 높은 성능을 보임.

결론

- 논문에서는 잠재 확산 모델을 제시했다. 이는 노이즈 제거 확산 모델의 훈련 및 샘플링 효율을 품질 저하 없이 크게 향상시키는 간단하고 효율적인 방법이다.

- Cross Attention 메커니즘을 기반으로 작업 별 아키텍처 없이 SOTA 모델들에 비해 파라미터 대비 괜찮은 결과를 보여주었다.

Open Questions

- 1. Stable Diffusion Model의 활용방안은?

- 2. 2개의 condition을 넣는 구조를 만들면 어떨까?
- (예 : 이미지 + text)

- 3. 학습되지 않은 이미지를 만들고 싶을 때는 어떻게 해야 할까?
- (예 : 한복)