

Accelerating Diffusion Models: Consistency Models and Hybrid Approach

Open DMQA Seminar
2023.12.15

조한샘

발표자 소개



- 조한샘
 - ✓ Data Mining & Quality Analytics Lab
 - ✓ 석·박통합과정 (2020.09~)
- 관심 연구 분야
 - ✓ Diffusion Models
 - ✓ Image Editing
- Contact
 - ✓ chosam95@korea.ac.kr

CONTENTS

- ◆ Consistency Models
 - Consistency Models
 - Latent Consistency Models
- ◆ Hybrid Approach
 - UFOGen

Introduction

Diffusion Models



"Michelangelo style statue of dog
reading news on a cell phone"

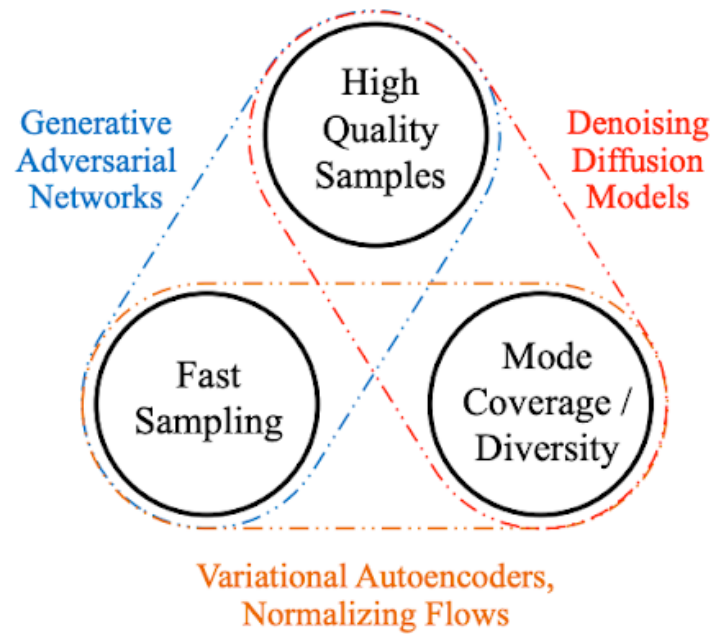


"A hamster driving a scooter"

<https://ml.cs.tsinghua.edu.cn/prolificdreamer/>
<https://emu-video.metademolab.com/>

Introduction

Diffusion Models



SDXL Turbo(1step)

<https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/>
Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2023). Adversarial Diffusion Distillation. arXiv preprint arXiv:2311.17042.

Consistency Models

Consistency Models

Consistency Models

- ICML 2023, OpenAI
- 2023년 12월 15일 기준 127회 인용

Consistency Models

Yang Song¹ Prafulla Dhariwal¹ Mark Chen¹ Ilya Sutskever¹

Abstract

Diffusion models have significantly advanced the fields of image, audio, and video generation, but they depend on an iterative sampling process that causes slow generation. To overcome this limitation, we propose *consistency models*, a new family of models that generate high quality samples by directly mapping noise to data. They support fast one-step generation by design, while still allowing multistep sampling to trade compute for sample quality. They also support zero-shot data editing, such as image inpainting, colorization, and super-resolution, without requiring explicit training on these tasks. Consistency models can be trained either by distilling pre-trained diffusion models, or as standalone generative models altogether. Through extensive experiments, we demonstrate that they outperform existing distillation techniques for diffusion models in one- and few-step sampling, achieving the new state-of-the-art FID of 3.55 on CIFAR-10 and 6.20 on ImageNet 64×64 for one-step generation. When trained in isolation, consistency models become a new family of generative models that can outperform existing one-step, non-adversarial generative models on standard benchmarks such as CIFAR-10, ImageNet 64×64 and LSUN 256×256 .

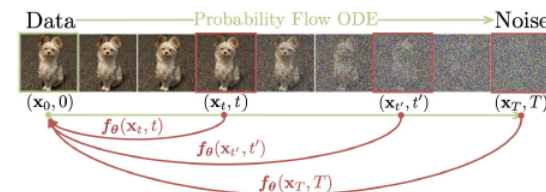


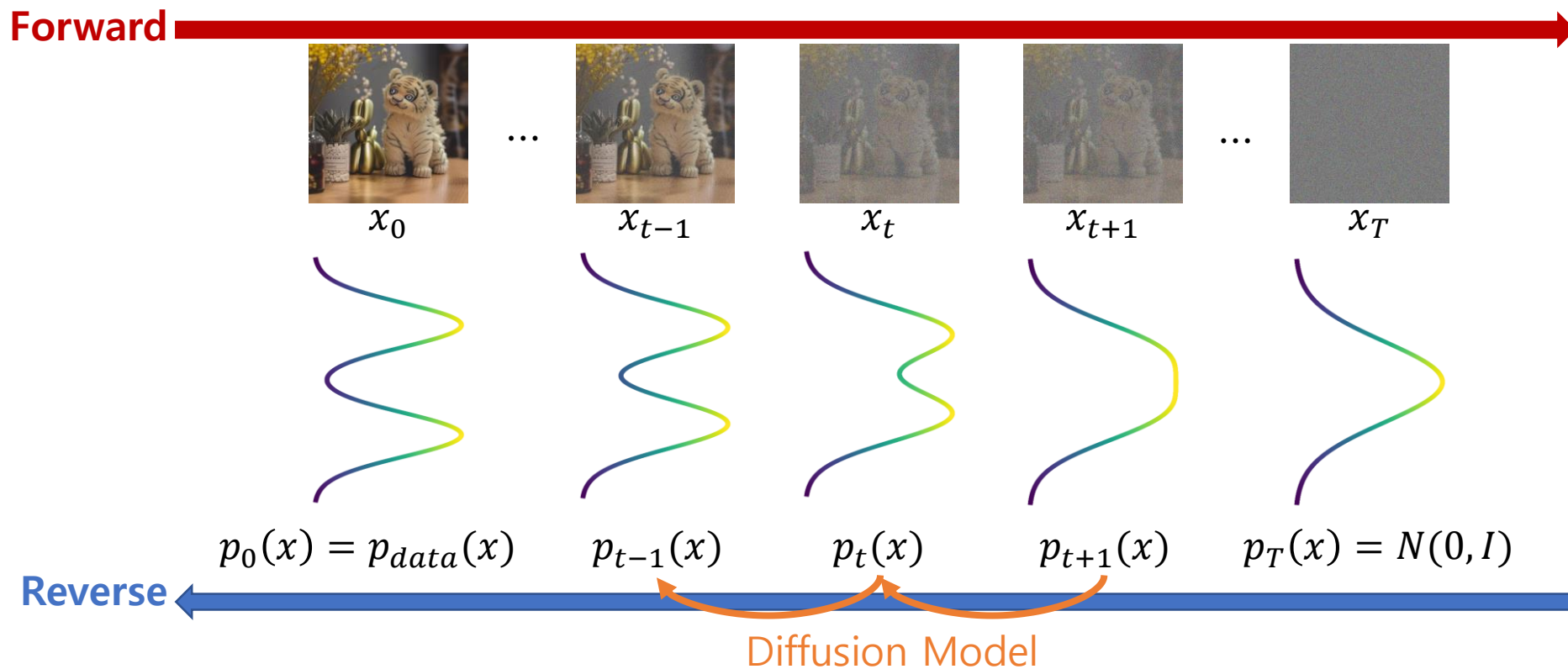
Figure 1: Given a **Probability Flow (PF) ODE** that smoothly converts data to noise, we learn to map any point (e.g., x_t , $x_{t'}$, and x_T) on the ODE trajectory to its origin (e.g., x_0) for generative modeling. Models of these mappings are called **consistency models**, as their outputs are trained to be consistent for points on the same trajectory.

