Lec6 Image Generation

·AIPKU·

人工智能引论实践课 计算机视觉小班

主讲人：刘家瑛

部分讲义来源与致谢

1. David Foster. Generative Deep Learning: Teaching Machines to Paint, Write, Compose and Play. 2019.
2. Alexei Efros. Tutorials on Generative Adversarial Networks. ICCV 2017 Tutorial.
3. Hung-yi Lee, Tutorial for Generative Adversarial Network (GAN), *http://speech.ee.ntu.edu.tw/~tlkagk/slide/Tutorial_HYLee_GAN.pdf*
4. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Nets. NeurIPS 2014.
5. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-image translation with conditional adversarial networks. CVPR 2017.
6. Karsten Kreis, Ruiqi Gao, Arash Vahdat. CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications. *https://cvpr2022-tutorial-diffusion-models.github.io/*

PART 01

生成模型

生成模型 (Generative Model)

- 形式化定义

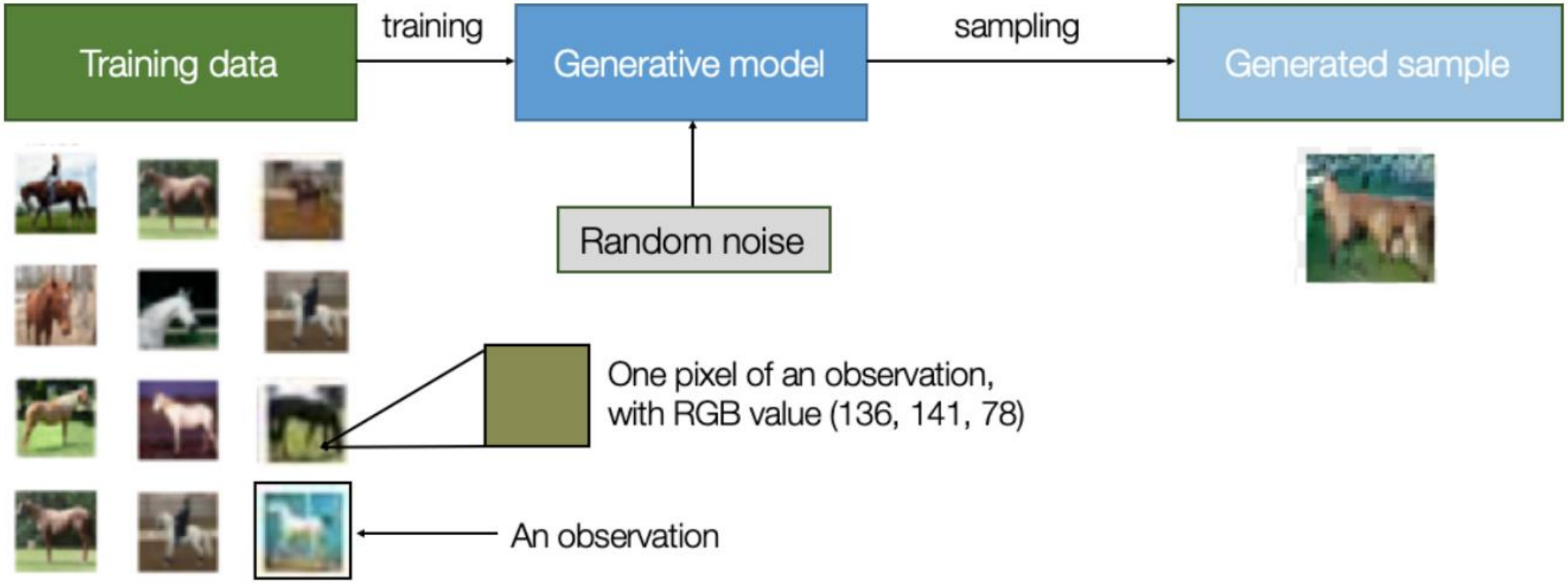
  从概率的角度描述了生成数据集的方法, 从该模型中采样即可生成新数据

- 计算机视觉中

  基于任务给定的条件, 随机生成符合训练用的数据分布的图像的模型

生成模型

- 生成模型过程


Training data
training
Generative model
sampling
Generated sample
Random noise
One pixel of an observation,
with RGB value (136, 141, 78)
An observation

判别模型 (Discriminative Model)

- 监督学习

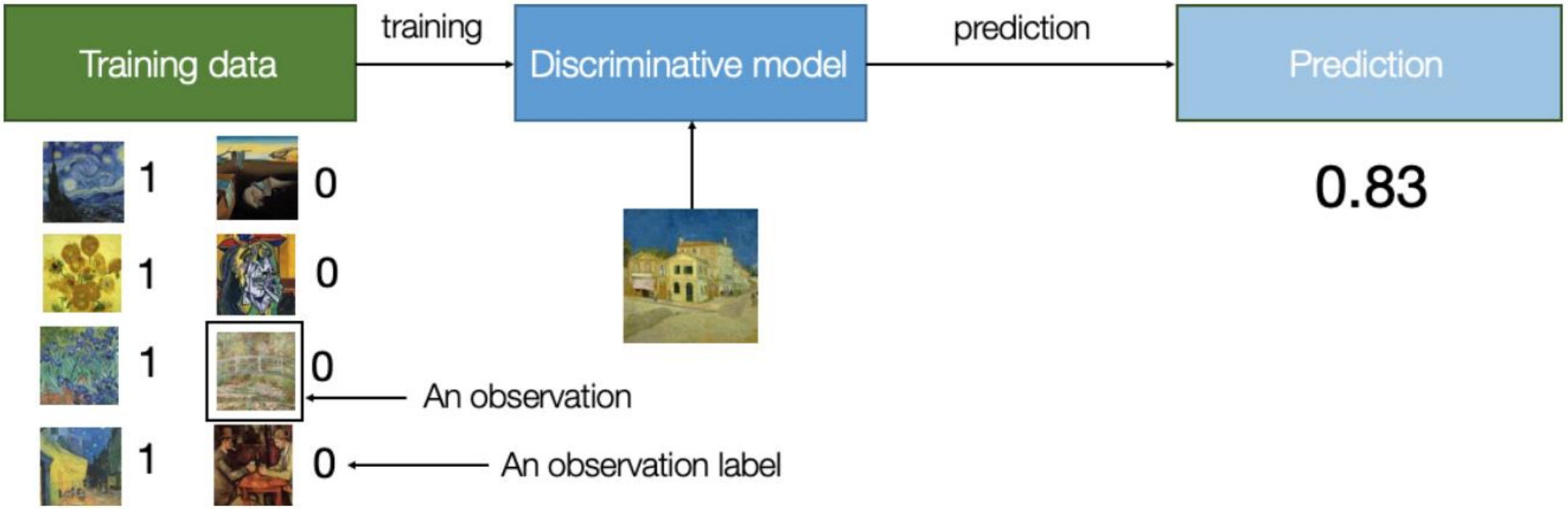
  能够通过带有标签的数据集将输入映射成输出的函数

- 计算机视觉中

  尝试区分真实图像和生成器创建的伪图像的分类器

判别模型 (Discriminative Model)

- 判别模型过程



生成与判别模型

- 数学表示

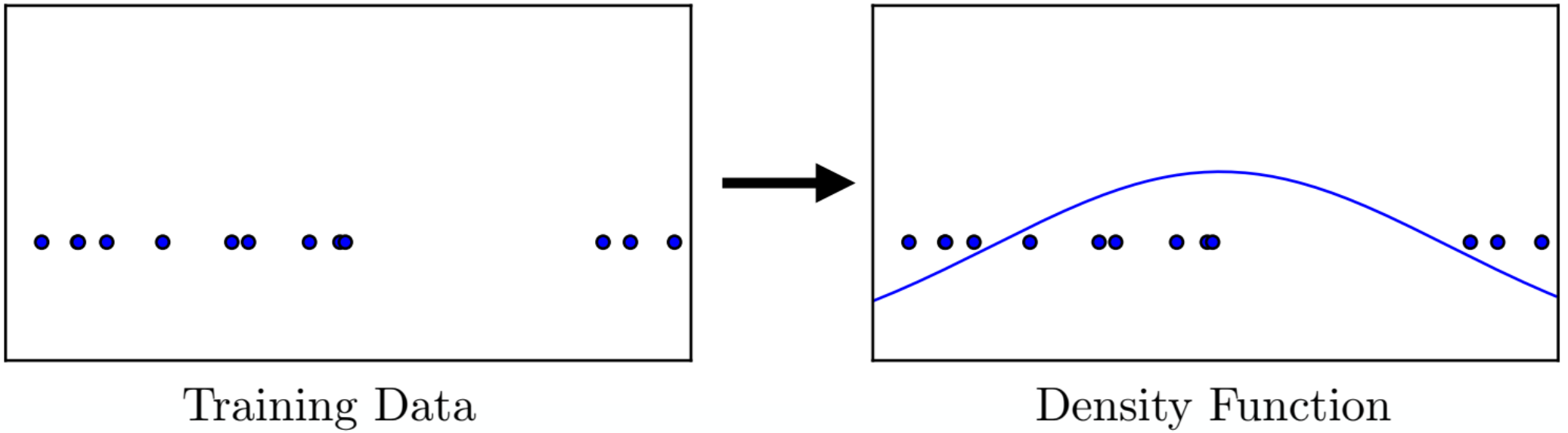
Discriminative modeling estimates $p(y|\mathbf{x})$—the probability of a label y given observation $\mathbf{x}$.

Generative modeling estimates $p(\mathbf{x})$—the probability of observing observation $\mathbf{x}$.

If the dataset is labeled, we can also build a generative model that estimates the distribution $p(\mathbf{x}|y)$.

生成模型

- 数据分布估计 (Density Estimation)



生成模型

- 样本生成 (Sample Generation)

Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

生成模型应用

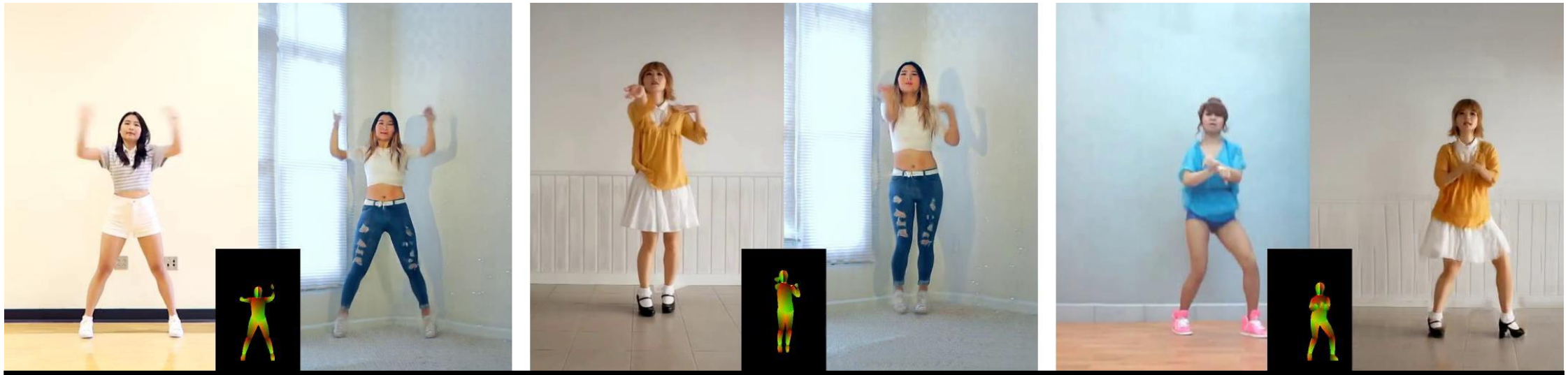
图像风格化

图像生成

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

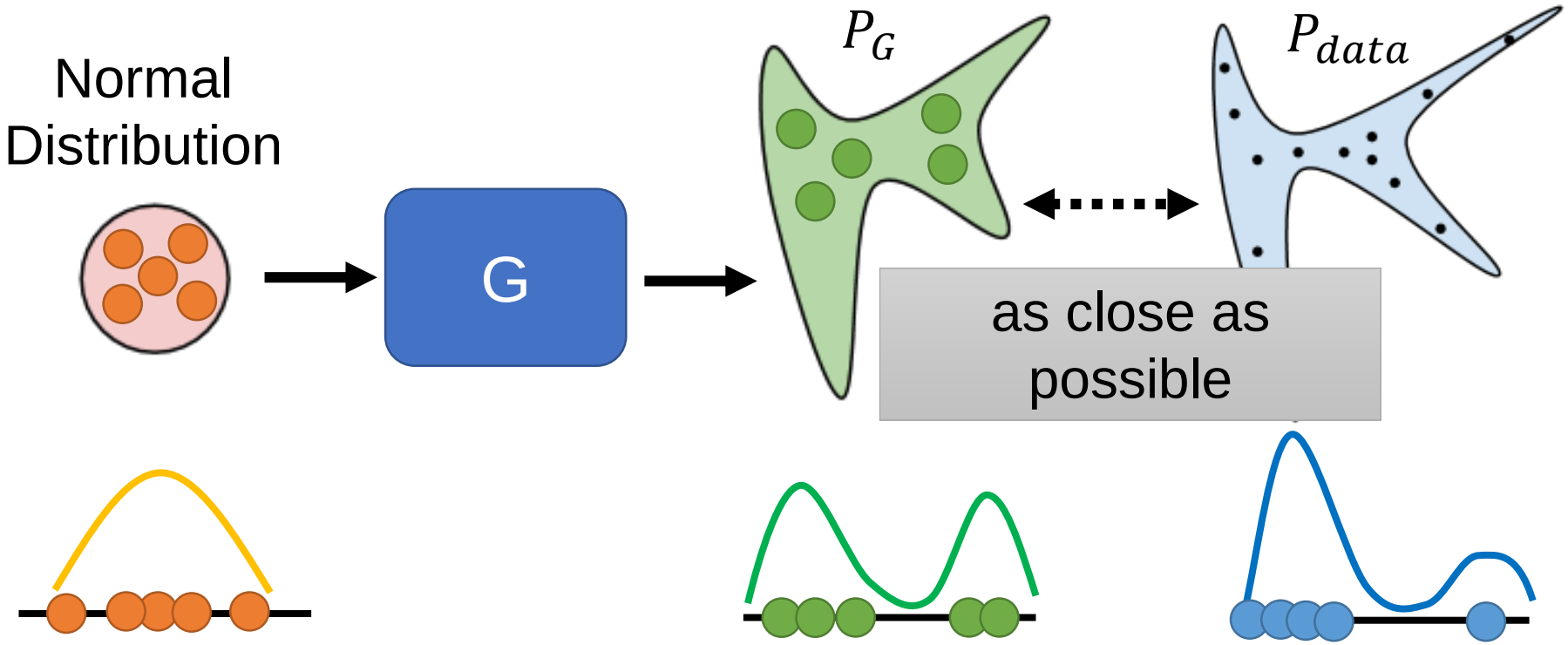
文本-图像生成



视频生成

生成模型的形式化目标

- 使得生成结果的分布与真实数据分布接近



生成模型的种类

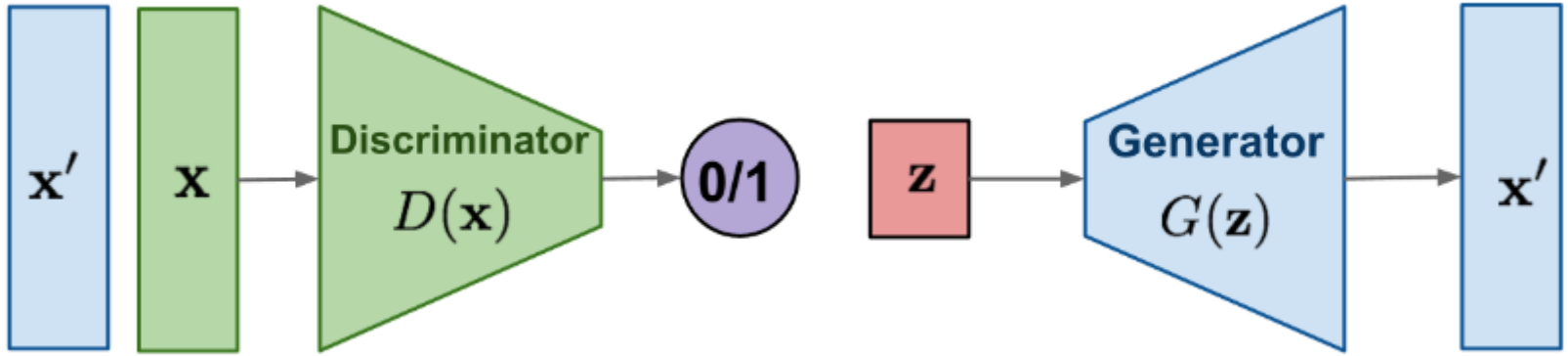
不同生成模型对前述目标**建模方式不同**

- **显式的方法:** 明确地定义并最大化关于 $p_\theta(\boldsymbol{x}_0)$ 的目标
  - 变分自编码器 (Variational Autoencoders, VAE)
  - 扩散模型 (Diffusion Models)
- **隐式的方法:** 使用模型对 $p_\theta(\boldsymbol{x}_0)$ 进行黑盒式的计算
  - 对抗生成网络 (Generative Adversarial Networks, GAN)

生成模型 — 对抗生成网络

对抗生成网络 (Generative Adversarial Networks, GAN)

- 使用一个 判别器 (Discriminator) 判定 生成器 (Generator) 的生成结果是否符合真实分布



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

生成模型 — 对抗生成网络

- 对抗生成网络 (Generative Adversarial Networks, GAN)

