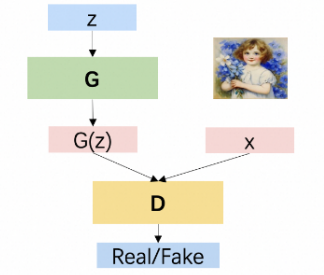
Sora底层技术

根据文生图的发展路线，我们把文生图的发展历程发展成如下4个阶段：

- 基于生成对抗网络（GAN）的模型
- 基于自回归（Autoregressive）的模型
- 基于扩散（diffusion）的模型
- 基于Transformers的扩散（diffusion）模型

基于生成对抗网络的（GAN）模型

文生图模型：GAN-based



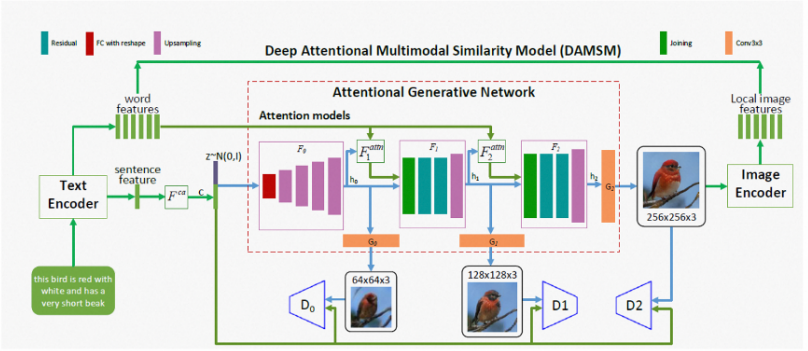
generator

$\tilde{x} \sim p_{\theta_1}(\tilde{x}|z)p(z),\ z \sim N(0,1)$

$d_{\theta_2}(x)$ discriminator loss

$x \sim p(x)$

Generative Adversarial Networks

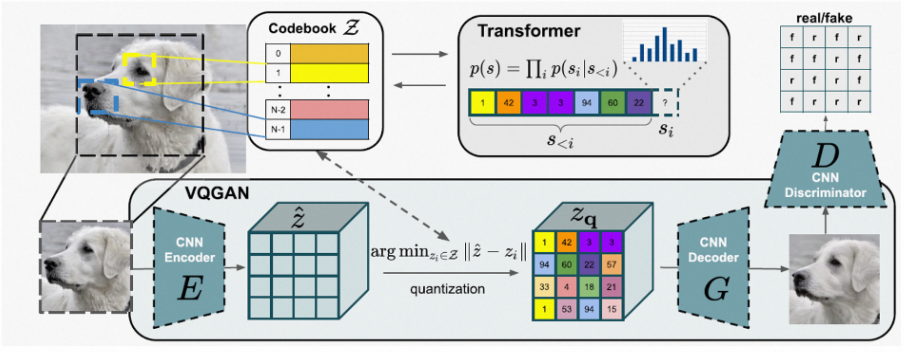


AttnGAN. Xu, Tao, et al 2017

优点	缺点
在一些窄分布（比如人脸）数据集上效果很好 采样速度快，方便嵌入到实时应用中	比较难训练、不稳定 有可能合成的样本都趋同（模式崩塌）

基于自回归方式的模型

文生图模型： Autoregressive Models



VQGAN. Esser, P. 2020

$\tilde{x} \sim p_\theta(x_N|x_1, \cdots, x_{N-1}) \cdots p_\theta(x_2|x_1)p_\theta(x_1)$

$p_\theta(x_1) \sim$ categorical distribution

$p_\theta(x_k|x_1, \cdots, x_{k-1}) \sim$ categorical distribution

自回归方式在自然语言中用的比较多（比如GPT）

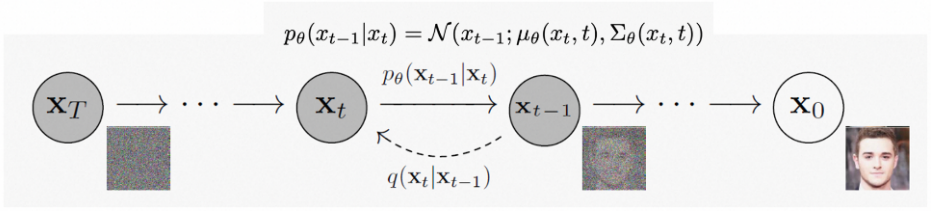
优点	缺点
训练稳定 大数据拟合能力强 有良好的扩展性	推理速度慢 框架灵活度低，难以修改