2022b;a). A key feature of diffusion models is the iterative sampling process which progressively removes noise from random initial vectors. This iterative process provides a flexible trade-off of compute and sample quality, as using extra compute for more iterations usually yields samples of better quality. It is also the crux of many zero-shot data editing capabilities of diffusion models, enabling them to solve challenging inverse problems ranging from image inpainting, colorization, stroke-guided image editing, to Computed Tomography and Magnetic Resonance Imaging (Song & Ermon, 2019; Song et al., 2021; 2022; 2023; Kawar et al., 2021; 2022; Chung et al., 2023; Meng et al., 2021).

Consistency Models

Preliminary: DDPM

- **Forward process:** data \rightarrow noise / 노이즈 스케줄에 따라 노이즈 추가
- **Reverse process:** noise \rightarrow data / 모델을 통해 노이즈 제거

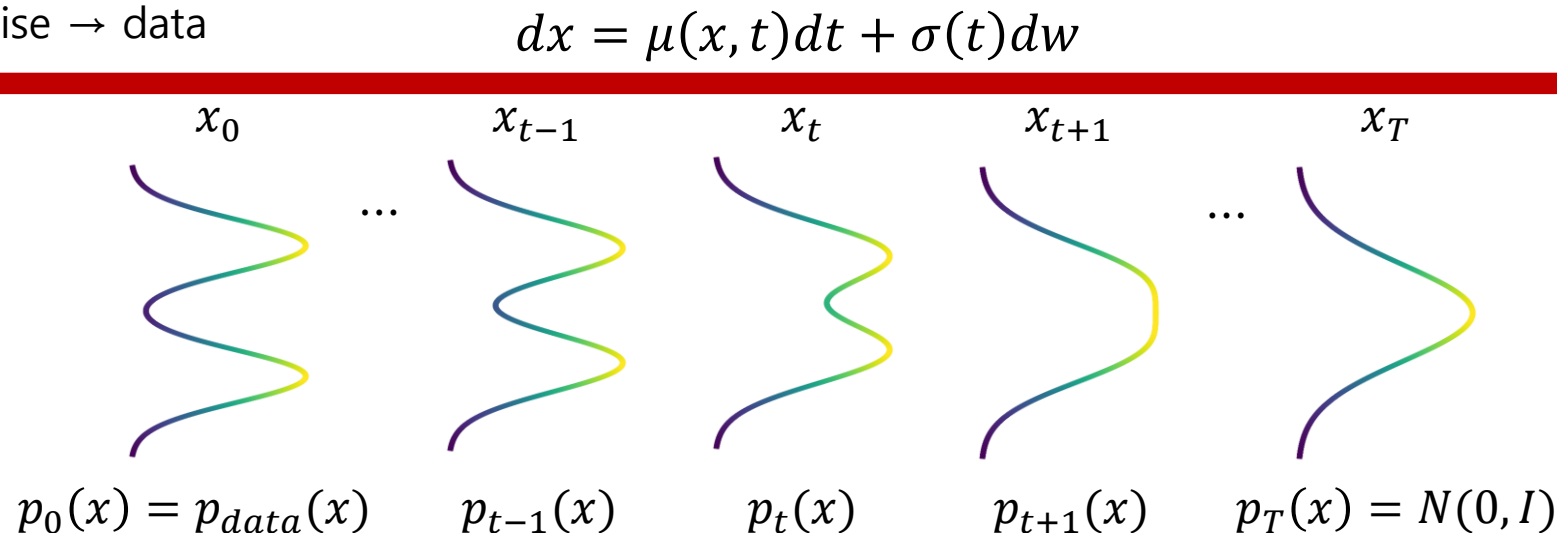


Consistency Models

Preliminary : Score-SDE

- Forward SDE: data \rightarrow noise
- Reverse SDE: noise \rightarrow data

Forward



Reverse

$dx = [\mu(x, t) - \sigma(t)^2 \nabla_x \log p_t(x)]dt + \sigma(t)d\bar{w}$ Diffusion Model

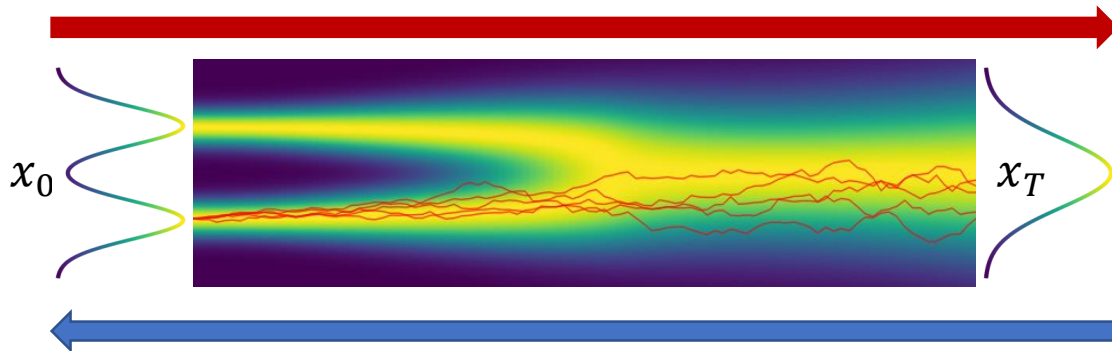
Consistency Models

Preliminary : Score-SDE

- Probability Flow ODE (PF ODE): Forward SDE와 동일한 확률 분포를 갖는 ODE

Forward SDE

$$dx = \mu(x, t)dt + \sigma(t)dw$$

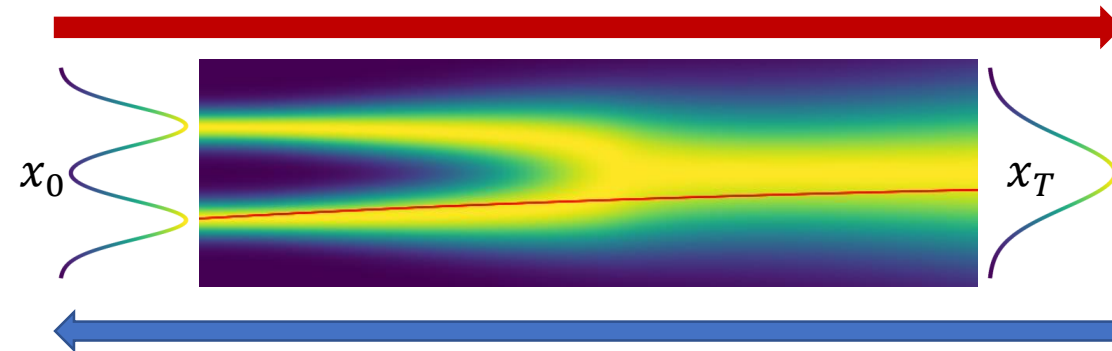


Reverse

$$dx = [\mu(x, t) - \sigma(t)^2 \nabla_x \log p_t(x)]dt + \sigma(t)d\bar{w}$$

PF ODE

$$dx = \left[\mu(x, t) - \frac{1}{2} \sigma(t)^2 \nabla_x \log p_t(x) \right] dt$$