  生成器和判别器互相对抗, 动态平衡

  判别器的目标为最大化下式:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

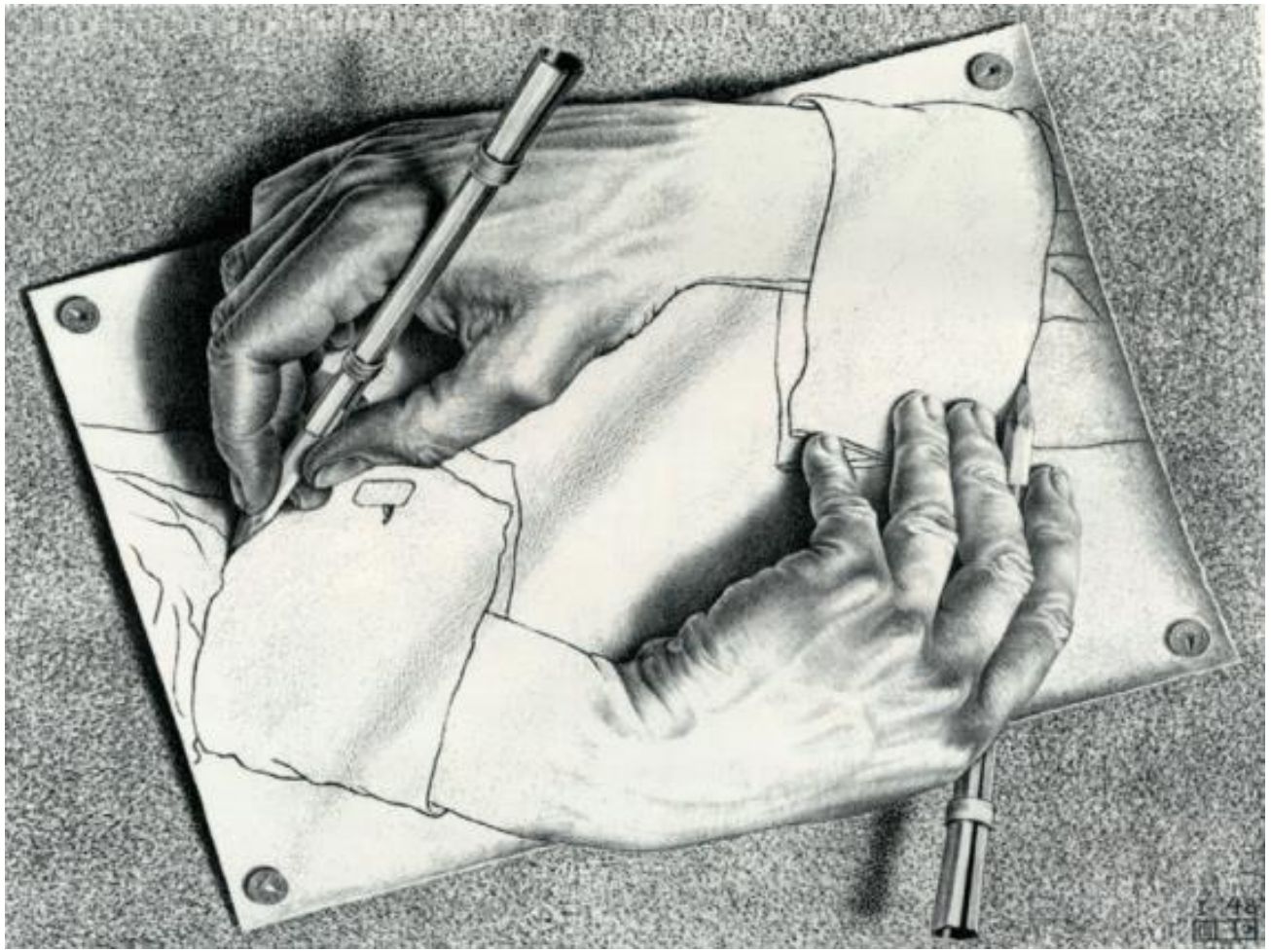
  生成器的目标为最小化下式:

$$\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

  优点: 强大、逻辑简单、易于构建

  缺点: 较难训练, 动态平衡较难把控

Reasons to Love GANs



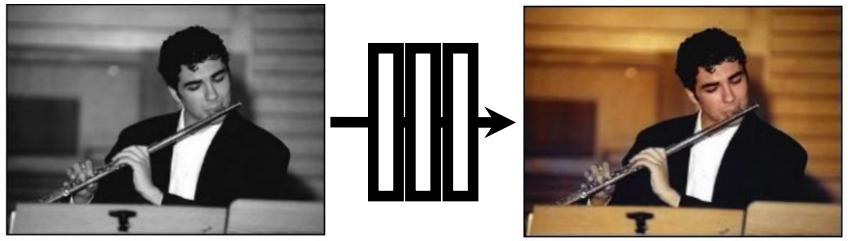
Alexei A. Efros UC Berkeley

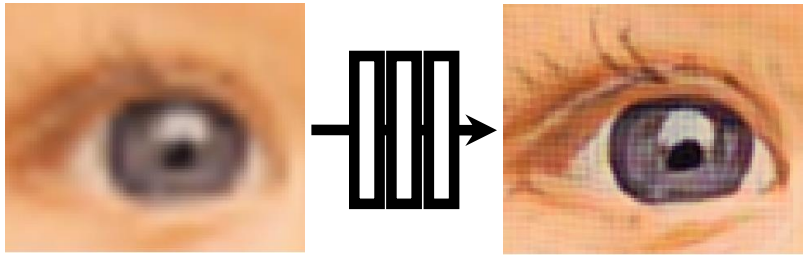
My 5 reasons to love GANs

1. GANs set up an arms race
2. GANs can be used as a “learned loss function”
3. GANs are “meta-supervisors”
4. GANs are great data memorizers
5. GANs are democratizing computer art

Generative Adversarial Network

Generated images



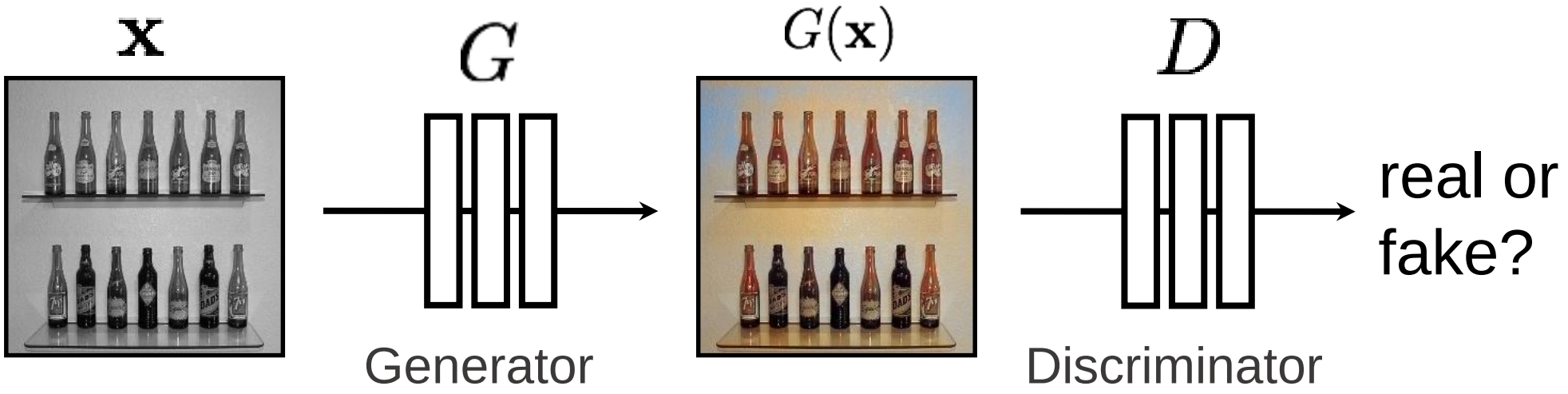


Generated vs. Real (classifier)

Real photos

…

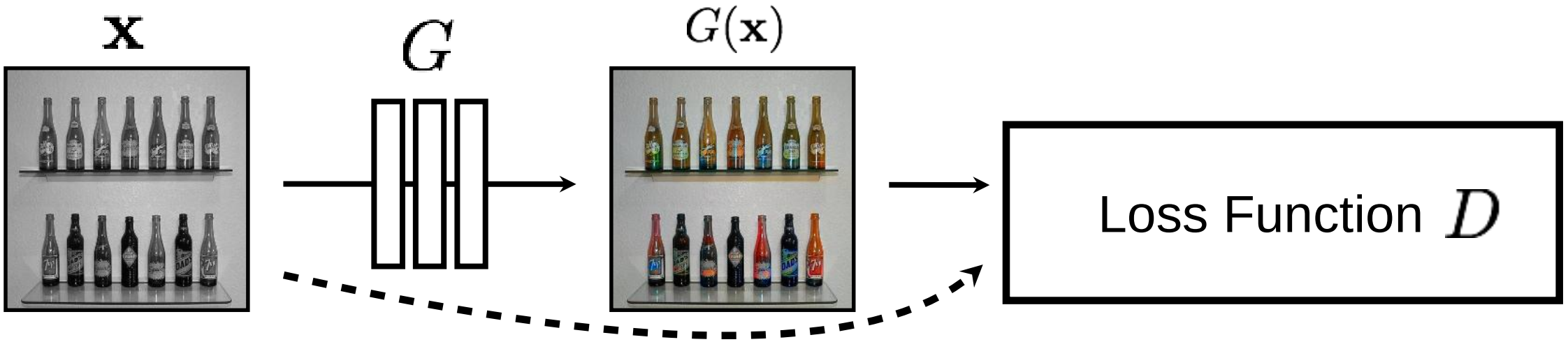
[Goodfellow et al. 2014]



G tries to synthesize fake images that fool D

D tries to identify the fakes

[Goodfellow et al. 2014]



G's perspective: **D** is a loss function.

Rather than being hand-designed, it is *learned*.

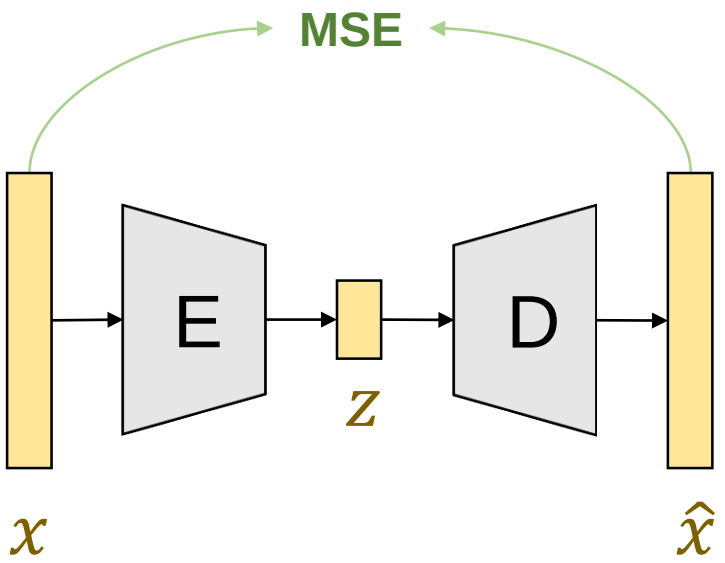
[Goodfellow et al. 2014]
[Isola et al. 2017]

生成模型 — 变分自编码器

- 自编码器 (Autoencoder)
  - 目的: 数据降维
  - 结构: 编码器 E + 解码器 D

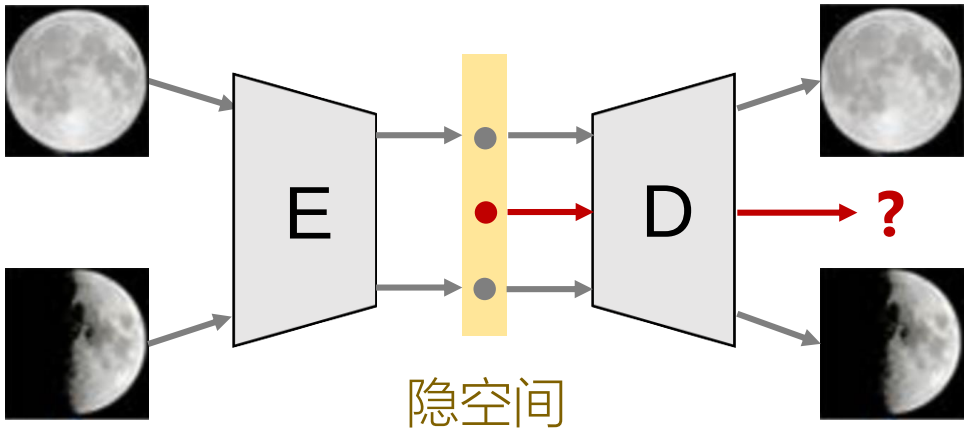
    具有窄的 bottleneck
  - 通过 $x \rightarrow \hat{x}$ 重建任务将 x 降维为 z

    训练损失函数可以使用 MSE

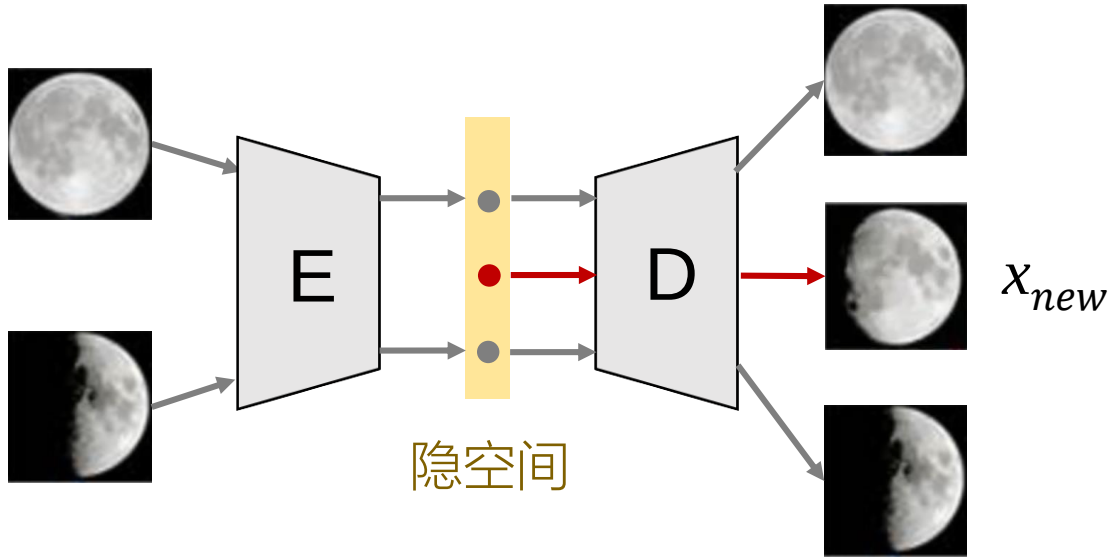


生成模型 — 变分自编码器

- 自编码器不是生成模型
  - 生成模型: $x_{new} \sim p(x)$



自编码器的效果



期望实现的效果

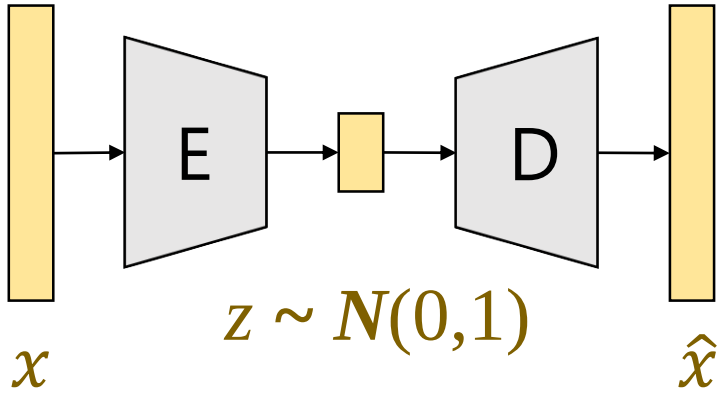
生成模型 — 变分自编码器

变分自编码器 (Variational Autoencoder, VAE)

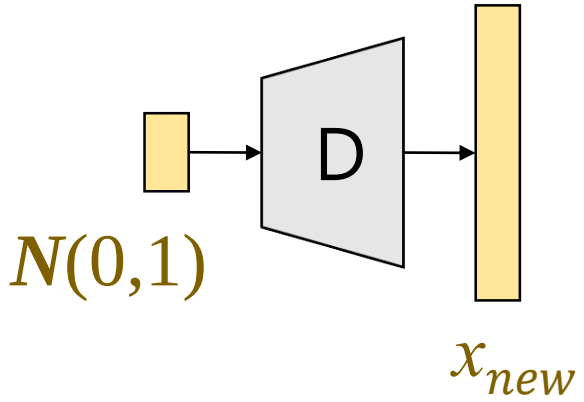
- z 能否变成一个先验分布?
  - 例如, 让 z 服从 $\boldsymbol{N}(0,1)$ 分布

    这样生成新的样本, 只需要从 $\boldsymbol{N}(0,1)$ 中采样, 然后经过解码器D解码

训练过程:



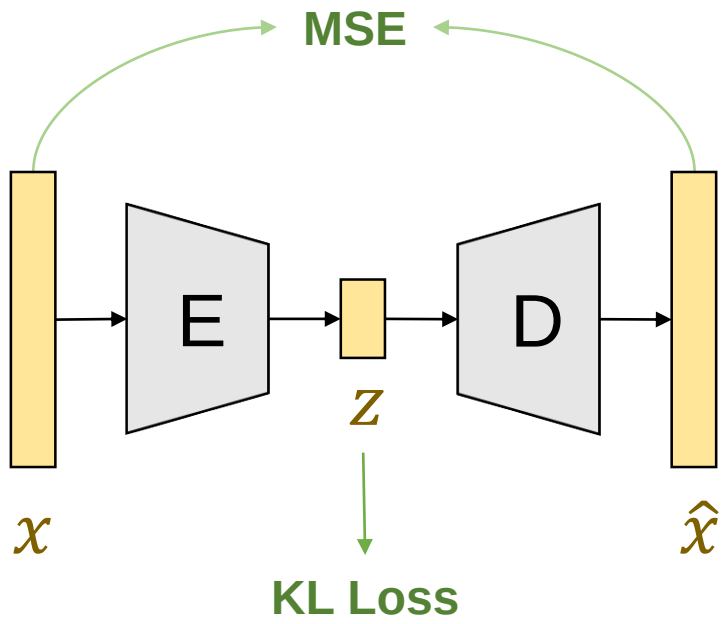
解码过程:



生成模型 — 变分自编码器

变分自编码器 (Variational Autoencoder, VAE)

- 目的: 生成模型
- 结构: 编码器 E + 解码器 D
- 损失函数: $L = L_{MSE} + L_{KL}$
  - MSE 学习如何重建
  - KL Loss 让 z 符合分布要求



Diederik P. Kingma, Max Welling: Auto-Encoding Variational Bayes. ICLR 2013

生成模型 — 变分自编码器

变分自编码器 (Variational Autoencoder, VAE)

- 学习目标: 通过最大化对数似然函数 $\mathbb{E}_{p^*(\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x})]$ 拟合数据分布

  根据隐变量模型，$p_{\boldsymbol{\theta}}(\mathbf{x})$ 的形式是 $p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$

- $p_{\boldsymbol{\theta}}(\mathbf{x})$ 的积分难以求解，因此根据变分分析的方法最大化其变分下界

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \underbrace{\mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]}_{\text{likelihood term}} - \underbrace{\mathrm{KL}[q_\eta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]}_{\text{KL penalty}}$$

对应重建损失函数　　对应 **KL Loss**

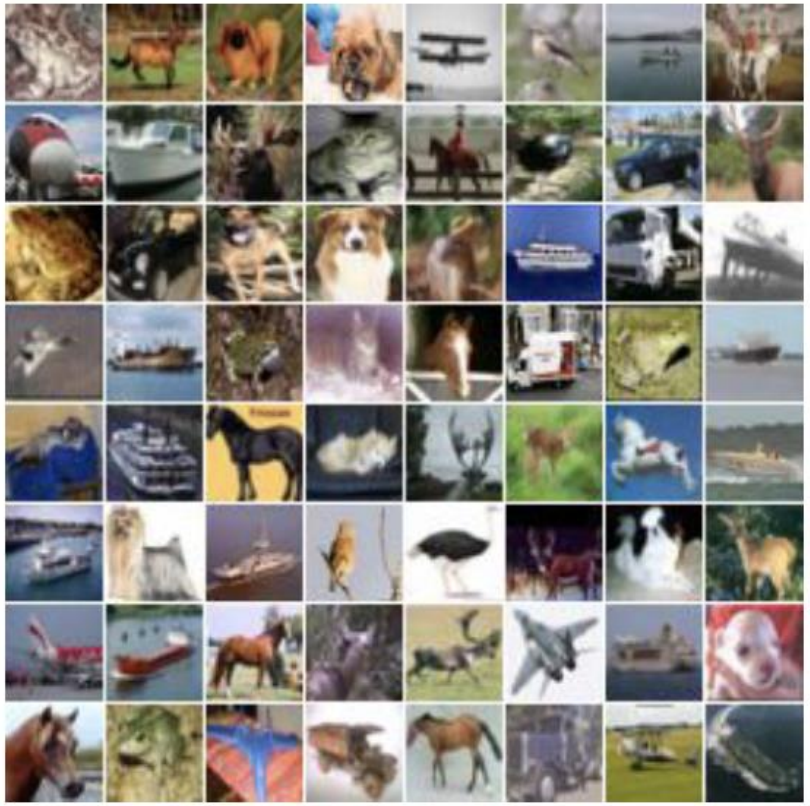
编码器分布：$q_\eta(\mathbf{z}|\mathbf{x})$

解码器分布：$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$

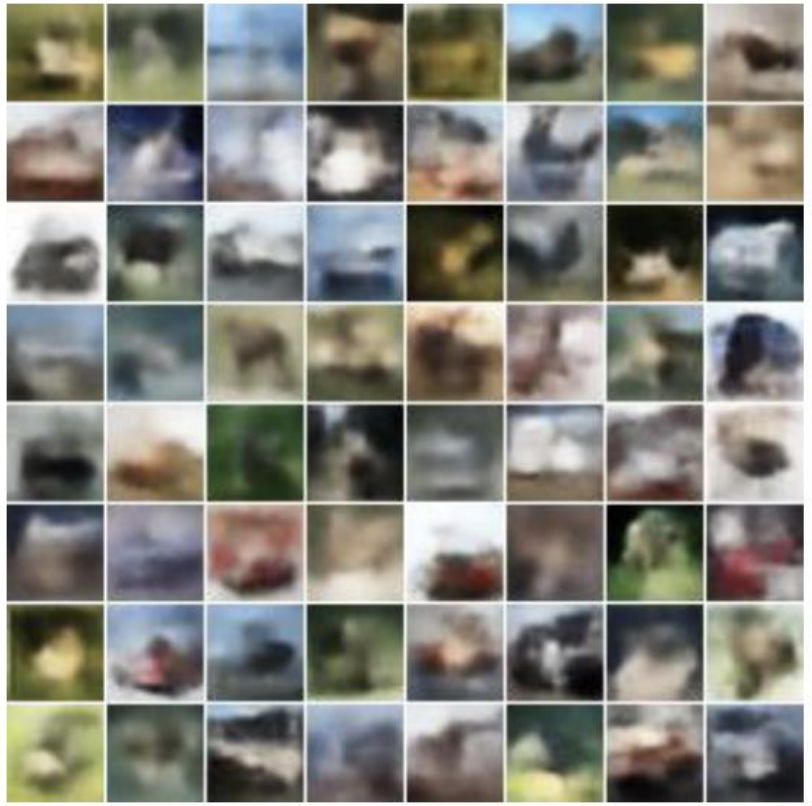
Mihaela Rosca, et al. Distribution Matching in Variational Inference. arXiv, 2018