基于扩散（diffusion）方式的模型

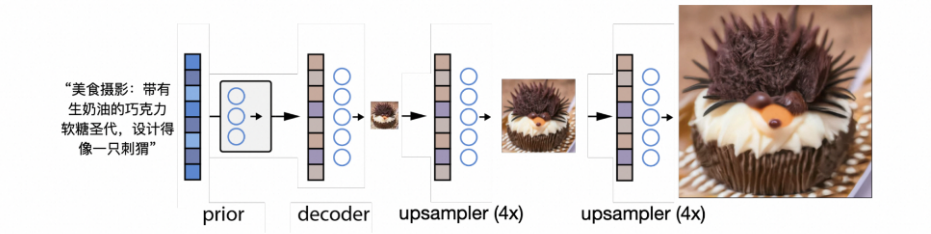
这是目前主流的技术，扩散模型也分为两个过程：

1. 前向过程：通过向原始数据不断加入高斯噪声来破坏训练数据，最终加噪声到一定步数之后，原始数据信息就完全被破坏，无限接近与一个纯噪声。
2. 反向过程：通过深度网络来去噪，来学习恢复数据。

训练完成之后，我们可以通过输入随机噪声，传递给去噪过程来生成数据。



DDPM. Ho, Jonathan, et al. 2020

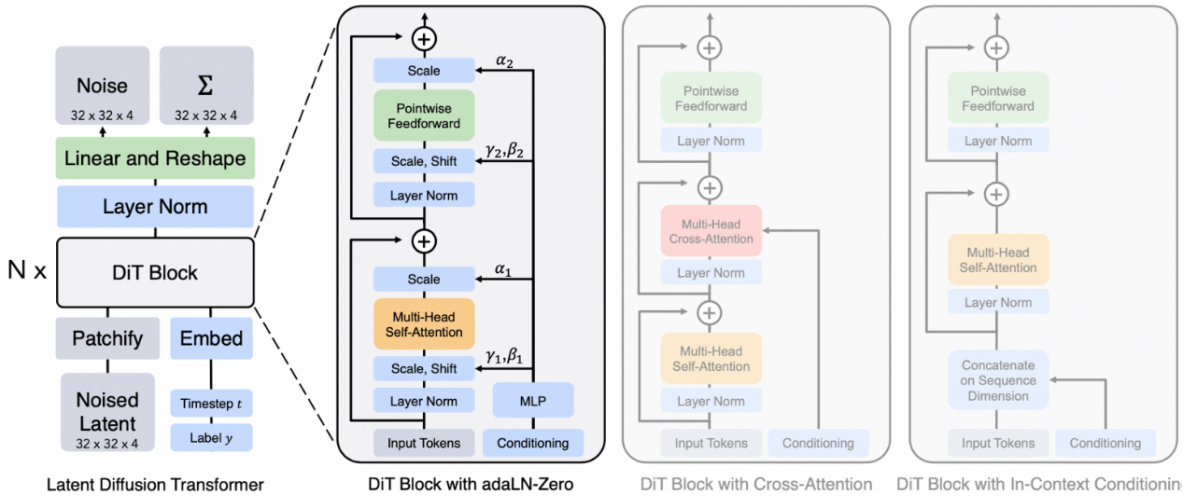


DALL·E 2, 2022.04

优点	缺点
大数据拟合能力强 复杂分布拟合能力强 有良好的扩展性、可编辑性 生成的图片质量高、多样	推理速度慢 缺乏语义解耦的隐空间

基于Transformers的架构的Diffusion模型

设计了一个简单而通用的基于Vision Transformers（ViT）的架构（U-ViT），替换了latent diffusion model中的U-Net部分中的卷积神经网络（CNN），用于diffusion模型的图像生成任务。



如上图，U-ViT遵循了Transformers的设计方法，将包括时间、条件和噪声图像patches在内的所有输入都视作为token。推理链路如下：

1. 输入一张$256 \times 256 \times 3$的图片,经过Encoder后得到对应的latent，压缩比为8，latent space推理时输入$32 \times 32 \times 4$的噪声，将latentspace的输入token化，图片使用patchify，label和timestep使用embedding。
2. 结合当前的step t , 输入label y， 经过N个Dit Block通过 MLP进行输出，得到输出的噪声以及对应的协方差矩阵
3. 经过T个step采样,得到32x32x4的降噪后的latent

在训练时，需要使得去躁后的latent和第一步得到的latent尽可能一致。

（未完待续）