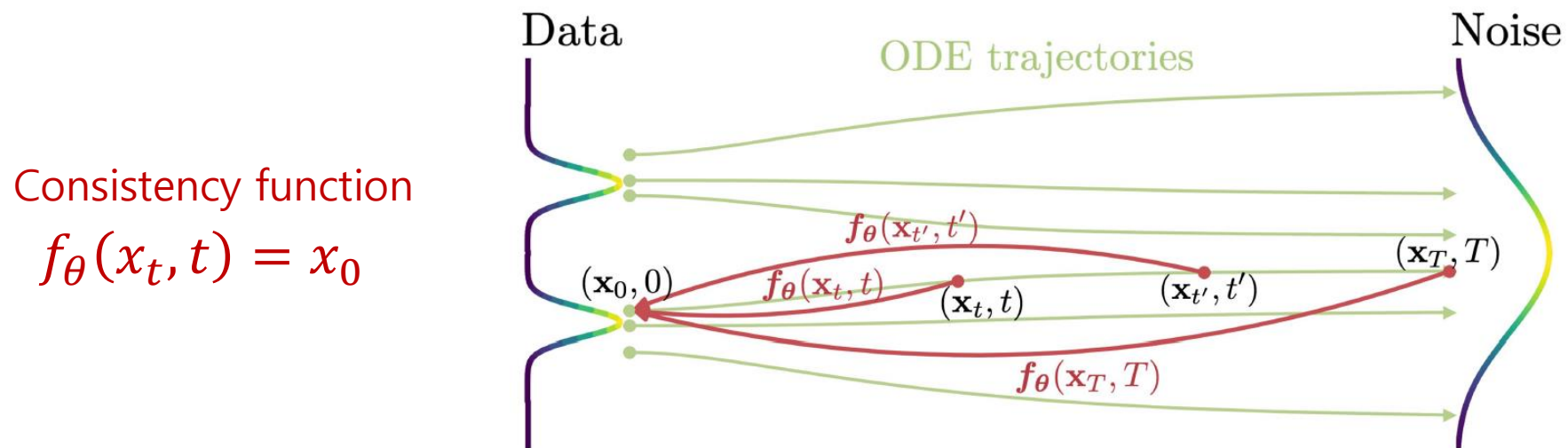


$$dx = \left[\mu(x, t) - \frac{1}{2} \sigma(t)^2 \nabla_x \log p_t(x) \right] dt$$

Consistency Models

Consistency Models

- Consistency function: 동일한 PF ODE trajectory에 위치하는 x_t 를 입력 받았을 때 x_0 를 return
- Self-consistency: 임의의 x_t 에 대해서 동일한 출력값을 가져야함 ($f_\theta(x_t, t) = f_\theta(x_{t'}, t')$ for all $t, t' \in [0, T]$)
- Boundary condition: x_0 를 입력하면 identity function이 되어야 한다 ($f_\theta(x_0, 0) = x_0$)

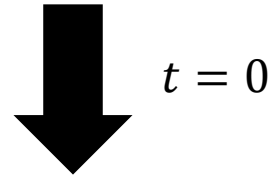


Consistency Models

Consistency Models

- $F_\theta(x_t, t)$: deep neural network
- $c_{skip}(t)$: $c_{skip}(0) = 1$ 을 만족하는 미분 가능한 함수
- $c_{out}(t)$: $c_{out}(0) = 0$ 을 만족하는 미분 가능한 함수

$$f_\theta(x_t, t) = c_{skip}(t)x_t + c_{out}(t)F_\theta(x_t, t)$$

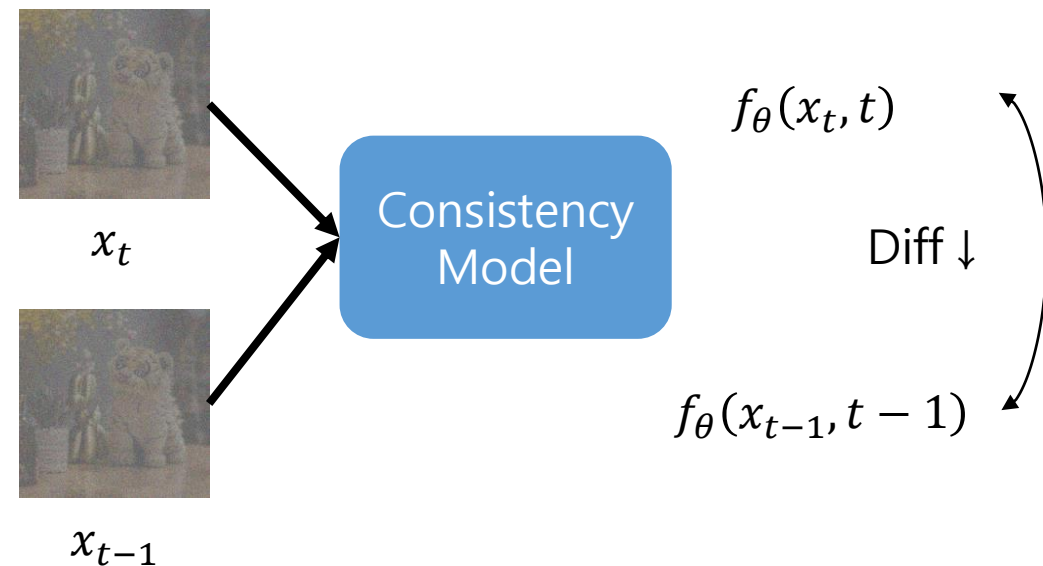
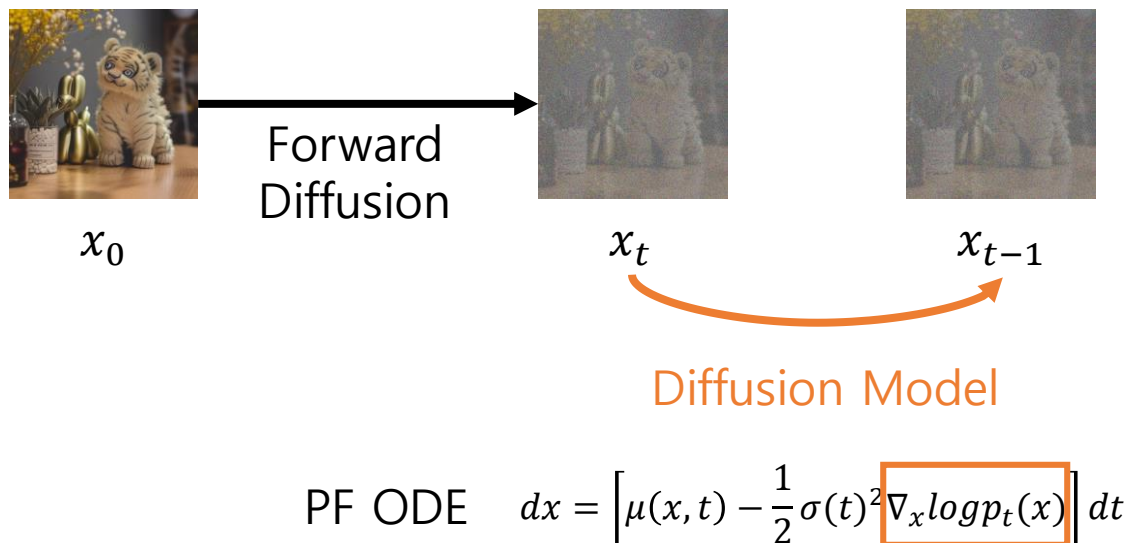


Boundary condition $f_\theta(x_0, t) = c_{skip}(0)x_0 + c_{out}(0)F_\theta(x_t, t)$

Consistency Models

Consistency Models – Consistency Distillation (CD)