生成模型 — 变分自编码器

- 生成效果



数 据



VAE

Mihaela Rosca, et al. Distribution Matching in Variational Inference. arXiv, 2018

生成模型 — 变分自编码器

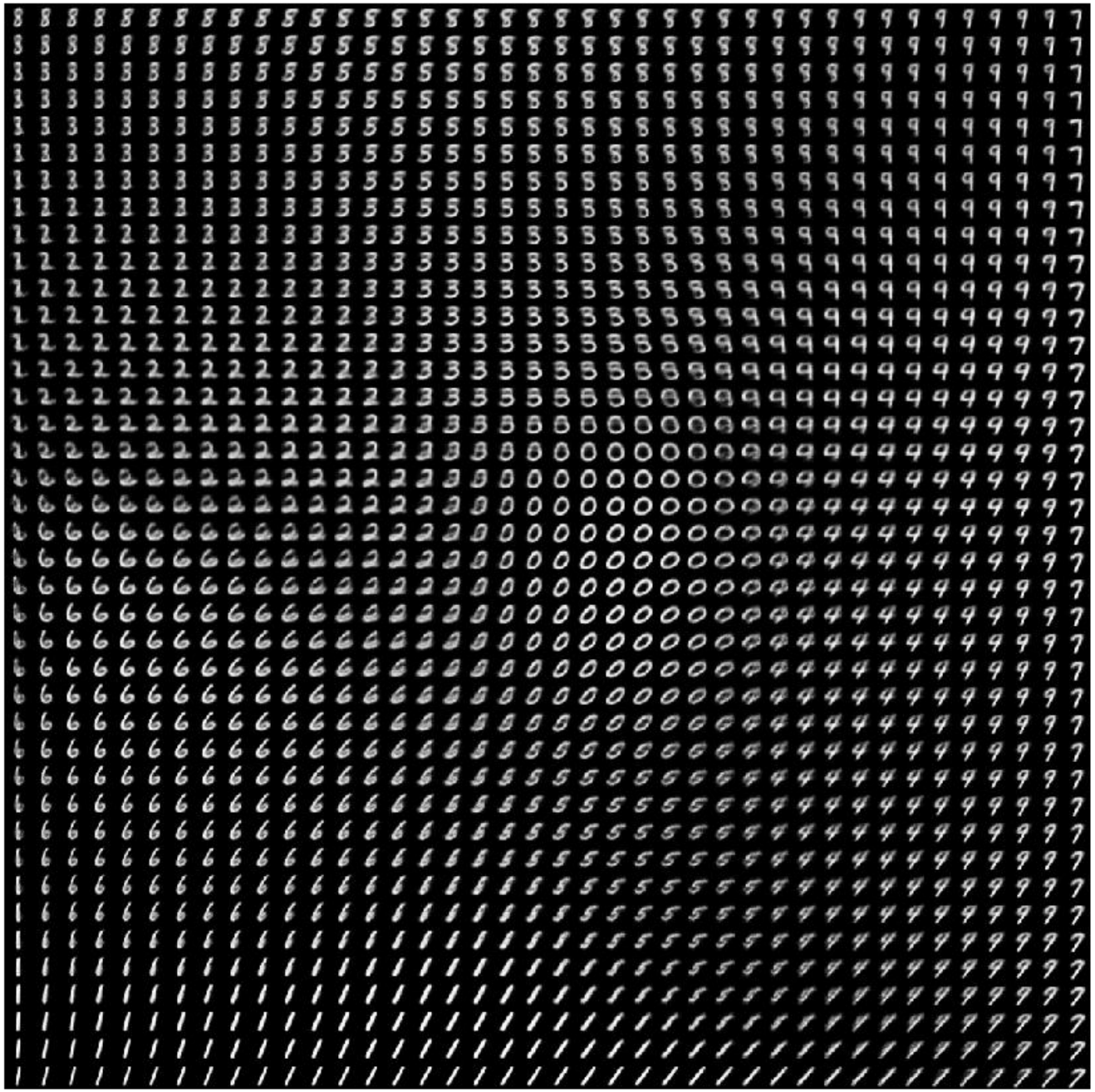
- 不同 z 的维度下的效果

(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Diederik P. Kingma, Max Welling: Auto-Encoding Variational Bayes. ICLR 2013

生成模型 — VAE

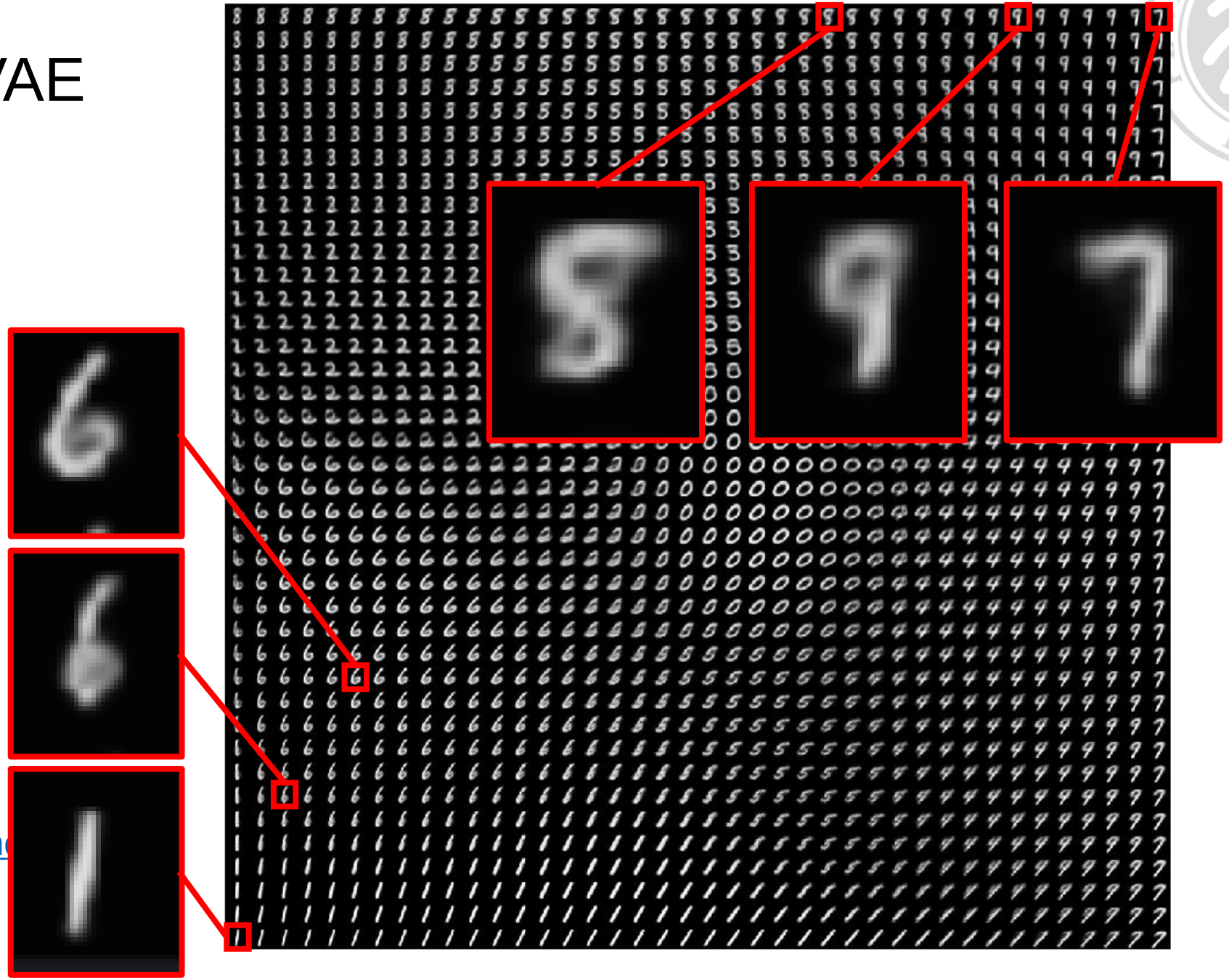
- 在隐空间采样



https://www.jeremyjordan.me/variational-autoencoders/

生成模型 — VAE

- 在隐空间采样

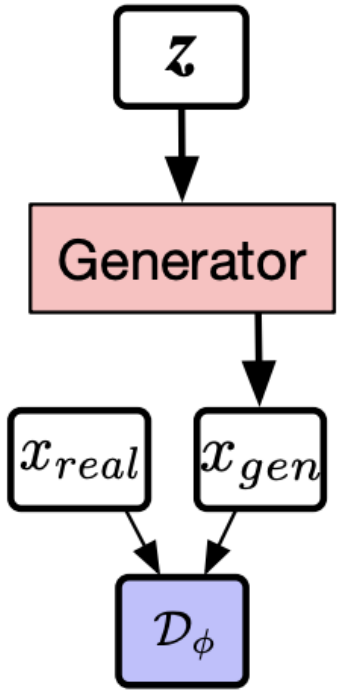


https://www.jeremyjordan.m
iational-autoencoders/

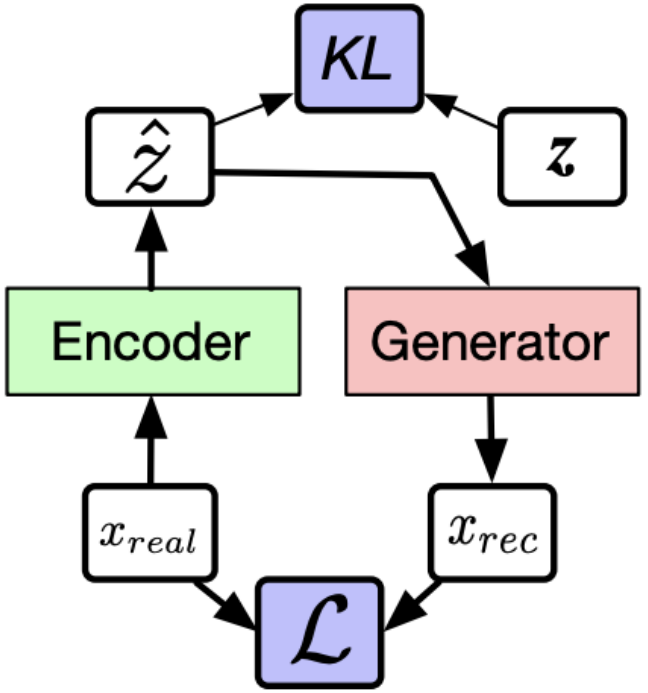
生成模型的区别

变分自编码器 (Variational Autoencoder, VAE)

- 与 GAN 的区别



GAN



VAE

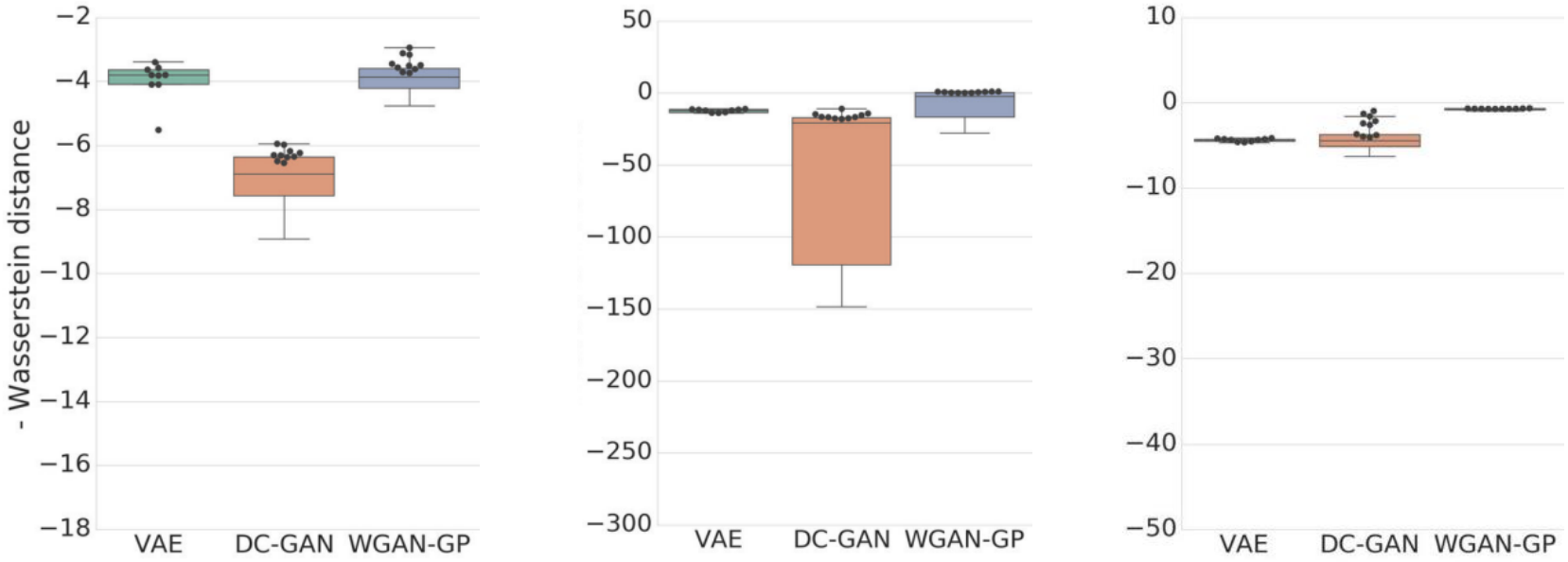
Mihaela Rosca, et al. Distribution Matching in Variational Inference. arXiv, 2018

生成模型的区别

变分自编码器 (Variational Autoencoder, VAE)

- 与 GAN 的生成性能比较
  - Wasserstein distance 越高越好

Mihaela Rosca, et al. Distribution Matching in Variational Inference. arXiv, 2018



ColorMNIST　　CelebA　　CIFAR-10

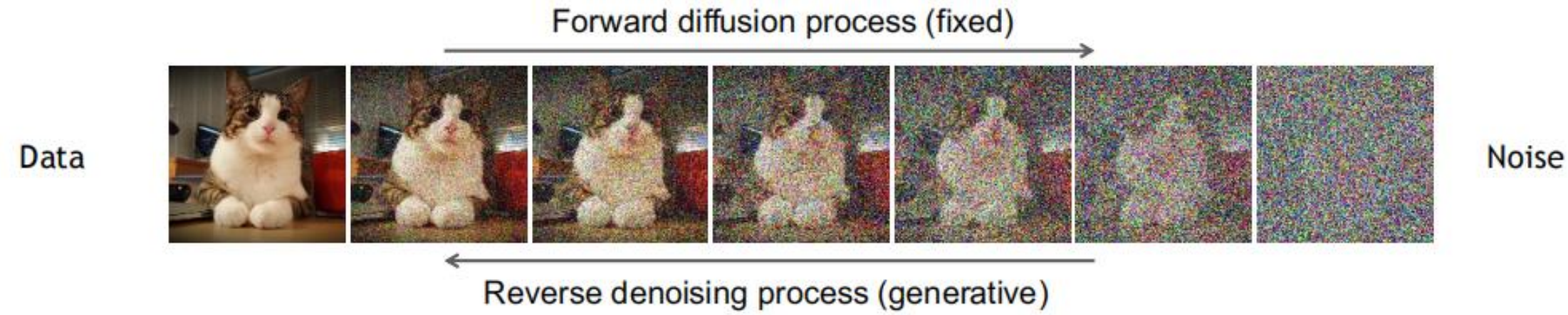
生成模型 — 扩散模型

扩散模型 (Diffusion Model)

- **扩散现象:** 一滴墨水逐渐冲淡在一杯清水中
- **启发:** 墨水 → 图像　　清水 → 高斯噪声

  能否建模该过程?

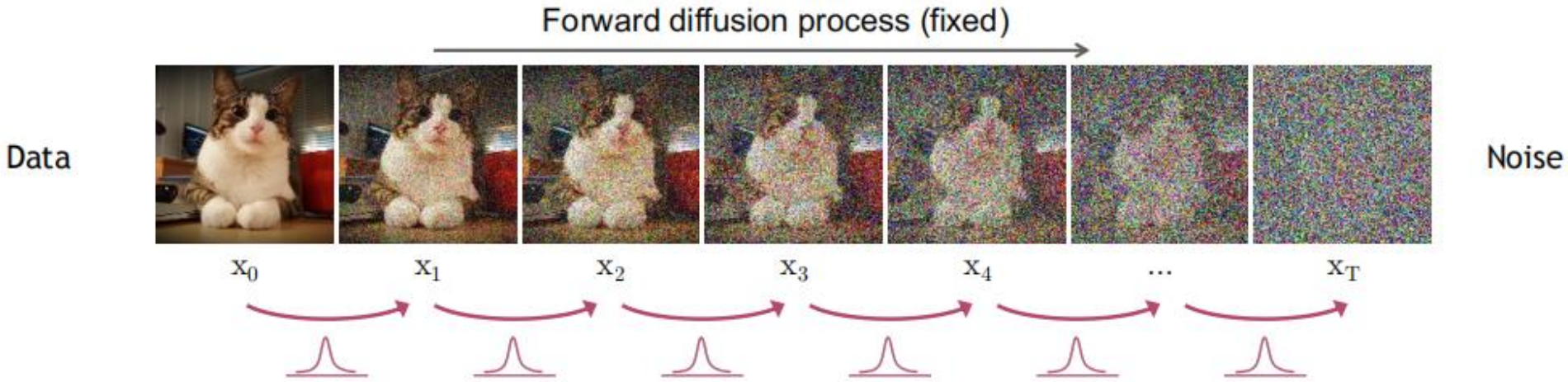
  能否逆转该过程?



生成模型 — 扩散模型

扩散模型 (Diffusion Model)

- 前向过程：墨水 “冲淡” 的过程, 从图像到高斯



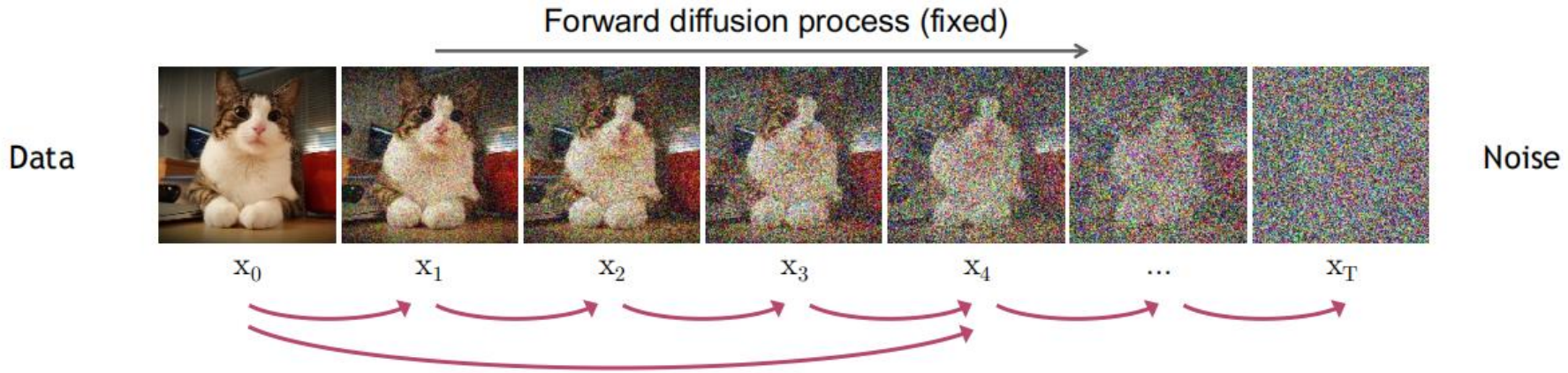
- 后一状态由前一状态加上一个小的高斯噪声的采样得到：

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I})$$

Jonathan Ho, et al. Denoising Diffusion Probabilistic Models. NIPS 2020

生成模型 — 扩散模型

扩散模型 (Diffusion Model)



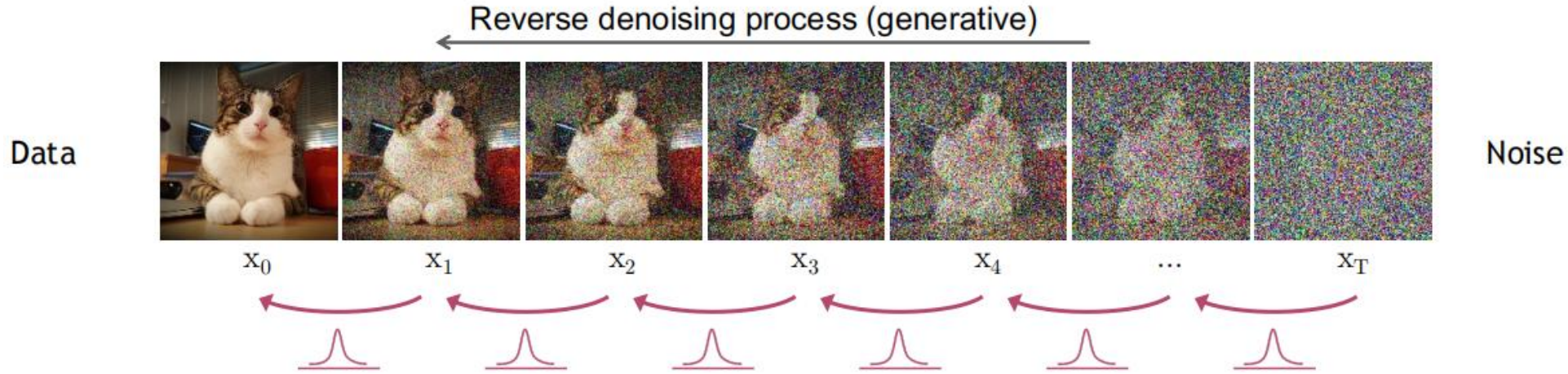
- 独立的高斯噪声可叠加：

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}))$$

生成模型 — 扩散模型

扩散模型 (Diffusion Model)

- 反向过程: “恢复墨水” 的过程, 从高斯到图像



- 前一状态由当前状态 “减去” 某一小的高斯噪声的采样得到

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

生成模型 — 扩散模型

扩散模型 (Diffusion Model)

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

- 怎么获得减去的高斯噪声?
- 使用神经网络进行拟合
  - 采用一个网络, 接收当前的噪声图像 $\mathbf{x}_t$ 和步数 t
  - 输出前一状态的噪声图像满足的高斯分布的均值 $\mu_\theta(\mathbf{x}_t, t)$

生成模型 — 扩散模型

扩散模型 (Diffusion Model)

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

- 如何搭建这一神经网络?
- 直觉上: 只需完成基本的 “去噪 (denoise)” 任务即可, 以均方差 MSE 为损失函数
  - 直接预测前一状态不够好, 前一状态与当前状态差别微小, 相对误差容易过大
  - 从 $\mathbf{x}_t$ 直接预测 $\mathbf{x}_0$ → 再从预测的 $\mathbf{x}_0$ 重建 $\mathbf{x}_{t-1}$

生成模型 — 扩散模型

- 基于 $\mathbf{x}_t$ 和 $\mathbf{x}_0$ 推导 $\mathbf{x}_{t-1}$ 满足的分布: 扩散模型的参数化

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

- 由贝叶斯原理, 得:

$$\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\
&\propto \exp\Big(-\frac{1}{2}\big(\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)\mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0)\big)\Big)
\end{aligned}$$

生成模型 — 扩散模型

- 推得 $\mathbf{x}_{t-1}$ 满足的高斯分布的均值和方差

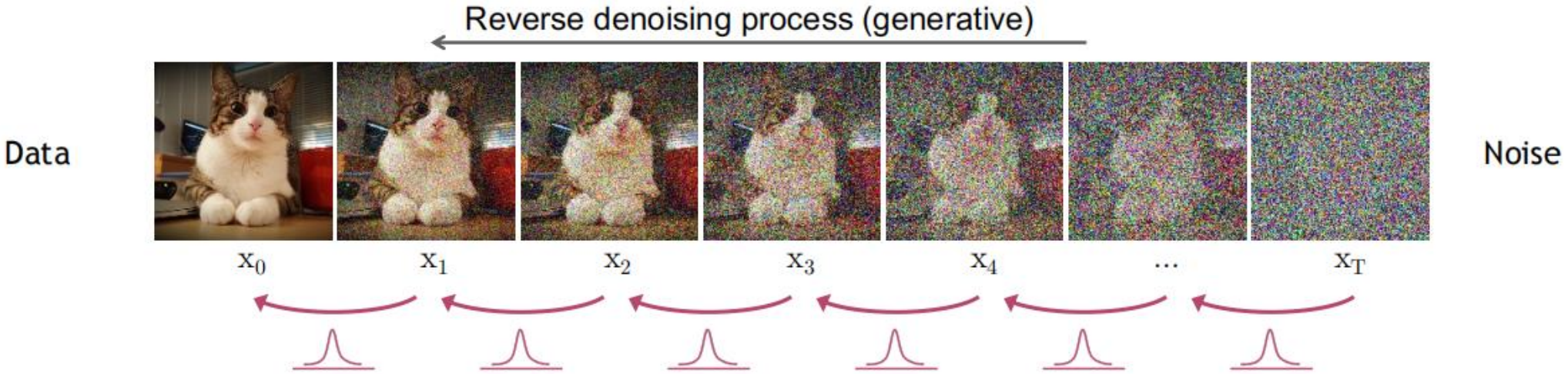
$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

$$\tilde{\beta}_t = 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0$$

- 考虑到 $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)$
- 神经网络以 $\mathbf{x}_t$ 和 t 为输入，输出 $\mathbf{x}_t$ 与 $\mathbf{x}_0$ 之间的高斯噪声 $\hat{\mathbf{z}}_t$
- 训练时随机采样 $\mathbf{z}_t$，从 $\mathbf{x}_0$ 构建 $\mathbf{x}_t$，将神经网络输出的 $\hat{\mathbf{z}}_t$ 和真实 $\mathbf{z}_t$ 之间的MSE作为损失函数

生成模型 — 扩散模型



- 采样步骤
  - 从高斯分布中随机采出起点 $\mathbf{x}_T$
  - 使用训练好的网络预测噪声 $\hat{\mathbf{z}}_T$
  - 计算前一状态 $\hat{\mathbf{x}}_{T-1}$ 满足的高斯分布的均值 $\mu_\theta(\mathbf{x}_T, T)$ 并采样 $\hat{\mathbf{x}}_{T-1}$
  - 重复上述步骤, 直至下标为 0

生成模型 — 扩散模型

为什么选择扩散模型

- 极大的模型容量和生成多样性
- 对训练用的图片种类几乎没有任何限制

  可以直接在大型数据集 ImageNet上构建

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. NIPS 2021

生成模型 — 扩散模型

为什么选择扩散模型

- 极高的生成质量

  在各类生成任务中达成最高水平

Model	FID	sFID	Prec	Rec
LSUN Bedrooms 256×256				
DCTransformer† [42]	6.40	6.66	0.44	**0.56**
DDPM [25]	4.89	9.07	0.60	0.45
IDDPM [43]	4.24	8.21	0.62	0.46
StyleGAN [27]	2.35	6.62	0.59	0.48
ADM (dropout)	**1.90**	**5.59**	**0.66**	0.51
LSUN Horses 256×256				
StyleGAN2 [28]	3.84	6.46	0.63	0.48
ADM	2.95	**5.94**	0.69	**0.55**
ADM (dropout)	**2.57**	6.81	**0.71**	**0.55**
LSUN Cats 256×256				
DDPM [25]	17.1	12.4	0.53	0.48
StyleGAN2 [28]	7.25	**6.33**	0.58	0.43
ADM (dropout)	**5.57**	6.69	**0.63**	**0.52**
ImageNet 64×64				
BigGAN-deep* [5]	4.06	3.96	**0.79**	0.48
IDDPM [43]	2.92	**3.79**	0.74	0.62
ADM	2.61	**3.77**	0.73	0.63
ADM (dropout)	**2.07**	4.29	0.74	**0.63**

Model	FID	sFID	Prec	Rec
ImageNet 128×128				
BigGAN-deep [5]	6.02	7.18	**0.86**	0.35
LOGAN† [68]	3.36			
ADM	5.91	**5.09**	0.70	**0.65**
ADM-G (25 steps)	5.98	7.04	0.78	0.51
ADM-G	**2.97**	**5.09**	0.78	0.59
ImageNet 256×256				
DCTransformer† [42]	36.51	8.24	0.36	**0.67**
VQ-VAE-2†‡ [51]	31.11	17.38	0.36	0.57
IDDPM‡ [43]	12.26	5.42	0.70	0.62
SR3†‡ [53]	11.30			
BigGAN-deep [5]	6.95	7.36	**0.87**	0.28
ADM	10.94	6.02	0.69	0.63
ADM-G (25 steps)	5.44	5.32	0.81	0.49
ADM-G	**4.59**	**5.25**	0.82	0.52
ImageNet 512×512				
BigGAN-deep [5]	8.43	8.13	**0.88**	0.29
ADM	23.24	10.19	0.73	**0.60**
ADM-G (25 steps)	8.41	9.67	0.83	0.47
ADM-G	**7.72**	**6.57**	0.87	0.42

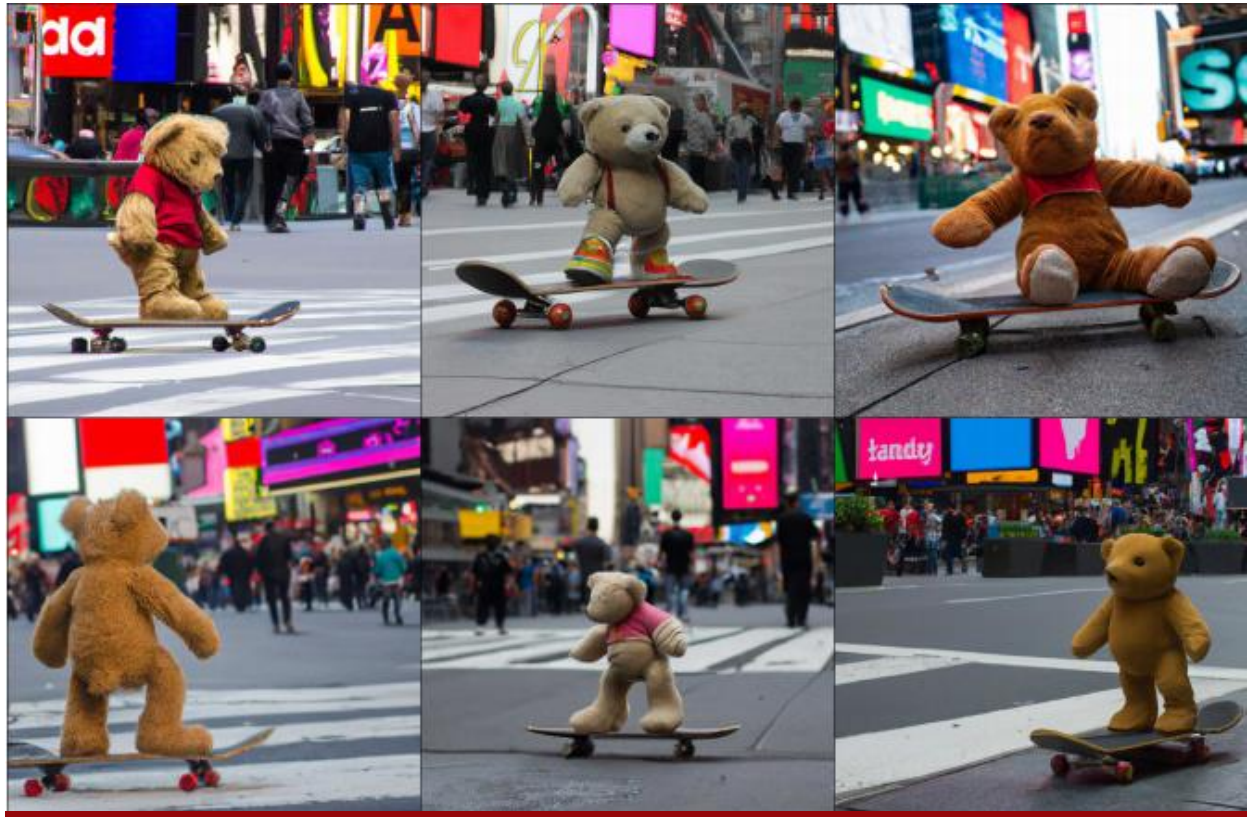
Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. NIPS 2021

生成模型 — 扩散模型

为什么选择扩散模型

- 对各种任务的适应性

  广泛适用于各种图像的条件生成任务



“A teddybear on a skateboard in Times Square.”

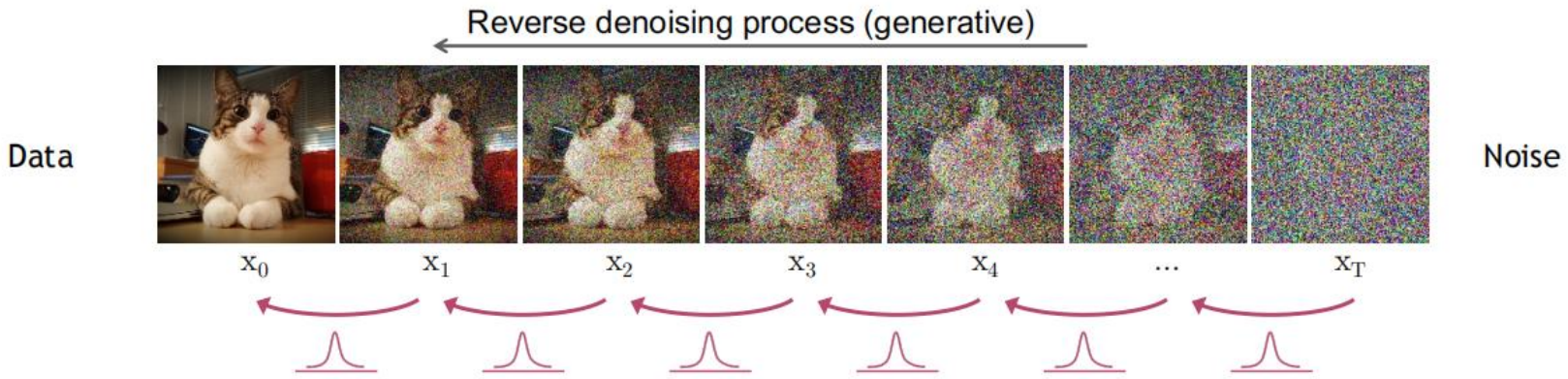
文本引导的图像生成结果
by DALL-E 2

Ramesh A, et al. Hierarchical text-conditional image generation with clip latents. arXiv 2022.

生成模型 — 扩散模型

扩散模型 (Diffusion Model)

- 缺点1: 巨大的训练代价
  - 逐步 “去噪” 的本质是构建了一系列不同的去噪网络, 每个网络都需要训练
  - 去噪的步数往往上千, 意味着需要训练上千个去噪网络
  - 常用的扩散模型训练代价经常超过 1000 个 GPU * Day



生成模型 — 扩散模型

扩散模型 (Diffusion Model)

- 缺点2: 缓慢的采样速度
  - 每一步 “去噪” 都需要将网络进行前向传播, 因此原生的扩散模型的采样速度比GAN慢上千倍
  - 机器学习领域的学者们已经提出了多种加速扩散模型的理论方法, 目前最快的方法仅需约十数次前向传播即可得到极高的质量

Cheng Lu, et al. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. arXiv 2022.

生成模型的表现

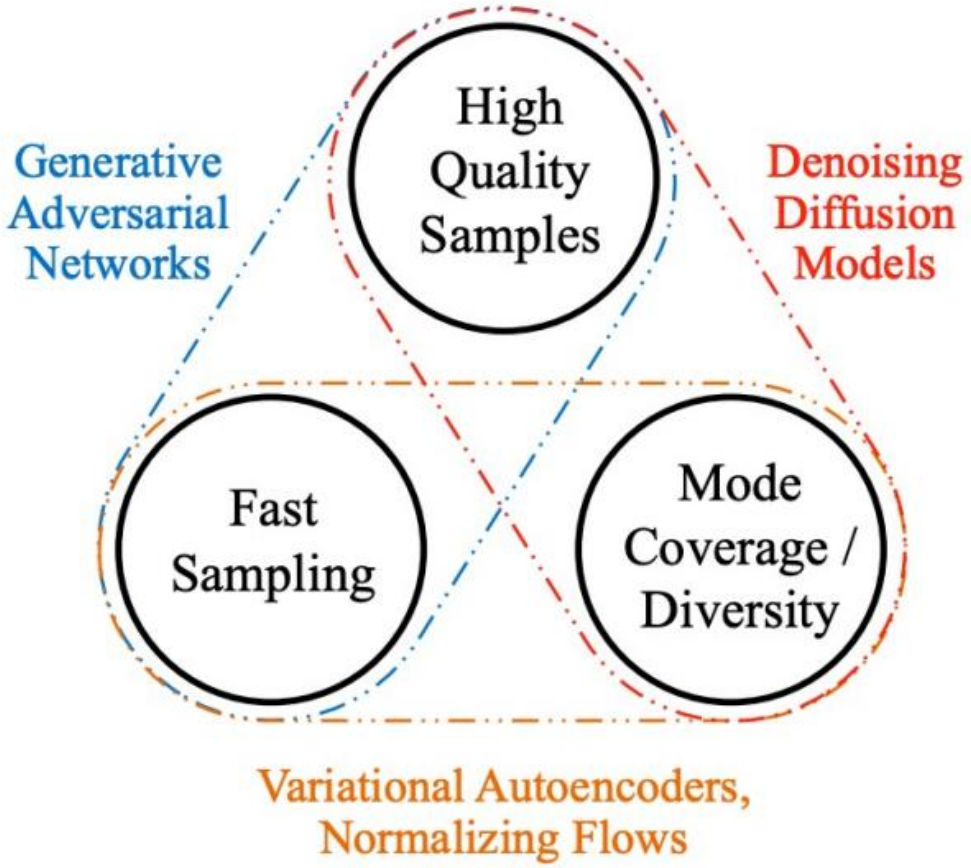
生成模型的三个难题

- 采样的速度
- 采样的质量
- 采样的多样性

上述目标互相制约, 难以同时达成

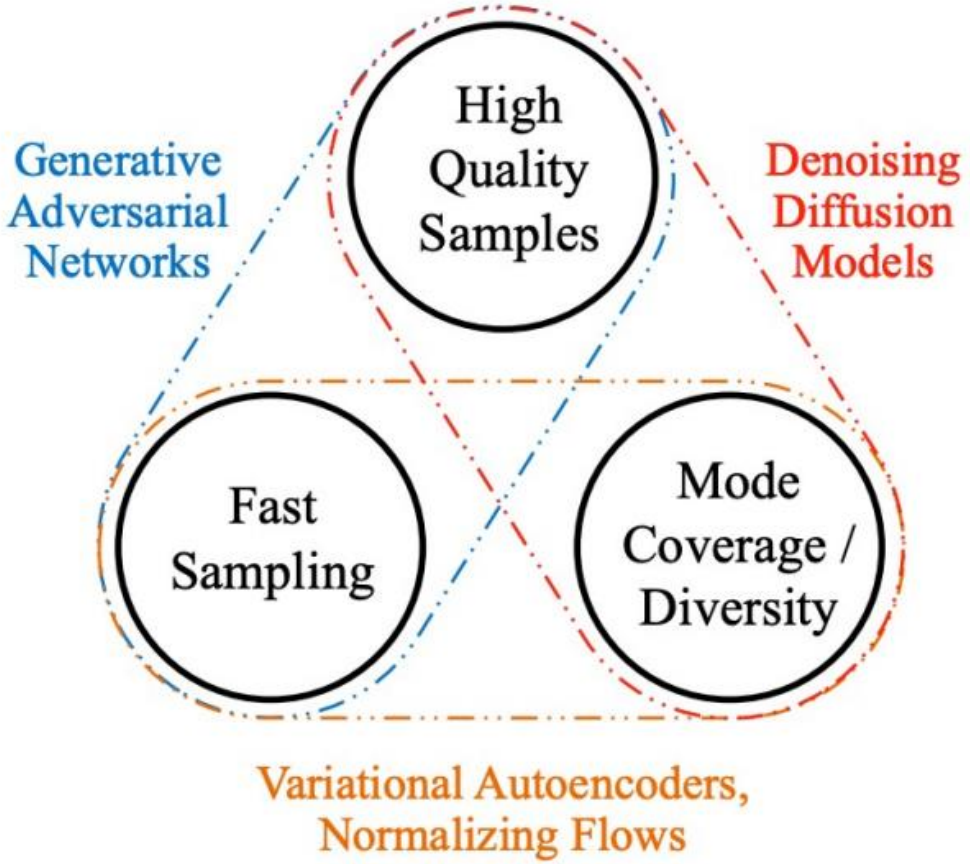
生成模型的表现

- GAN, VAE和Diffusion Models 三种常用模型各自的特性：



Karsten Kreis, et al. Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022.