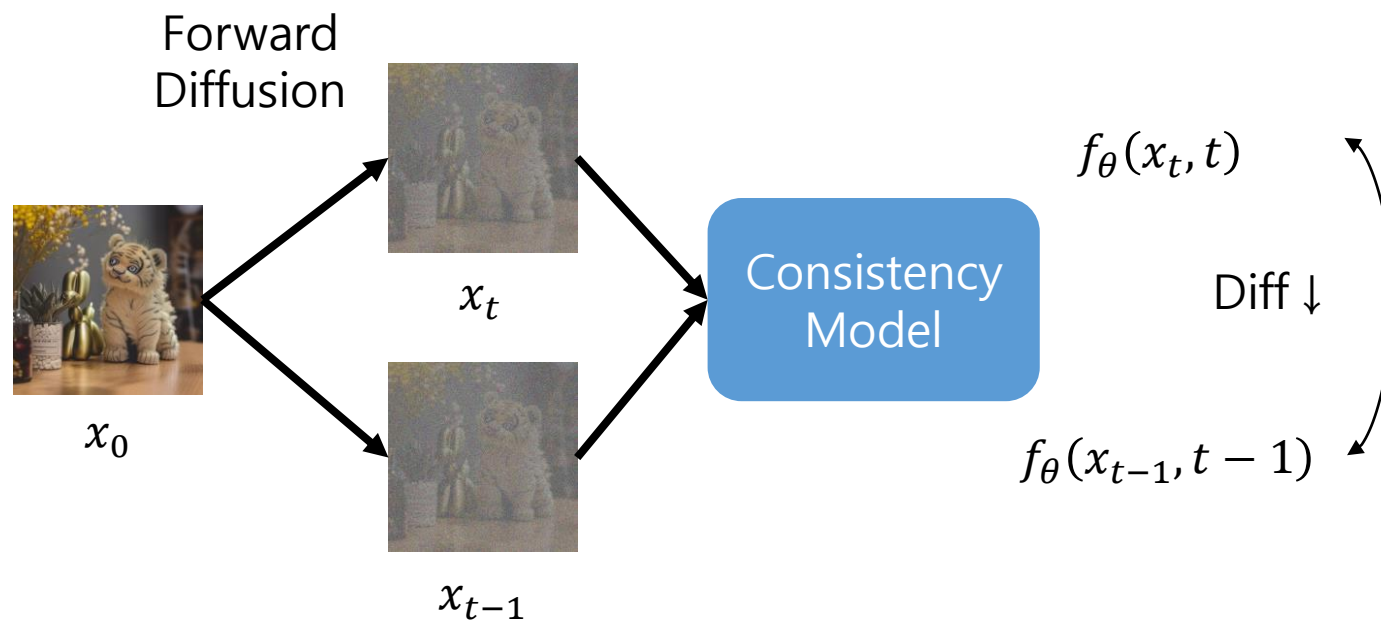
- 이미지에 forward diffusion을 통해 노이즈 추가
- x_t 로부터 동일한 PF ODE 상에 존재하는 x_{t-1} 획득
- Self consistency를 만족하기 위해 $f_\theta(x_t, t)$ 와 $f_\theta(x_{t-1}, t-1)$ 차이 최소화



Consistency Models

Consistency Models – Consistency Training (CT)

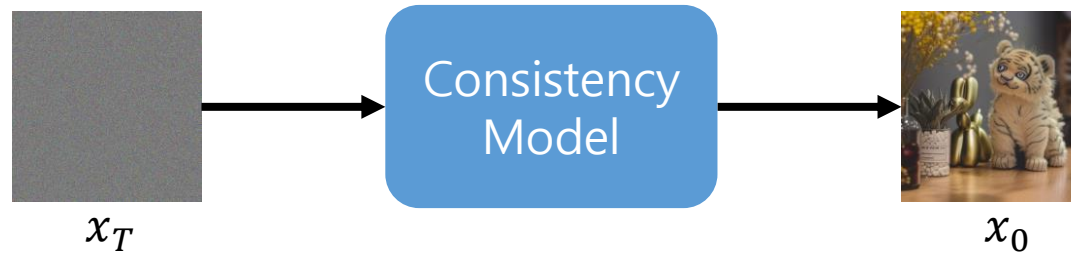
- 사전에 학습된 diffusion model 없이 consistency model 학습 가능
- 단독으로 학습이 가능하기 때문에 새로운 형태의 생성모델로 볼 수 있음
- CD와 CT의 loss가 동일함을 증명 (Theorem 2)



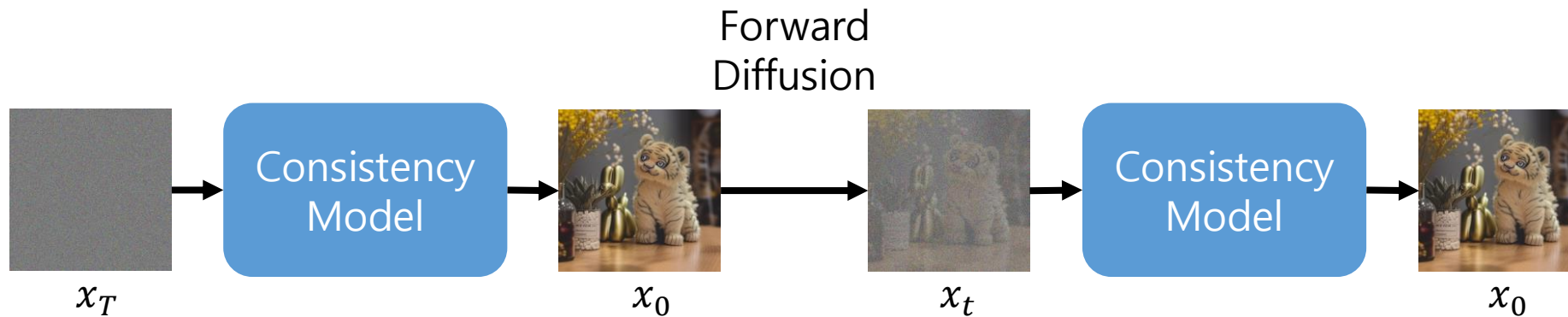
Consistency Models

Consistency Models - Sampling

- One-step sampling: x_T 를 $f(x,t)$ 에 넣으면 이미지 생성 가능



- Multistep Sampling: 노이즈를 추가한 이후 다시 x_0 생성



Consistency Models

Consistency Models - Experiments

- 기존 diffusion model 대비 좋은 성능

| METHOD | NFE (↓) | FID (↓) | IS (↑) |
|---|---------|-------------|-------------|
| Diffusion + Samplers | | | |
| DDIM (Song et al., 2020) | 50 | 4.67 | |
| DDIM (Song et al., 2020) | 20 | 6.84 | |
| DDIM (Song et al., 2020) | 10 | 8.23 | |
| DPM-solver-2 (Lu et al., 2022) | 10 | 5.94 | |
| DPM-solver-fast (Lu et al., 2022) | 10 | 4.70 | |
| 3-DEIS (Zhang & Chen, 2022) | 10 | 4.17 | |
| Diffusion + Distillation | | | |
| Knowledge Distillation* (Luhman & Luhman, 2021) | 1 | 9.36 | |
| DFNO* (Zheng et al., 2022) | 1 | 4.12 | |
| 1-Rectified Flow (+distill)* (Liu et al., 2022) | 1 | 6.18 | 9.08 |
| 2-Rectified Flow (+distill)* (Liu et al., 2022) | 1 | 4.85 | 9.01 |
| 3-Rectified Flow (+distill)* (Liu et al., 2022) | 1 | 5.21 | 8.79 |
| PD (Salimans & Ho, 2022) | 1 | 8.34 | 8.69 |
| CD | 1 | 3.55 | 9.48 |
| PD (Salimans & Ho, 2022) | 2 | 5.58 | 9.05 |
| CD | 2 | 2.93 | 9.75 |

EDM
(35step)



CT
(1step)



CT
(2step)



Consistency Models

Consistency Models - Experiments

- Zero-shot image editing도 가능



(a) *Left: The gray-scale image. Middle: Colorized images. Right: The ground-truth image.*



(b) *Left: The downsampled image (32×32). Middle: Full resolution images (256×256). Right: The ground-truth image (256×256).*