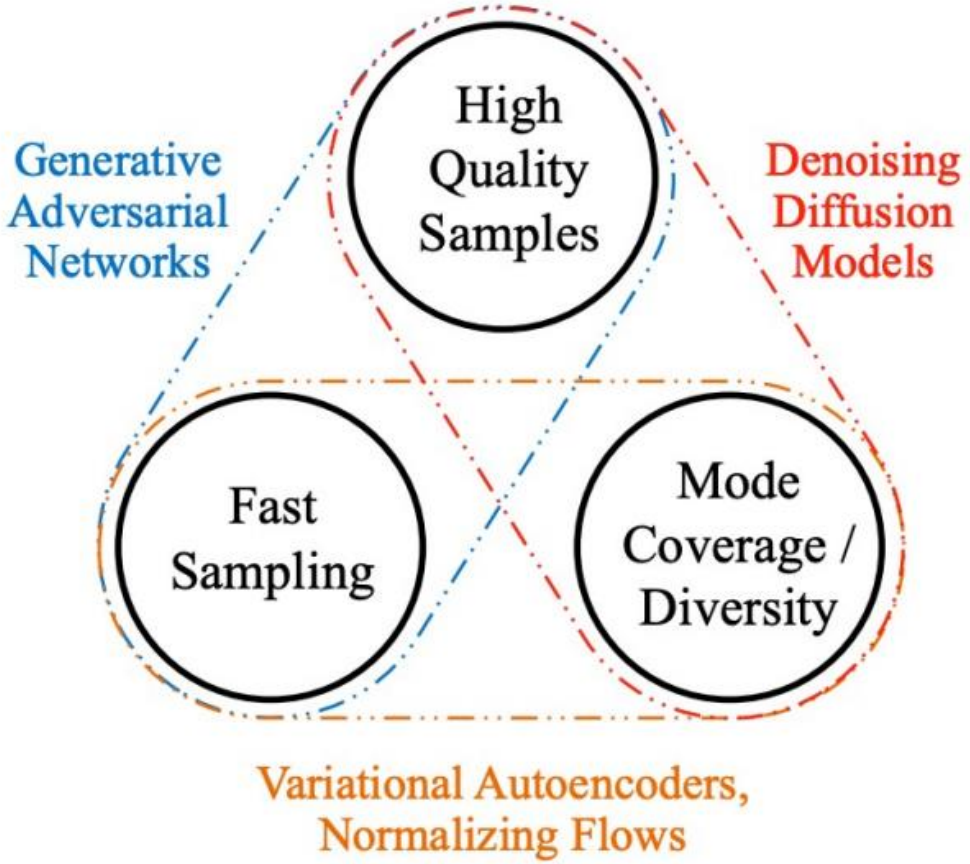
生成模型的表现



GAN

- 生成质量高, 易于采样的特性突出
- 往往需要指定域的图像进行训练, 否则很容易训练失败

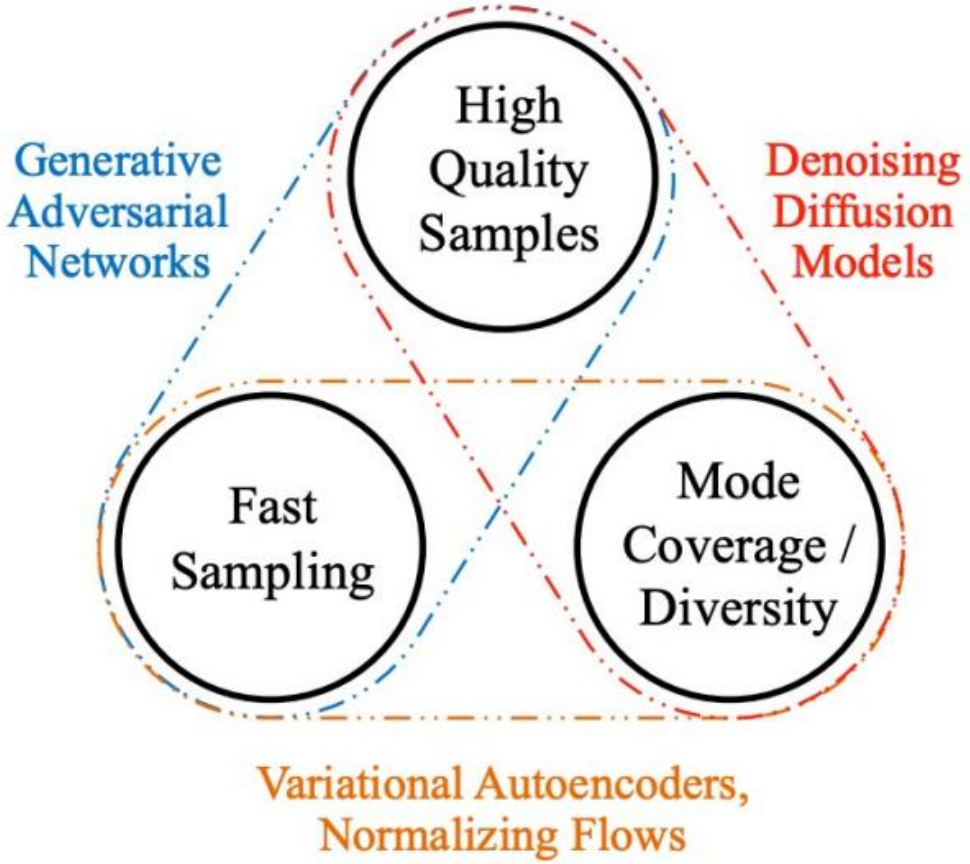
生成模型的表现



VAE

- 易于训练的模型，采样速度也非常快
- 理论基础使得它可以生成多样化的数据
- 它的采样质量往往十分普通

生成模型的表现



Diffusion Model

- 理论基础为其提供了极高的生成质量和极大的生成多样性
- 采样非常耗时，对设备要求很高

PART 02

2D生成技术与呈现

二维生成的任务

- 高质量图像生成
- 图到图转换
- 文本-图像跨模态生成

高质量图像生成

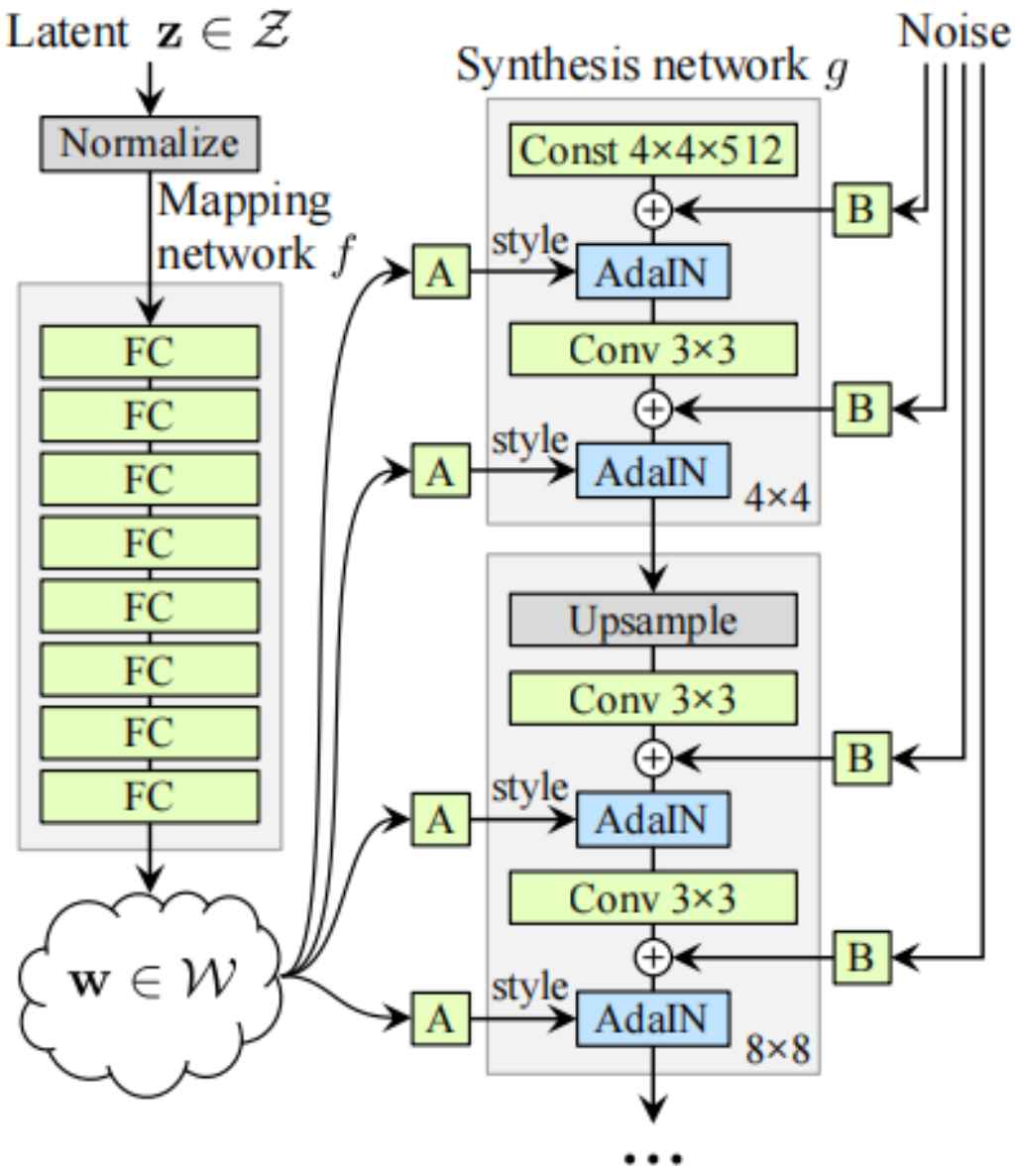
StyleGAN

- 生成高分辨率的图像
- 可实现对图像不同层级的控制
- 引入随机性, 生成图像多样

高质量图像生成

StyleGAN 网络结构概览

- Mapping network — 映射层
- Synthesis network — 合成层

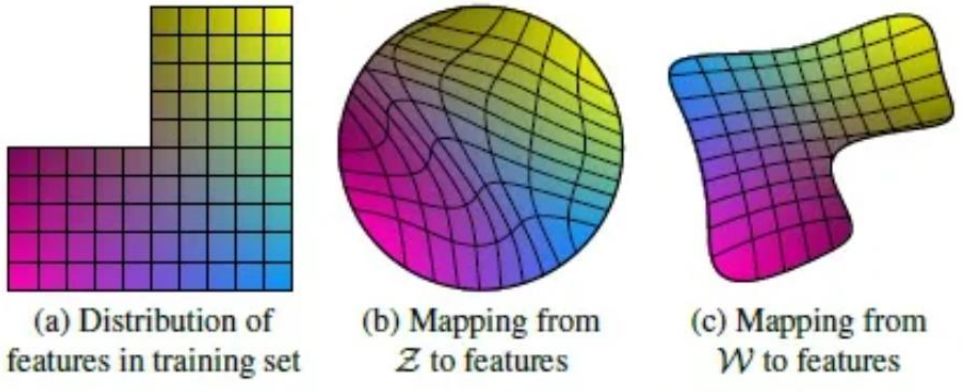


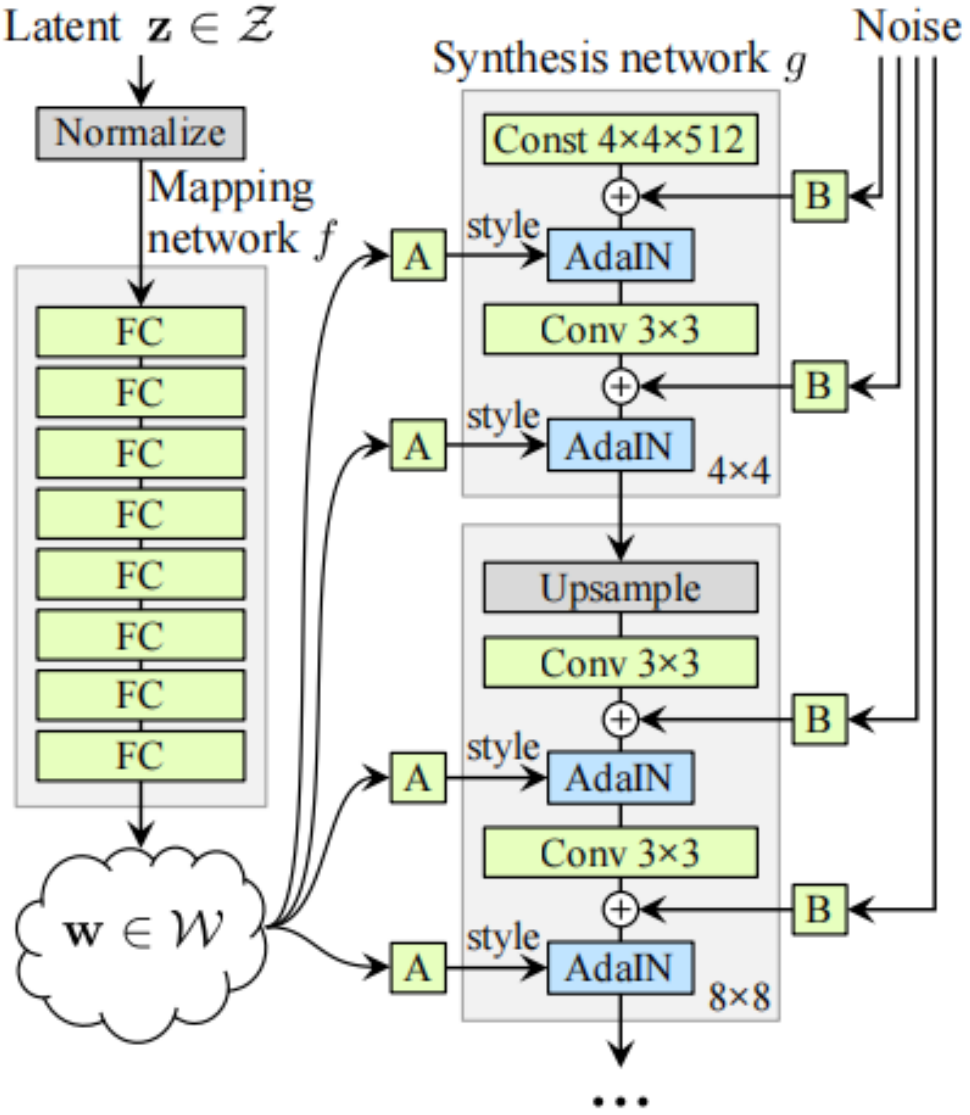
Karras T, et al. A Style-Based Generator Architecture for Generative Adversarial Networks. CVPR 2019.

高质量图像生成

StyleGAN

- Mapping Network
  - 将图像编码 $\mathbf{z}$ 映射为新编码 $\mathbf{w}$
  - 将均匀的高斯分布映射到真实分布





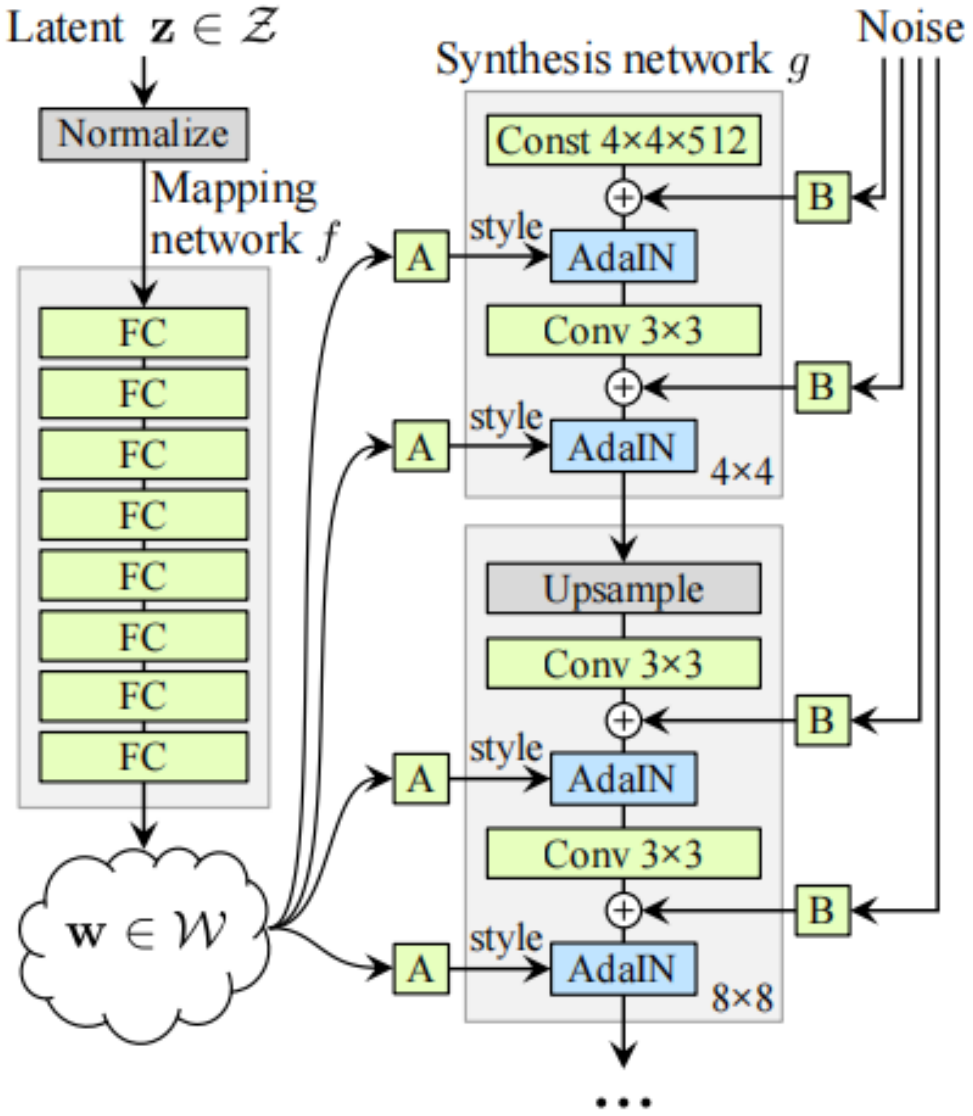
高质量图像生成

StyleGAN

- Synthesis network
  - 根据编码 $\mathbf{w}$ 和噪声生成图像
  - 使用 AdaIN 进行图像调整

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

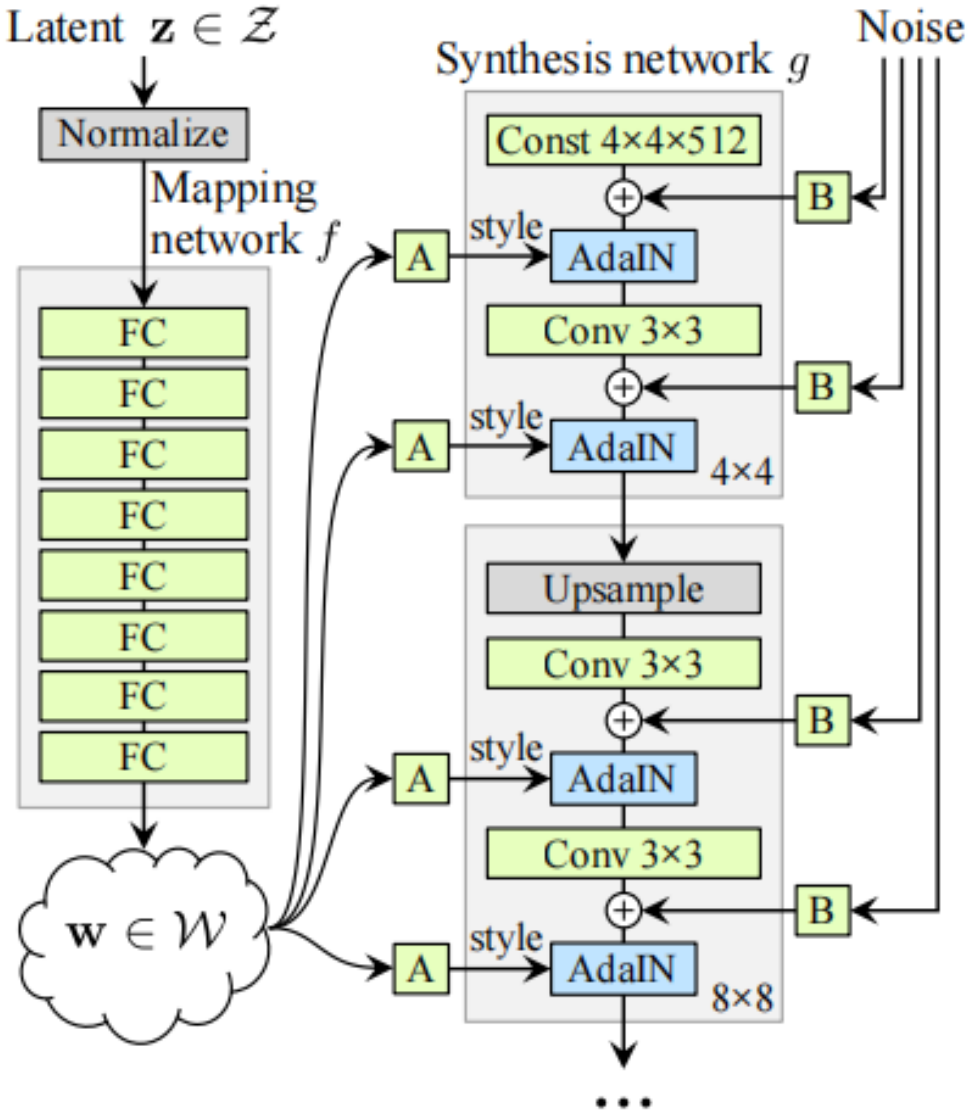
  - 多层迭代生成, 保证图像分辨率
  - 使用噪声增强生成多样性



高质量图像生成

StyleGAN 应用 —— 风格混合

- 采用两个不同的**w**控制生成
- 在某些层使用编码 B
- 在其余层使用编码 A
- 使用编码 B 的位置影响图像的特征



高质量图像生成

StyleGAN

- 基础层 (4×4 – 8×8)
- 人脸的身份特征

  （年龄、脸型等）

高质量图像生成

StyleGAN

- 中间层 (16×16 – 32×32)
- 人脸的发型、姿态

高质量图像生成

StyleGAN

- 细调层 (64×64 – 1024×1024)
- 人脸的肤色

高质量图像生成

StyleGAN

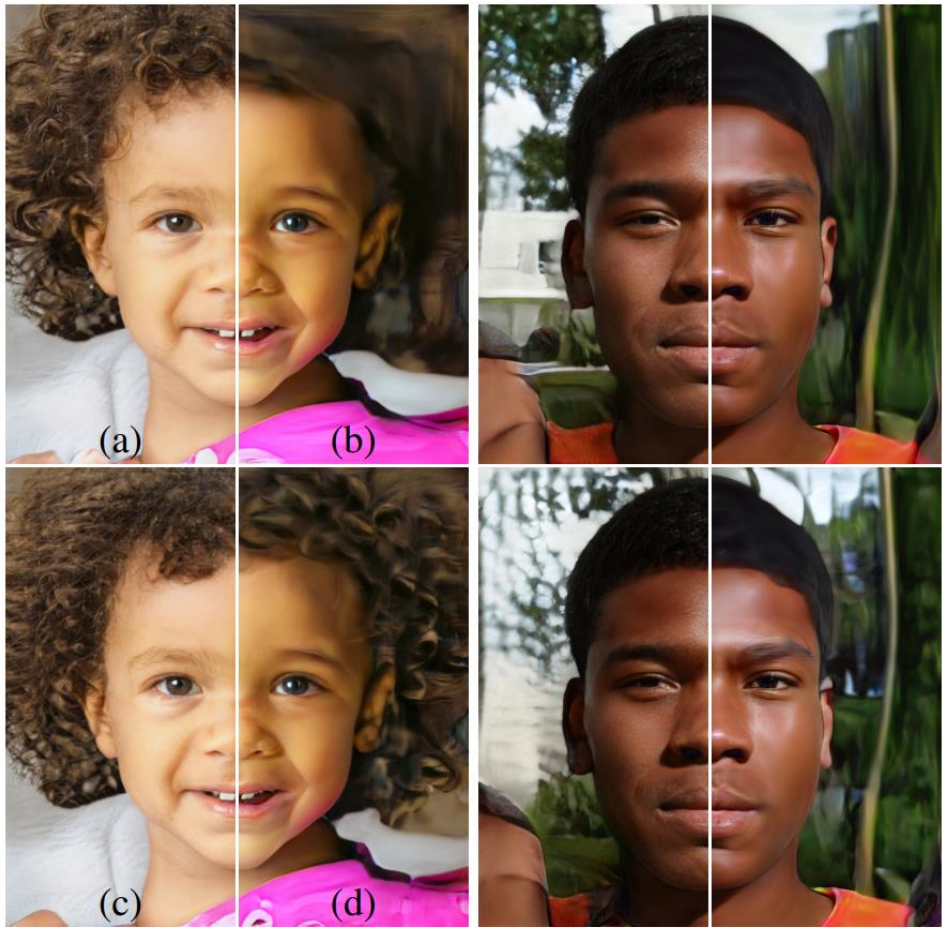
- 分析: 随机性的效果
- 随机性可以使生成的人脸的细节部分更随机、更自然

(a) Generated image (b) Stochastic variation (c) Standard deviation

高质量图像生成

StyleGAN

- 在不同层添加随机性的效果比较



（a）所有层噪声

（b）不加噪声

（c）细调层噪声

（d）基础层噪声

高质量图像生成

StyleGAN 的缺陷 —— “水滴形伪影”

- 原因: AdaIN操作带来的正则化伪影。
- 解决方法: 不使用AdaIN

Karras T, et al. Analyzing and Improving the Image Quality of StyleGAN, CVPR 2020.

高质量图像生成

StyleGAN2

- 目标: 解决StyleGAN存在的一系列问题, 最主要的就是伪影

Karras T, et al. Analyzing and Improving the Image Quality of StyleGAN. CVPR 2020.

高质量图像生成

StyleGAN2

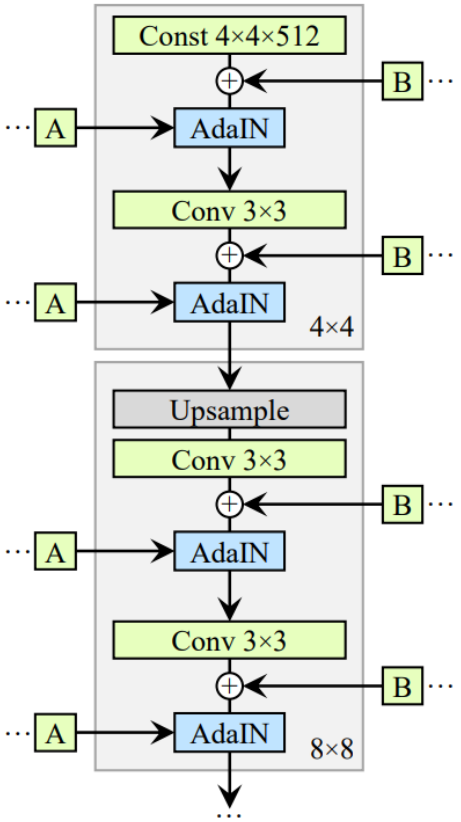
- 修改StyleGAN1结构
- 回顾 AdaIN

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

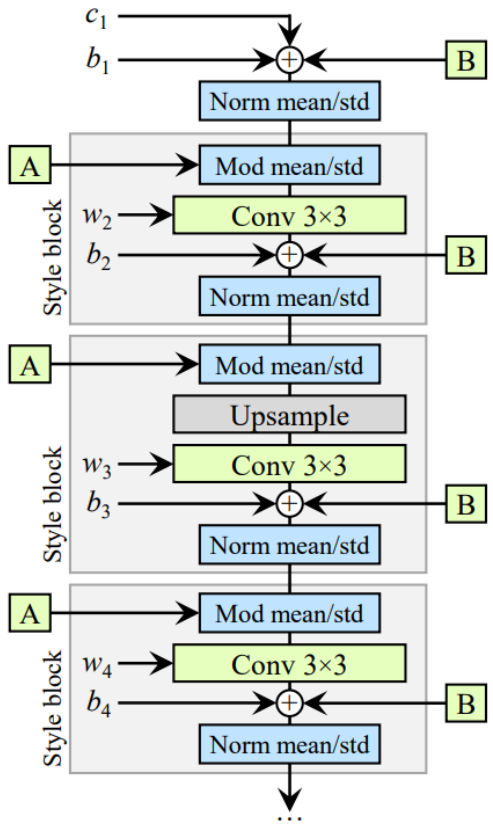
- 细化展示StyleGAN1:

  标注出卷积层的参数

  卷积核参数$\mathbf{w}$，偏移量 $\mathbf{b}$



StyleGAN1

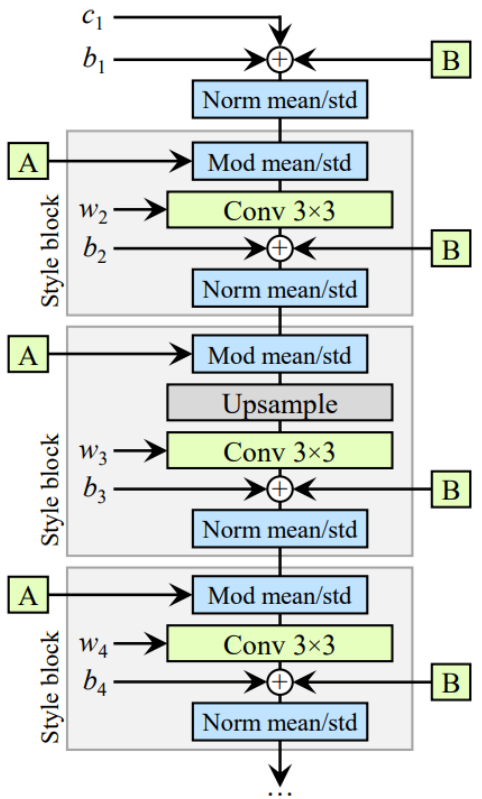


StyleGAN1 (细化)

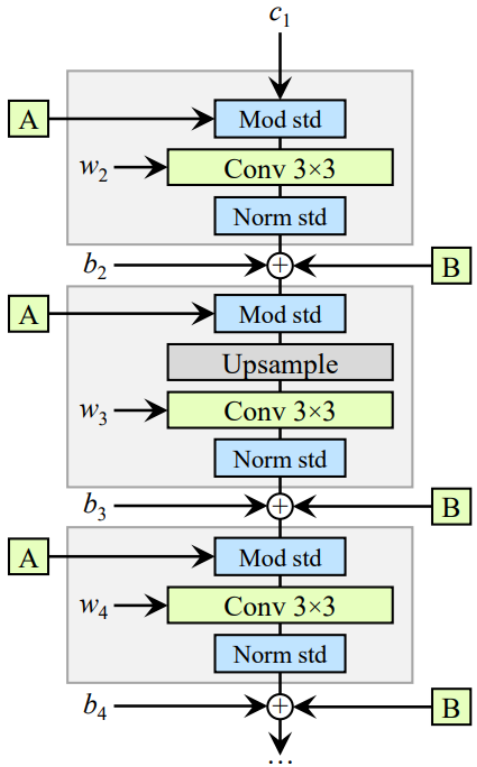
高质量图像生成

StyleGAN2 的具体修改

- 删除了模型对平均值的操作
- 将加入噪声移动至标准化模块后
- 开始时不添加噪声



StyleGAN1 (细化)



基于StyleGAN1
修改AdaIN后

高质量图像生成

StyleGAN2 —— 修改卷积参数

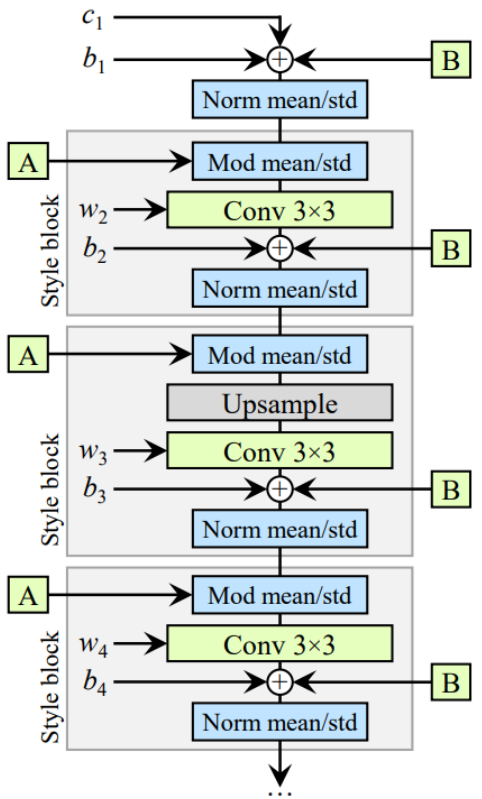
- AdaIN结构的等价性

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

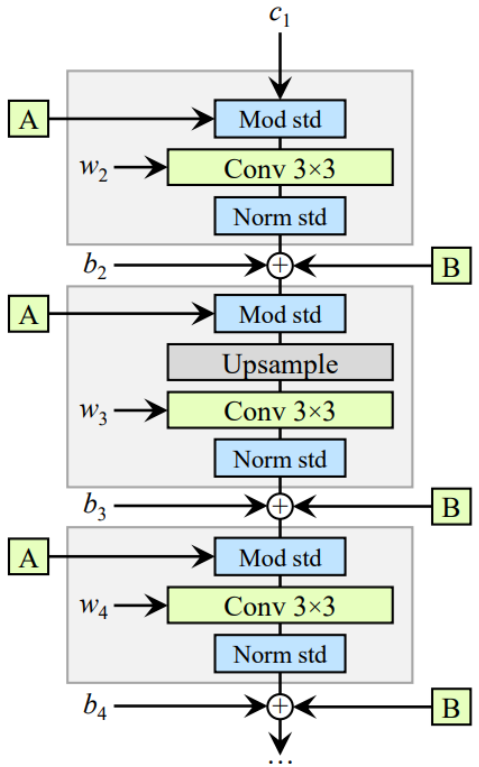
$$w'_{ijk} = s_i \cdot w_{ijk},$$

- 调适过程等价于对卷积核加参数

  也就是对$\mathbf{w}$的修改



StyleGAN1 (细化)



基于StyleGAN1
修改AdaIN后

高质量图像生成

StyleGAN2 —— 修改卷积参数

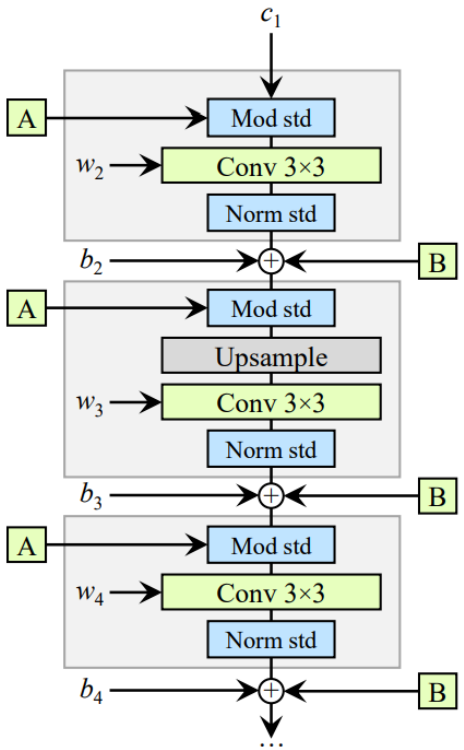
- 卷积核的方差

$$\sigma_j = \sqrt{\sum_{i,k} {w'_{ijk}}^2},$$

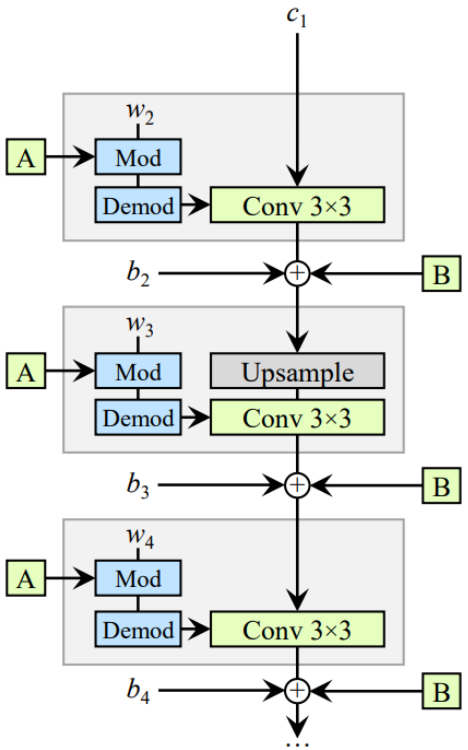
- 将标准化转移到卷积层

$$w'_{ijk} = s_i \cdot w_{ijk},$$

$$w''_{ijk} = w'_{ijk} \Big/ \sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon},$$