(c) *Left: A stroke input provided by users. Right: Stroke-guided image generation.*

Consistency Models

Latent Consistency Models

- Arxiv 10월 6일 공개, Tsinghua Univ.
- 2023년 12월 15일 기준 9회 인용

LATENT CONSISTENCY MODELS: SYNTHESIZING HIGH-RESOLUTION IMAGES WITH FEW-STEP INFERENCE

Simian Luo* Yiqin Tan* Longbo Huang[†] Jian Li[†] Hang Zhao[†]
Institute for Interdisciplinary Information Sciences, Tsinghua University
{luosm22, tyq22}@mails.tsinghua.edu.cn
{longbohuang, lijian83, hangzhao}@tsinghua.edu.cn

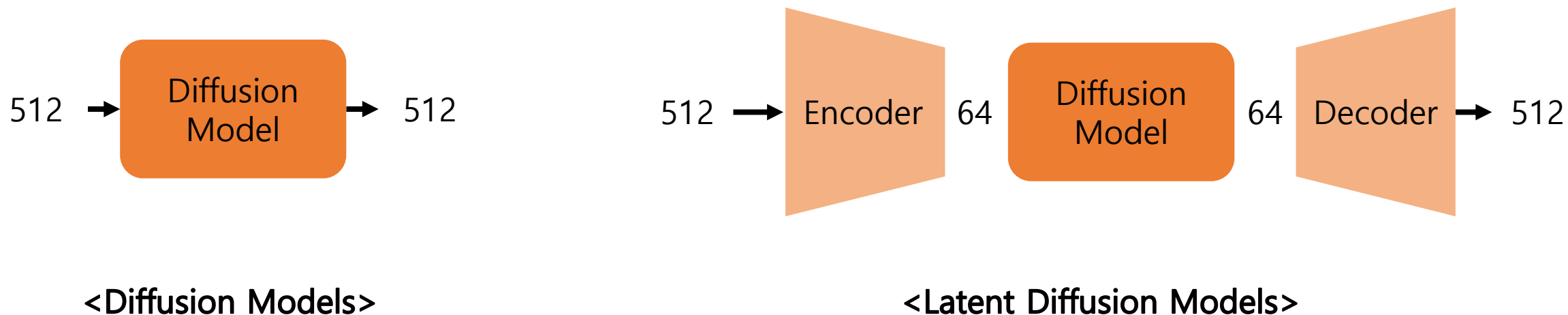
ABSTRACT

Latent Diffusion models (LDMs) have achieved remarkable results in synthesizing high-resolution images. However, the iterative sampling process is computationally intensive and leads to slow generation. Inspired by Consistency Models (Song et al., 2023), we propose Latent Consistency Models (LCMs), enabling swift inference with minimal steps on any pre-trained LDMs, including Stable Diffusion (Rombach et al., 2022). Viewing the guided reverse diffusion process as solving an augmented probability flow ODE (PF-ODE), LCMs are designed to directly predict the solution of such ODE in latent space, mitigating the need for numerous iterations and allowing rapid, high-fidelity sampling. Efficiently distilled from pre-trained classifier-free guided diffusion models, a high-quality 768×768 2~4-step LCM takes only 32 A100 GPU hours for training. Furthermore, we introduce Latent Consistency Fine-tuning (LCF), a novel method that is tailored for fine-tuning LCMs on customized image datasets. Evaluation on the LAION-5B-Aesthetics dataset demonstrates that LCMs achieve state-of-the-art text-to-image generation performance with few-step inference. Project Page: <https://latent-consistency-models.github.io/>

Consistency Models

Diffusion to Latent Diffusion

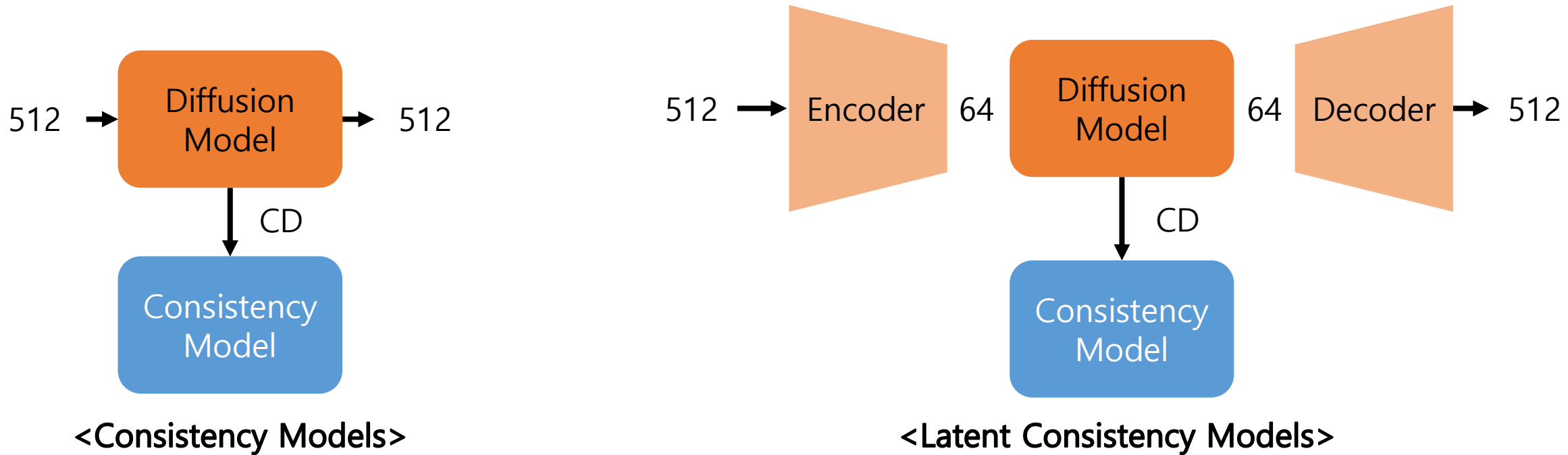
- 고해상도 이미지를 생성하기 위해서 diffusion 과정을 latent space상에서 진행
- Text-to-image 모델로 확장
- Diffusion models → Latent diffusion models



Consistency Models

Consistency to Latent Consistency

- 고해상도 이미지를 생성하기 위해서 diffusion 과정을 latent space상에서 진행
- Text-to-image 모델로 확장
- Consistency models → Latent consistency models



Consistency Models

Latent Consistency Models

- Classifier-free Guidance (CFG): condition이 들어간 output에 큰 가중치를 부여 이미지의 퀄리티를 향상하는テクニック

$$\tilde{\epsilon}_{\theta}(z_t, w, c, t) = (1 + w)\epsilon_{\theta}(z_t, c, t) - w\epsilon_{\theta}(z_t, \emptyset, t)$$

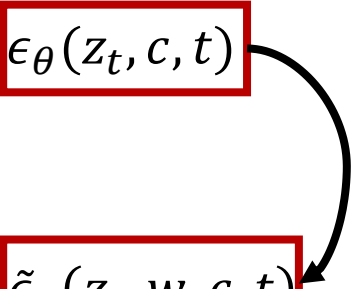
CFG

Conditional
diffusion model

Unconditional
diffusion model

- Augmented PF ODE: CFG가 포함된 diffusion model의 아웃풋을 활용하여 CD 진행

PF ODE

$$\frac{dx}{dt} = f(t)z_t + \frac{g^2(t)}{2\sigma_t} \epsilon_{\theta}(z_t, c, t)$$