基于StyleGAN1
修改AdaIN后

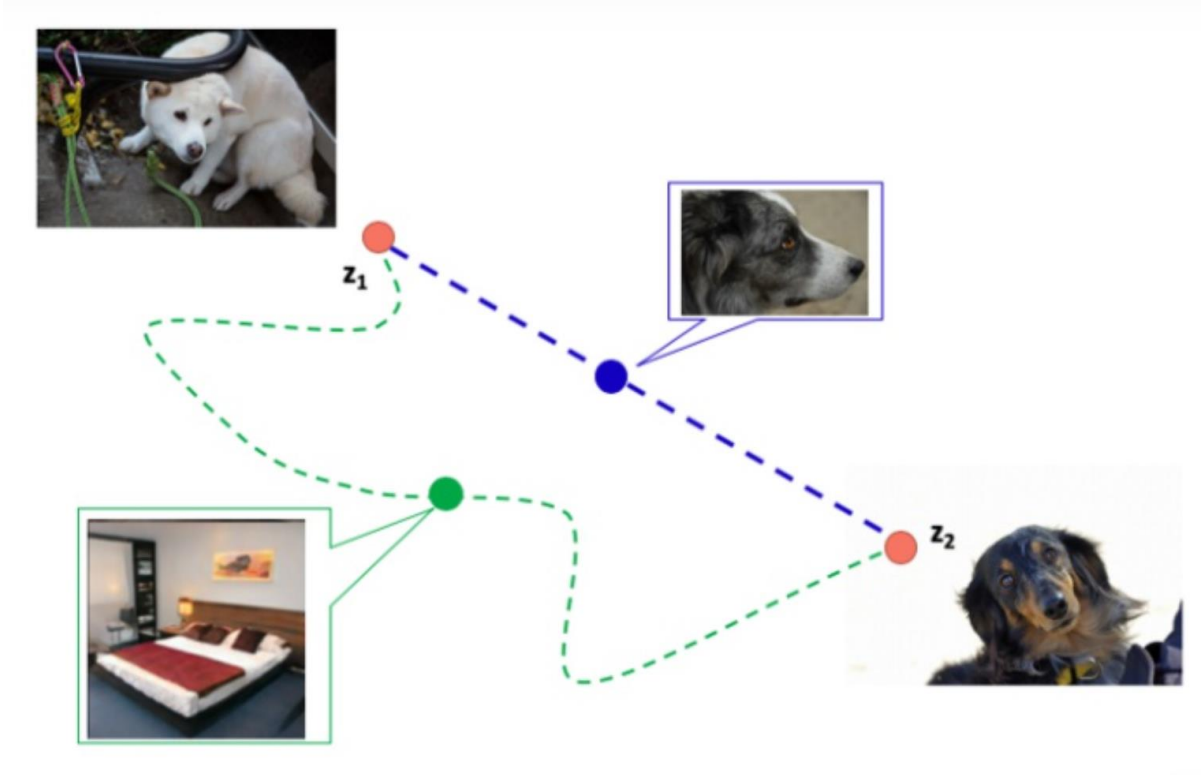


StyleGAN2

高质量图像生成

StyleGAN2 —— 最短路径损失

- 希望网络可以学到生成目标的最短路径



高质量图像生成

StyleGAN2 —— 最短路径损失

- 希望网络可以学到生成目标的最短路径

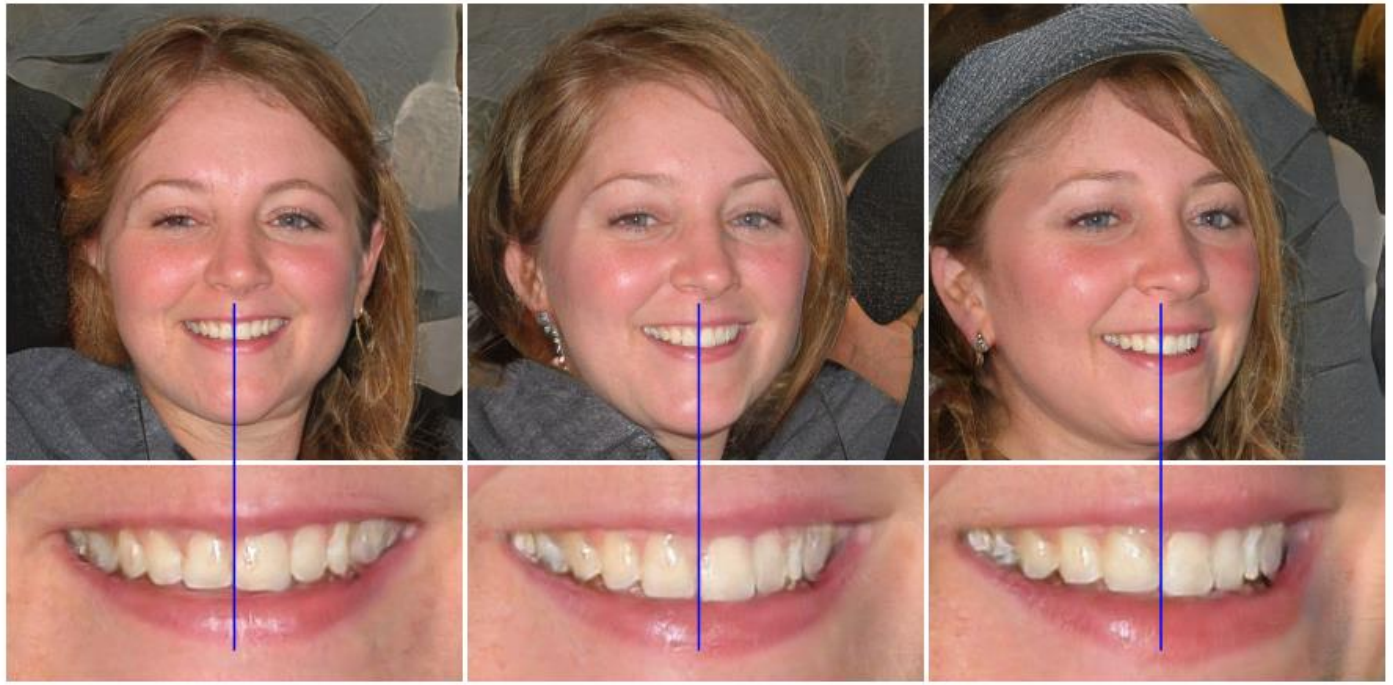
$$\mathbb{E}_{\mathbf{w},\mathbf{y}\sim\mathcal{N}(0,\mathbf{I})}\left(\left\|\mathbf{J}_{\mathbf{w}}^{T}\mathbf{y}\right\|_{2}-a\right)^{2}$$

- 前者是 $\mathbf{g}(\mathbf{w})\times\mathbf{y}$ 的导数值
- 后者是动态学习的最优平均值

高质量图像生成

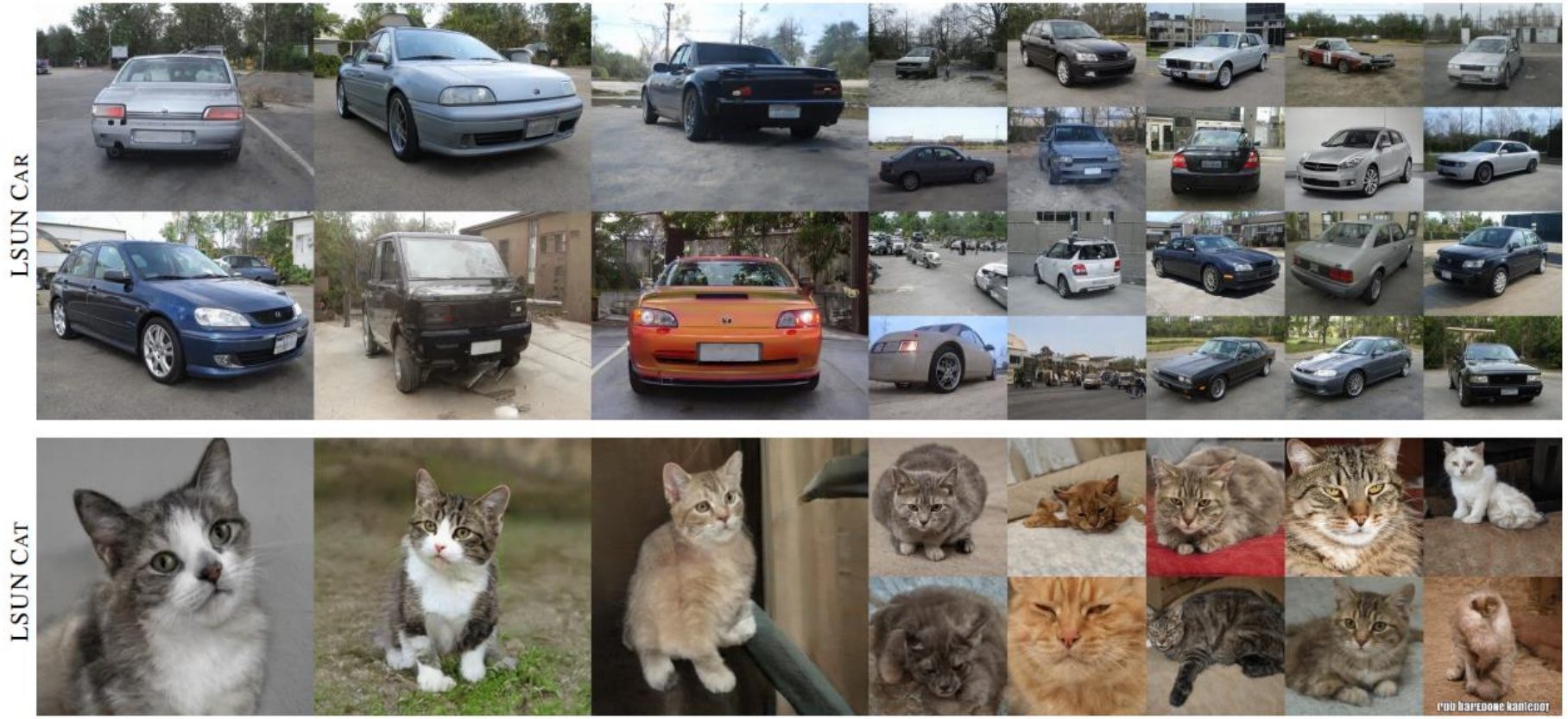
StyleGAN2 —— 其它优化

- 减少正则化损失参与
- 移除逐步学习过程



高质量图像生成

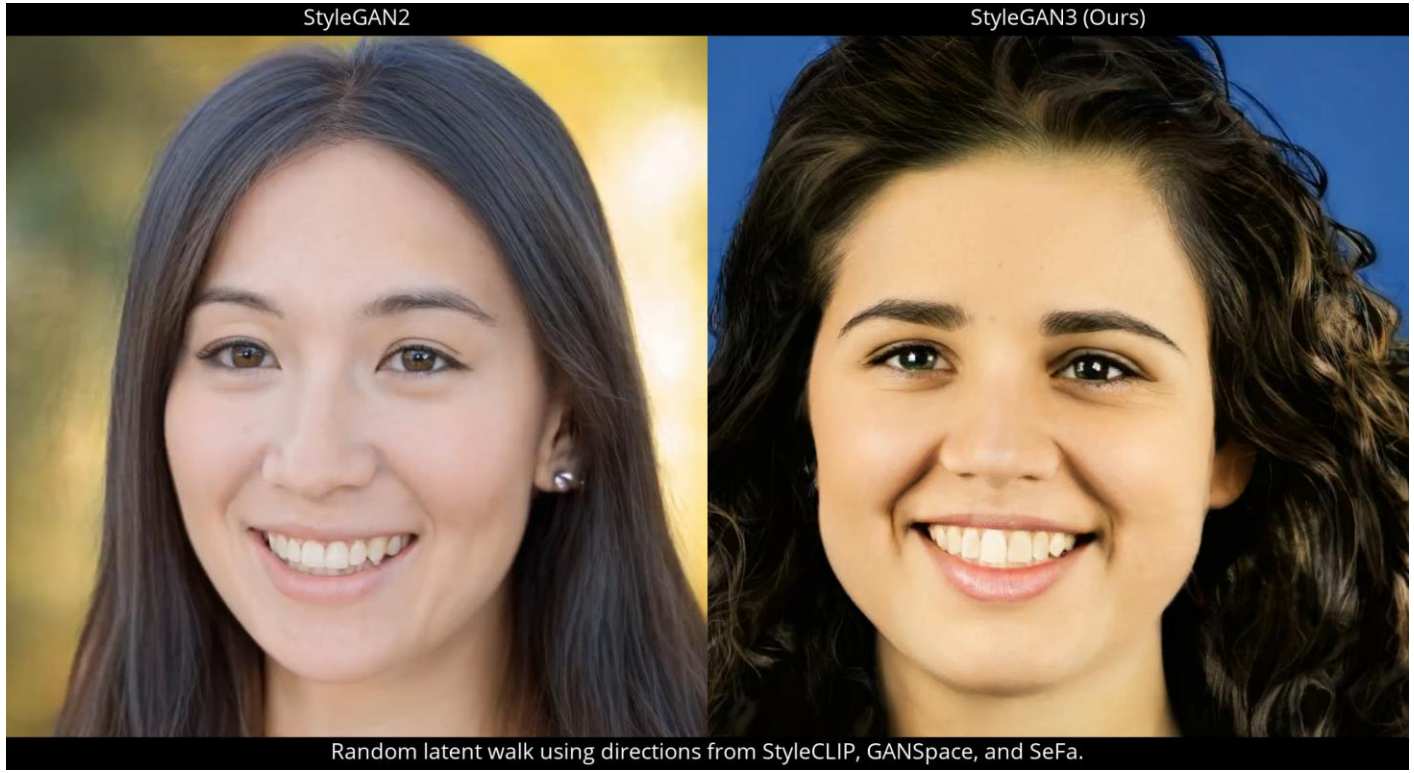
StyleGAN2 —— 其它结果



高质量图像生成

StyleGAN2 的缺陷

- 纹理粘附: 图片渐变时, 纹理留在原地



高质量图像生成

StyleGAN3 —— 解决移动一致性

Karras T, et al. Alias-Free Generative Adversarial Networks. NIPS 2021.

高质量图像生成

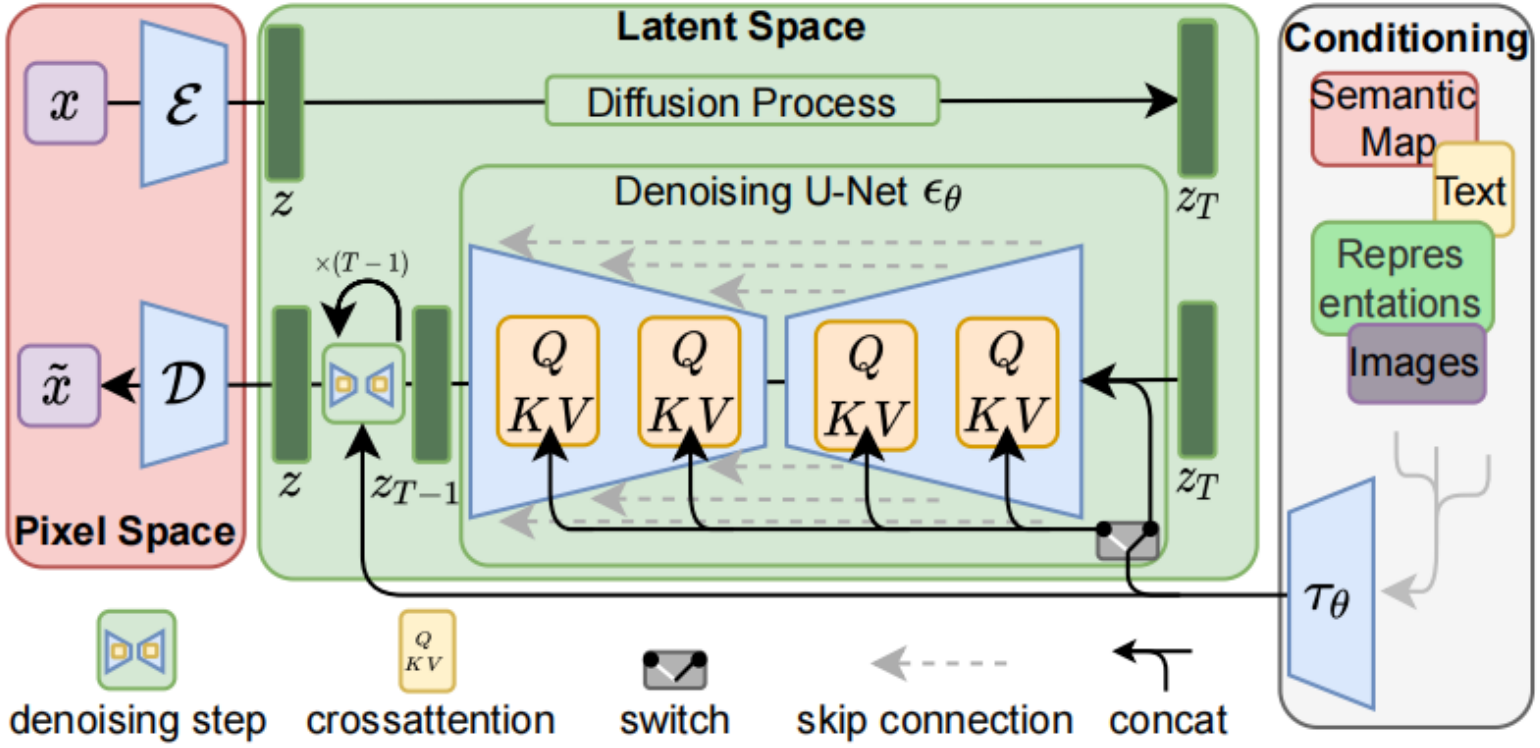
Stable Diffusion

- 针对扩散模型的缺点进行改进
- 扩散模型的缺点:
  - 训练代价非常巨大（上千GPU*Day）
  - 采样速度非常缓慢（无GPU几乎无法采样, 普通GPU采样常耗时数分钟）

高质量图像生成

Stable Diffusion

- 解决运算量问题的方法: 对图像的低尺度特征构建扩散模型

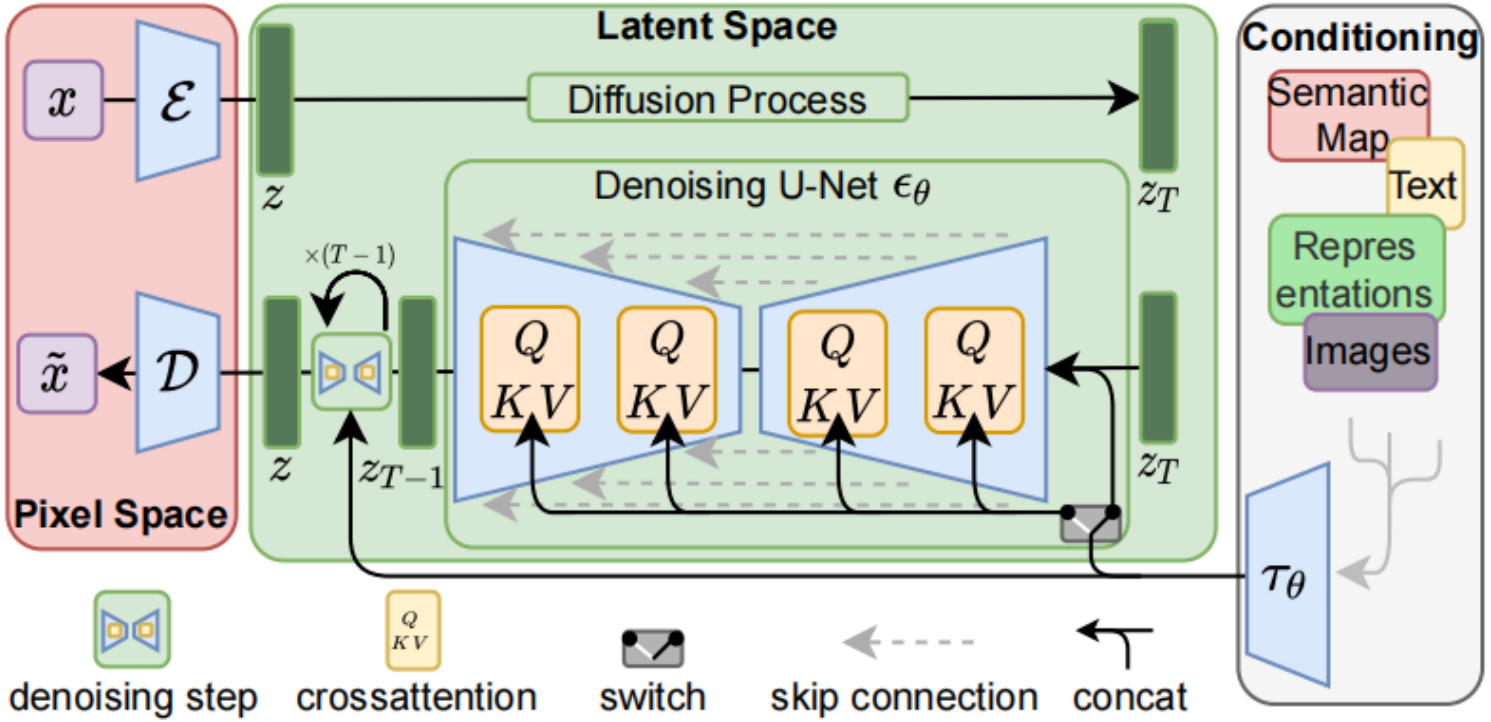


Rombach R, et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022.