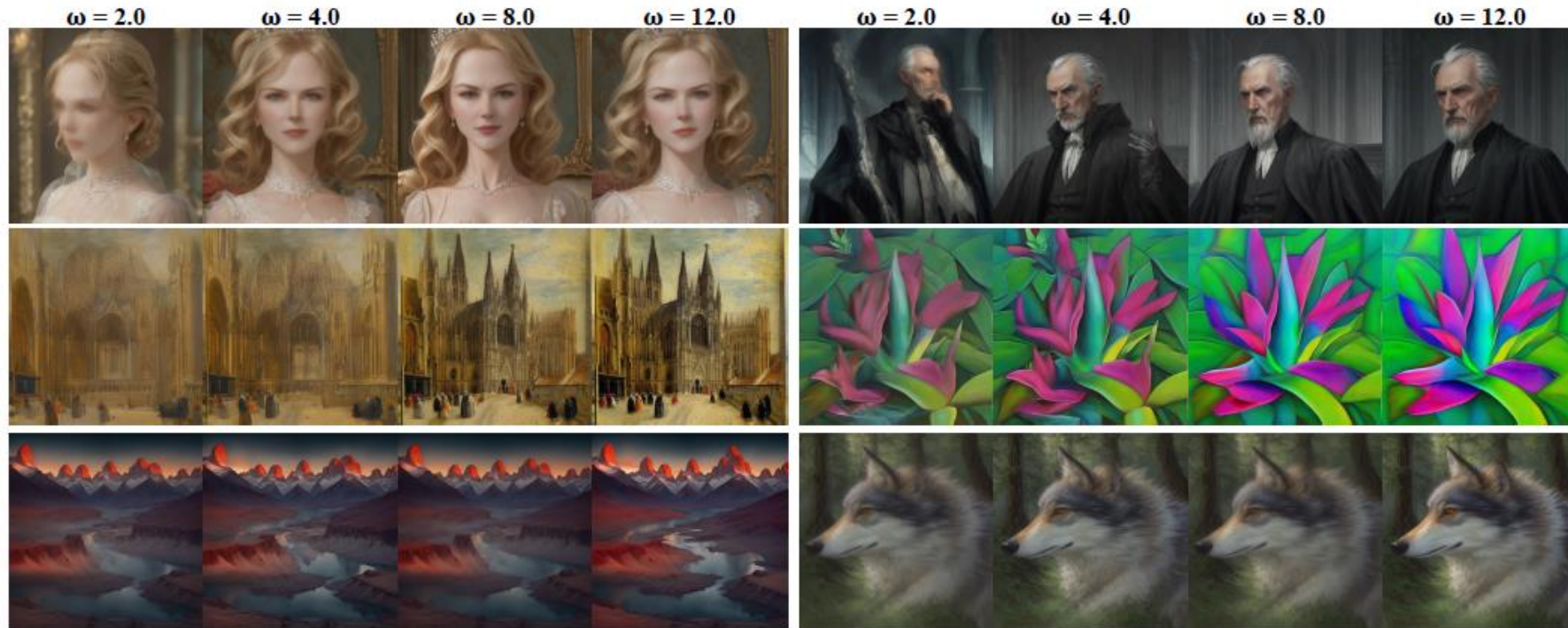
Augmented PF ODE

$$\frac{dx}{dt} = f(t)z_t + \frac{g^2(t)}{2\sigma_t} \tilde{\epsilon}_{\theta}(z_t, w, c, t)$$

Consistency Models

Latent Consistency Models - Experiments

- 기존 diffusion model처럼 CFG scale에 따라 이미지 퀄리티가 높아지는 모습



LCM 4step



Hybrid Approach

Hybrid Approach

UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs

- Arxiv 11월 14일 공개, Google
- 2023년 12월 15일 기준 3회 인용



UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs

Yanwu Xu^{*,[1,2]}†, Yang Zhao^[1]‡, Zhisheng Xiao^[1]‡, Tingbo Hou^[1]

¹ Google

{yanwuxu, yzhaoeric, zsxiao, tingbo}@google.com

² Department of Electrical Computer Engineering, Boston University

yanwuxu@bu.edu

Abstract

Text-to-image diffusion models have demonstrated remarkable capabilities in transforming text prompts into coherent images, yet the computational cost of the multi-step inference remains a persistent challenge. To address this issue, we present UFOGen, a novel generative model designed for ultra-fast, one-step text-to-image generation. In contrast to conventional approaches that focus on improving samplers or employing distillation techniques for diffusion models, UFOGen adopts a hybrid methodology, integrating diffusion models with a GAN objective. Leveraging a newly introduced diffusion-GAN objective and initialization with pre-trained diffusion models, UFOGen excels in efficiently generating high-quality images conditioned on textual descriptions in a single step. Beyond traditional text-to-image generation, UFOGen showcases versatility in applications. Notably, UFOGen stands among the pioneering models enabling one-step text-to-image generation and diverse downstream tasks, presenting a significant advancement in the landscape of efficient generative models.

ing [5, 13, 65]. Yet, despite their impressive generative quality and wide-ranging utility, diffusion models have a notable limitation: they rely on iterative denoising to generate final samples, which leads to slow generation speeds. The slow inference and the consequential computational demands of large-scale diffusion models pose significant impediments to their deployment.

In the seminal work by Song *et al.* [56], it was revealed that sampling from a diffusion model is equivalent to solving the probability flow ordinary differential equation (PF-ODE) associated with the diffusion process. Presently, the majority of research aimed at enhancing the sampling efficiency of diffusion models centers on the ODE formulation. One line of work seeks to advance numerical solvers for the PF-ODE, with the intention of enabling the solution of the ODE with greater discretization size, ultimately leading to fewer requisite sampling steps [2, 35, 36, 55]. However, the inherent trade-off between step size and accuracy still exists. Given the highly non-linear and complicated trajectory of the PF-ODE, it would be extremely difficult to reduce the number of required sampling steps to a minimal level.

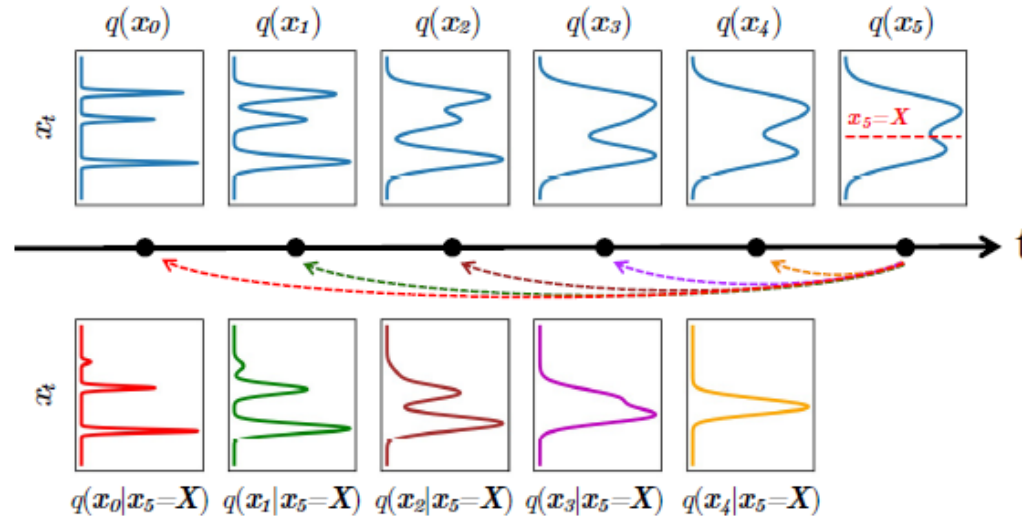
Hybrid Approach

Preliminary: DDGAN

- Reverse process의 step이 커지면 Gaussian이라는 가정이 깨짐
- Adversarial loss를 활용해 step 사이즈가 커질 때의 분포를 추정



Marginal Diffused
Data Distributions



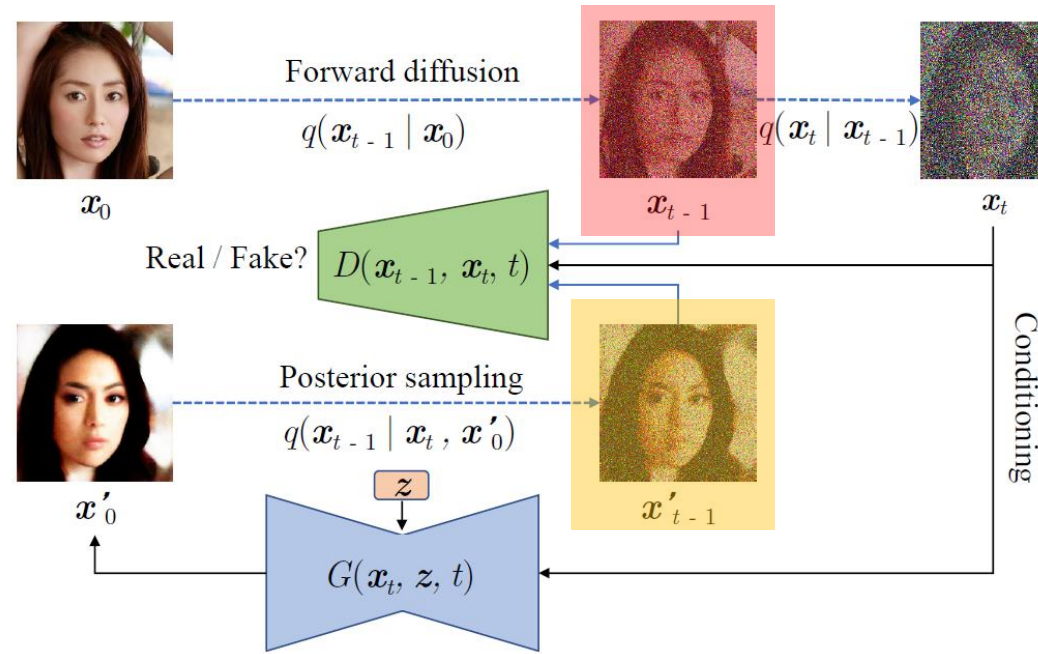
$$D_{adv}(q(x_{t-1}|x_t)||p_{\theta}(x'_{t-1}|x_t))$$

Hybrid Approach

Preliminary: DDGAN

- x_{t-1} : 데이터에서 얻어진 노이즈 이미지
- x'_{t-1} : Generator가 생성한 노이즈 이미지
- Discriminator 진짜 가짜 구분 \rightarrow Generator($p_\theta(x'_{t-1}|x_t)$)가 $q(x_{t-1}|x_t)$ 와 분포 학습

$$D_{adv}(q(x_{t-1}|x_t) || p_\theta(x'_{t-1}|x_t))$$



Hybrid Approach

Preliminary: SIDDMs

- DDGAN의 학습 불안전성을 개선하기 위해서 새로운 loss function을 도입

DDGAN

$$D_{adv}(q(x_{t-1}|x_t)||p_{\theta}(x'_{t-1}|x_t))$$



SIDDMs

$$D_{adv}(q(x_{t-1})||p_{\theta}(x'_{t-1})) + \lambda_{KL}KL(p_{\theta}(x_t|x'_{t-1})||q(x_t|x_{t-1}))$$

Hybrid Approach

UFOGen – Loss

- **Term1**: $q(x_{t-1})$ 과 $p_\theta(x'_{t-1})$ align \rightarrow $q(x_0)$ 과 $p_\theta(x'_0)$ align 하는것과 동일 (Appendix A.2.1.)
- **Term2**: x_{t-1}, x'_{t-1} 차이를 줄이는 것은 x_0, x'_0 차이를 줄이는것과 동일 (Appendix A.2.2.)
- 실험적으로 x_0, x'_0 차이를 줄이는 것이 좋은 성능을 보임

$$D_{adv}(q(x_{t-1})||p_\theta(x'_{t-1})) + \lambda_{KL} KL(p_\theta(x_t|x'_{t-1})||q(x_t|x_{t-1}))$$

Generator가 데이터 분포와 align



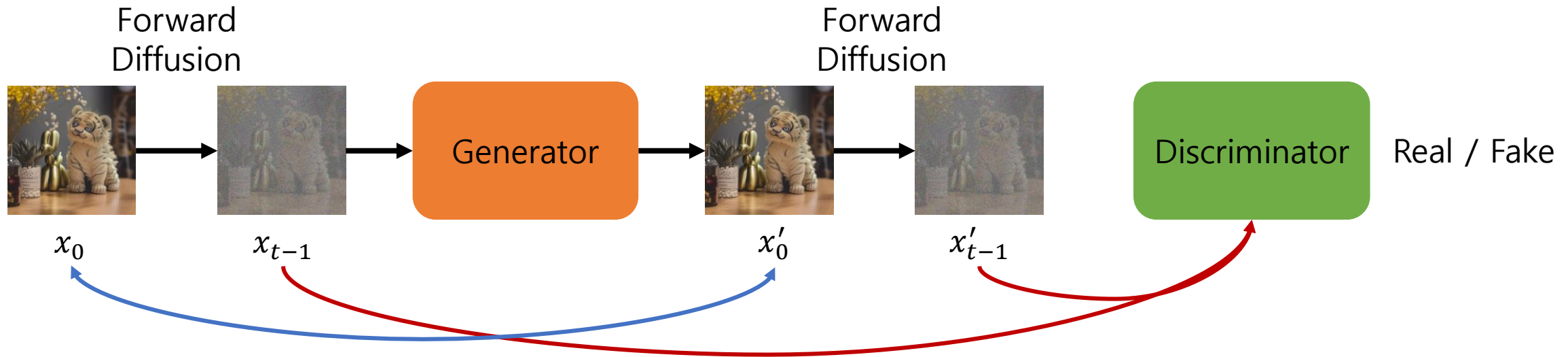
이미지 reconstruction

$$\frac{(1 - \beta_t)\bar{\alpha}_{t-1}\|x'_0 - x_0\|^2}{2\beta_t}$$

Hybrid Approach

UFOGen – Training

- Generator와 discriminator는 사전 학습된 diffusion model에서부터 학습 시작



$$D_{adv}(q(x_{t-1})||p_{\theta}(x'_{t-1})) + \lambda_{KL}KL(p_{\theta}(x_t|x'_{t-1})||q(x_t|x_{t-1})) \rightarrow \frac{(1 - \beta_t)\bar{\alpha}_{t-1}\|x'_0 - x_0\|^2}{2\beta_t}$$