高质量图像生成

Stable Diffusion

- 额外训练一个图像的编码器、解码器
- 将去噪的对象从原图像变成对图像编码得到的特征，该特征的尺度小于原图

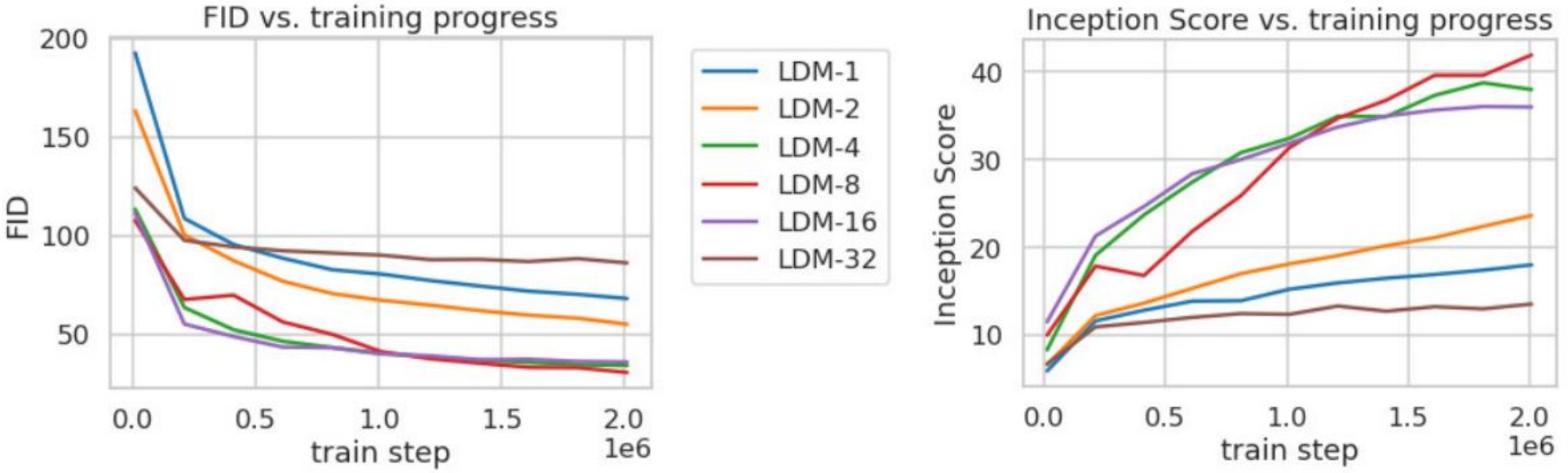


Rombach R, et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022.

高质量图像生成

Stable Diffusion

- 在不至于损失太多采样质量的情况下，图像抽取的特征至多有多小?



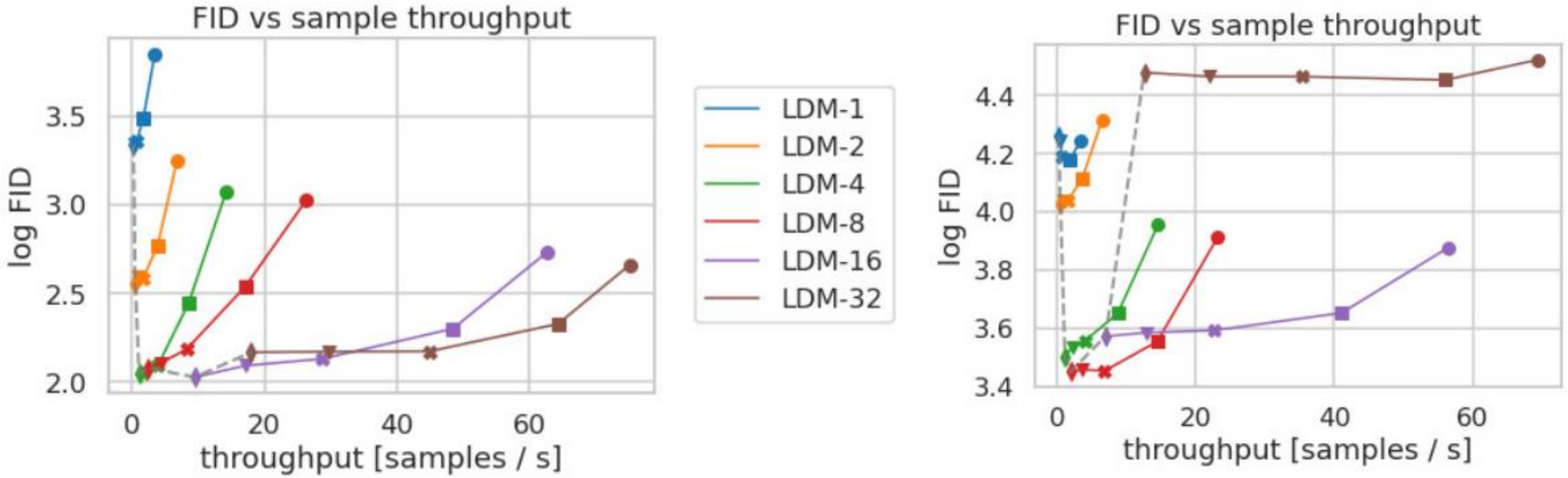
训练过程中的性能曲线，LDM-X 代表尺度为对原图尺度的1/X的特征构建扩散模型

可以看到 LDM-4/8 表现相对更优

高质量图像生成

Stable Diffusion

- 在不至于损失太多采样质量的情况下，图像抽取的特征至多有多小?



训练好的模型分别在 CelebA-HQ 和 ImageNet 数据集上采样的质量

同样也是 LDM-4/8 表现更好

高质量图像生成

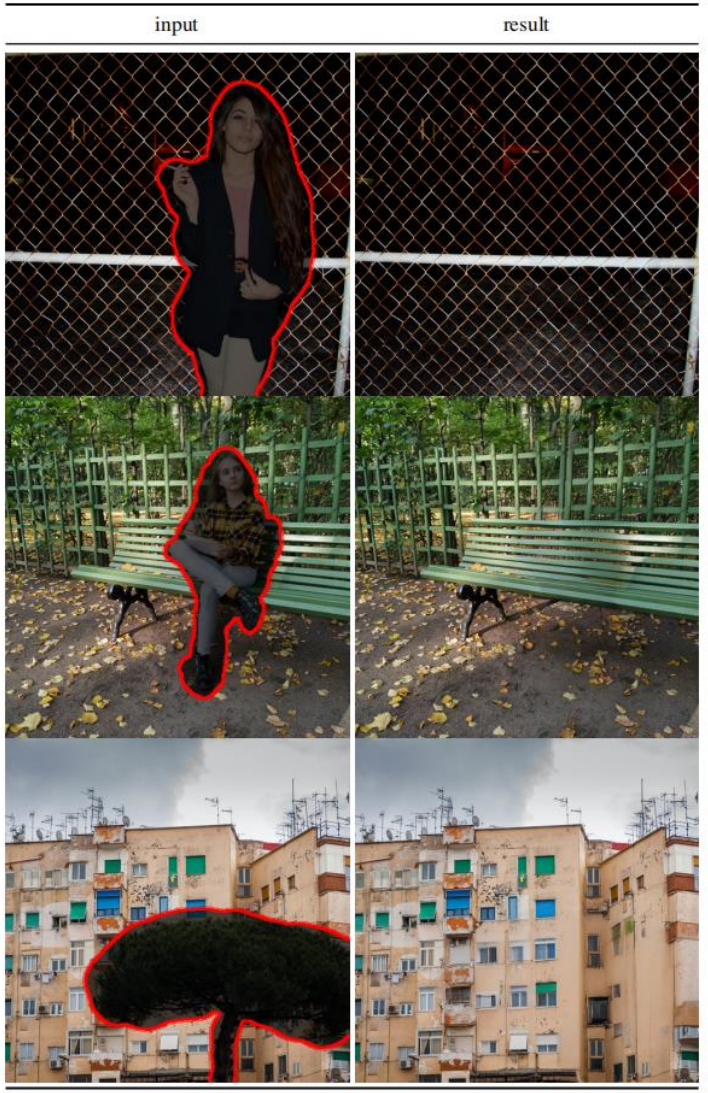
Stable Diffusion

- 得益于对低尺度特征搭建扩散模型的优点, LDM-8 可以提升数十倍的训练和采样速度:
  - 训练可在单张 80G A100 显卡上完成
  - 在普通的 GPU 上采样只需约数十秒

高质量图像生成

- Stable Diffusion 可用于多样化的任务

语义分割图→图像

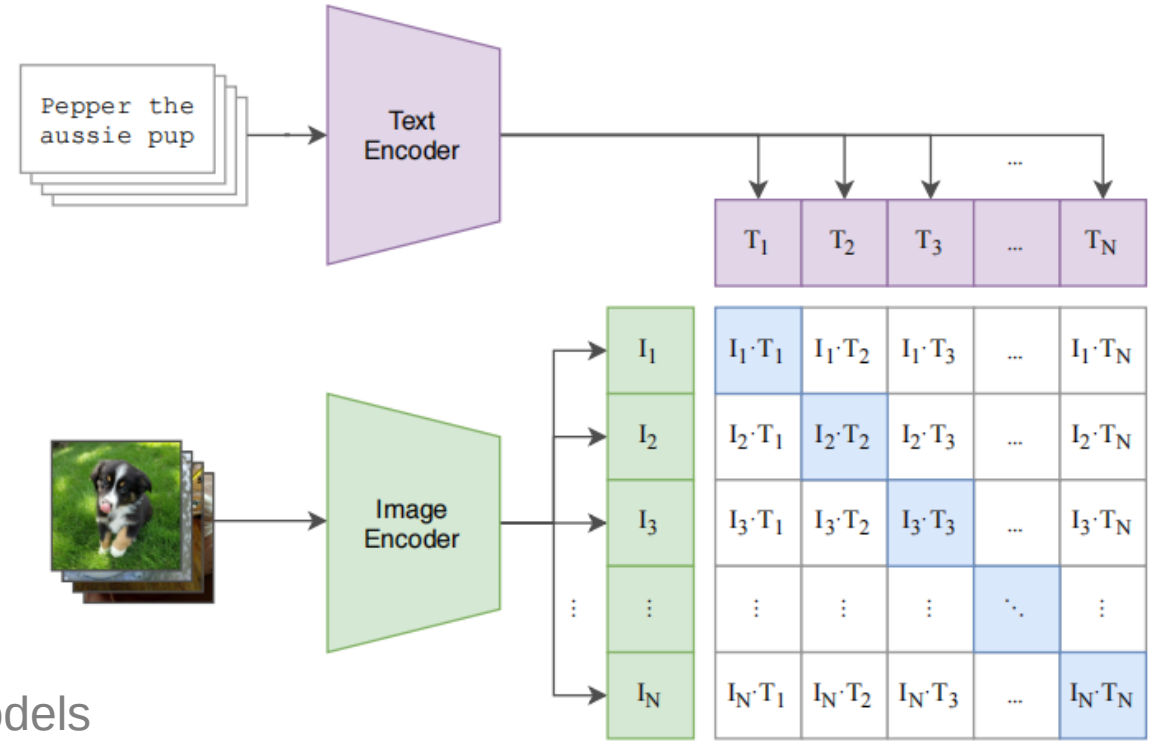


图像补全

Rombach R, et al. High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

文本-图像跨模态生成

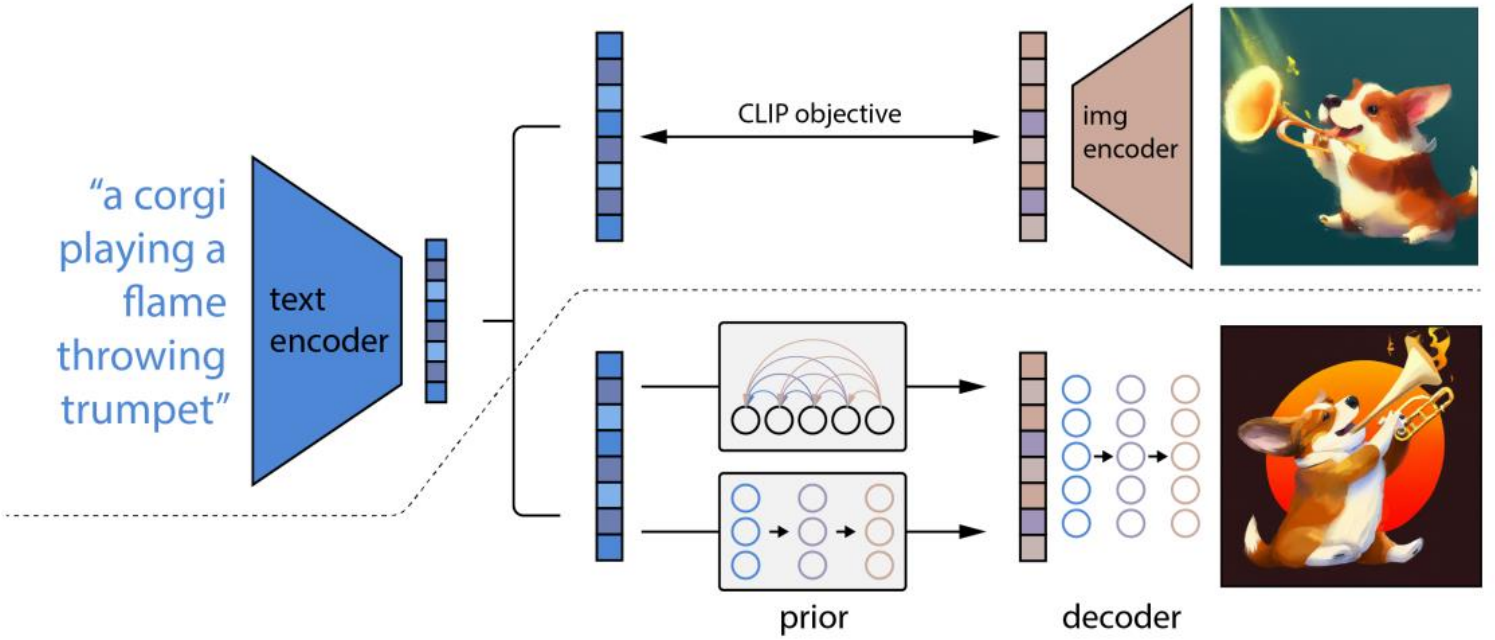
- CLIP (Contrastive Language Image Pretraining)
  - 用于文本-图像语义的跨模态对齐用的工具
  - 基于对比学习训练
  - 跨模态生成的技术基础



Radford A, et al. Learning transferable visual models from natural language supervision, ICML 2021.

文本-图像跨模态生成

- DALL-E 2
  - 利用扩散模型的强大能力直接进行模态跨越
  - 构建了两个扩散模型:
    - 跨越CLIP的文本隐空间和图像隐空间
    - 基于CLIP图像隐向量进行图像生成



Ramesh A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv 2022.

文本-图像跨模态生成

- DALL-E 2
  - 扩散模型难以一次性生成高质量图像，

    因此其后续附加了多个上采样用扩散模型 → 最终连接成完整的高质量

    文本-图像跨模态生成框架
  - 技术简单，效果惊艳
  - 进一步证明了扩散模型强大的容量和生成性能

文本-图像跨模态生成

- DALL-E 2

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Ramesh A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv 2022.

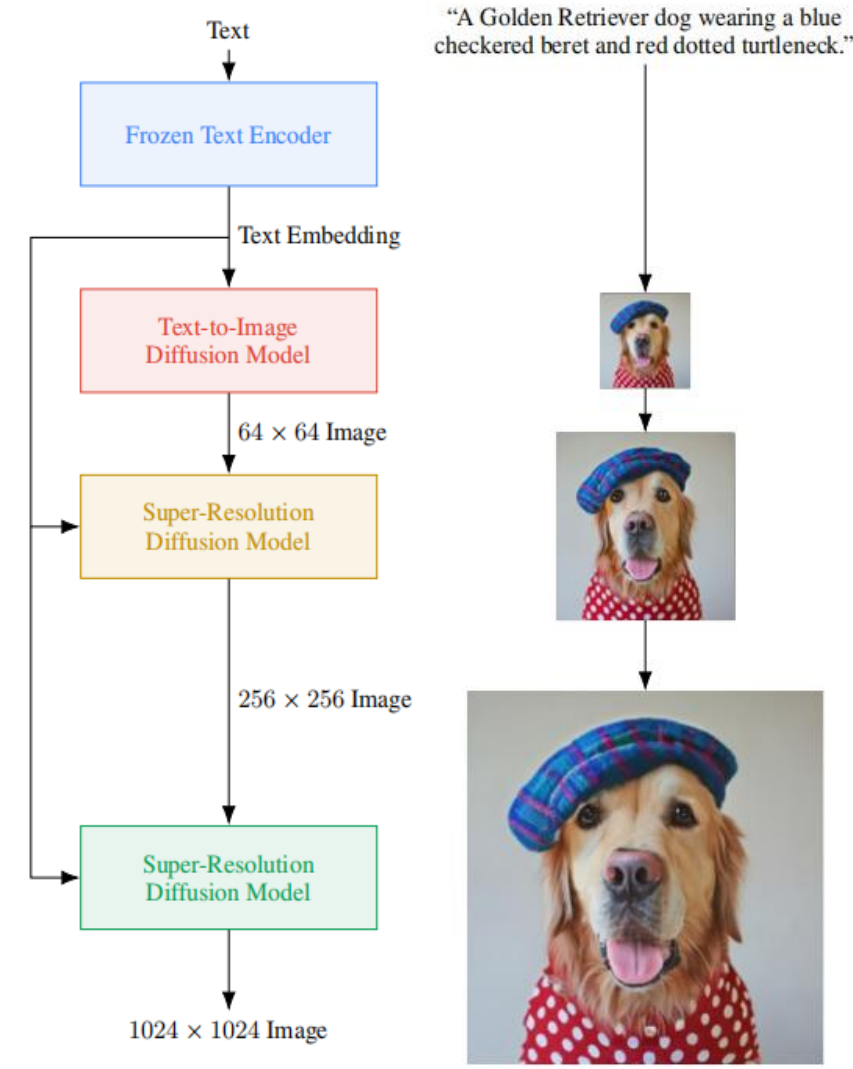
文本-图像跨模态生成

- DALL-E 2

(a) A high quality photo of a dog playing in a green field next to a lake.

文本-图像跨模态生成

- Imagen
  - DALL-E 2的缺点: 使用了 “两次跨越”, 不够直接
  - 解决这一问题:
    - Imagen直接将文本隐向量生成为图片, 仅一次跨越
    - 后同样接上采样扩散模型, 实现高分辨率图像生成



Saharia C, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, arXiv 2022.

文本-图像跨模态生成

- Imagen
  - 技术比 DALL-E 2 更加简单, 性能更加强大

Teddy bears swimming at the Olympics 400m Butter-fly event.

A cute corgi lives in a house made out of sushi.

文本-图像跨模态生成

- Imagen
  - 技术比 DALL-E 2 更加简单, 性能更加强大

A brain riding a rocketship heading towards the moon.

A dragon fruit wearing karate belt in the snow.