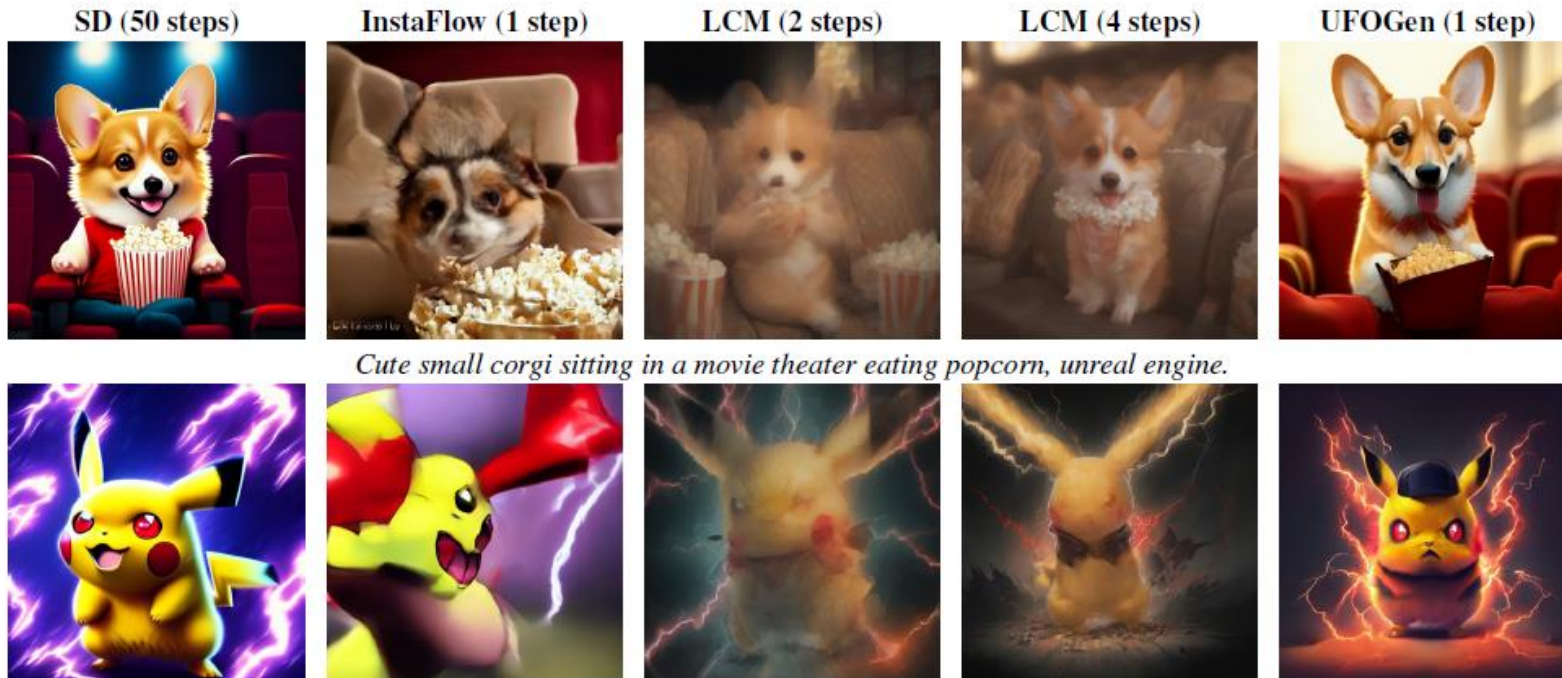
Generator가 데이터 분포와 align

이미지 reconstruction

Hybrid Approach

UFOGen – Experiments

- 1step generation에서도 우수한 성능을 보임



Cute small corgi sitting in a movie theater eating popcorn, unreal engine.

A Pikachu with an angry expression and red eyes, with lightning around it, hyper realistic style.

Hybrid Approach

UFOGen – Experiments

- Image-to-Image: input 이미지에 노이즈 추가 후 Generator를 통해 이미지 생성
- Controllable generation: T2I-adapter를 통해 추가 condition을 활용한 생성

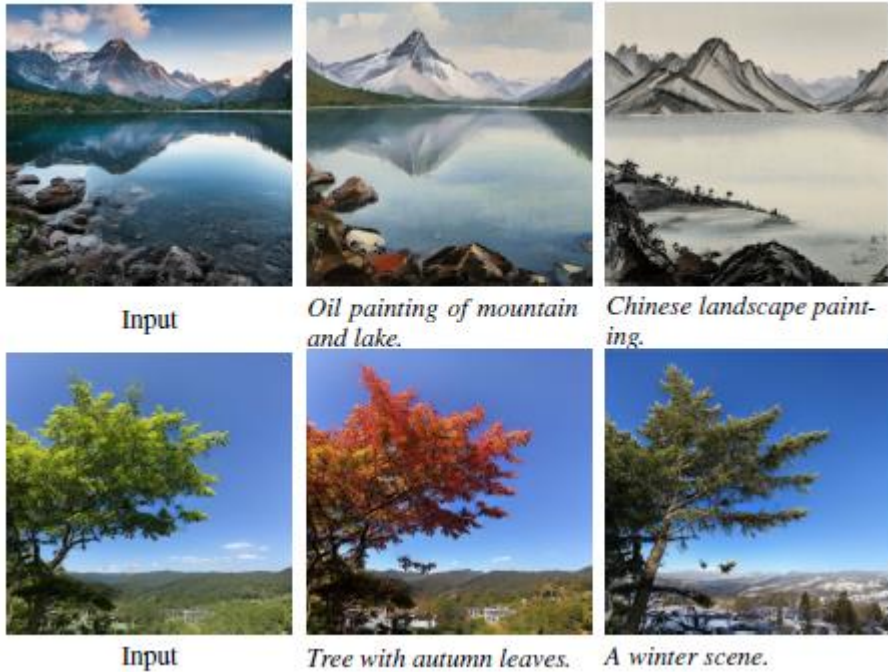
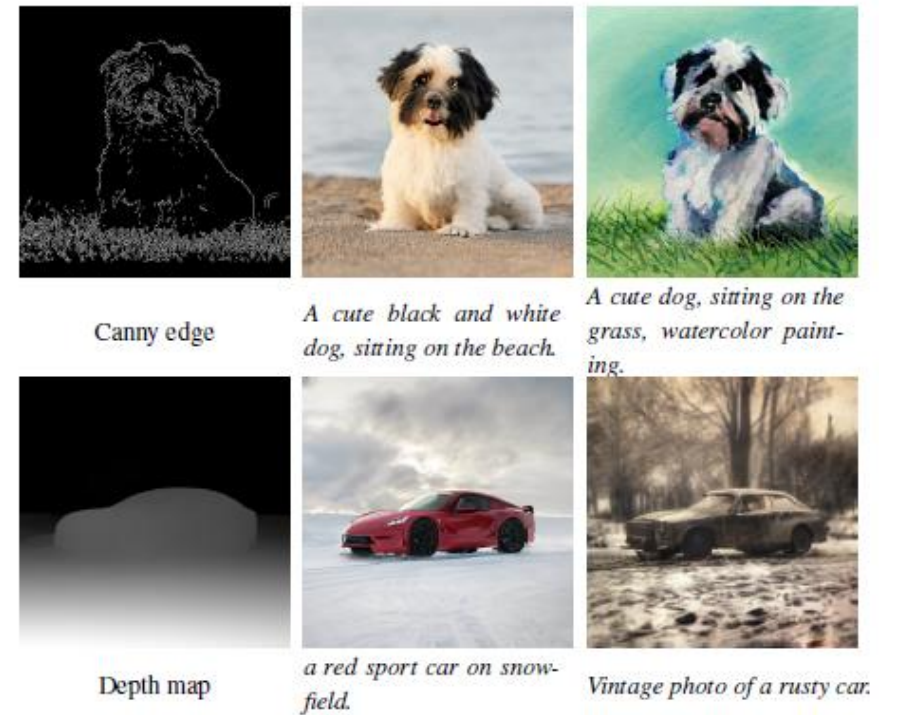


Image-to-Image



Controllable generation

Conclusion

Accelerating Diffusion Models

- Consistency Models
 - ✓ Consistency Models
 - x_t 를 입력 받아 x_0 를 예측하는 consistency model 제안
 - CD: diffusion model을 통해 학습 / CT: 사전정보 없이 학습
 - ✓ Latent Consistency Models
 - Consistency model을 latent space에서 진행함으로써 고해상도 이미지 생성
 - Consistency model을 Text-to-image 모델로 확장
- Hybrid Approach
 - ✓ UFOGen
 - Reverse process의 step 사이즈가 커질 때 Gaussian 분포 가정이 깨짐 (DDGAN)
 - Generator가 데이터 분포를 추정