去噪扩散概率模型

乔纳森·何加州大学
伯克利分校
jonathanho@berkeley.edu

Ajay Jain加州
大学伯克利分校
ajayj@berkeley.edu

Pieter Abbeel加州大
学伯克利分校
pabbeel@cs.berkeley.edu

抽象的

我们使用扩散概率模型（一类受非平衡热力学启发的隐变量模型）呈现了高质量的图像合成结果。我们的最佳结果是通过在一个加权变分界限上进行训练获得的，该界限是根据扩散概率模型与基于朗之万动力学的去噪得分匹配之间的新联系而设计的。我们的模型自然地采用了一种渐进式有损解压缩方案，这可以被解释为自回归解码的泛化。在非条件 CIFAR10 数据集上，我们获得了 9.46 的初始得分和 3.17 的最佳 FID 得分。在256x256 LSUN 上，我们获得了与 ProgressiveGAN 类似的样本质量。我们的实现可在https://github.com/hojonathanho/diffusion 获取。

1 简介

近年来，各类深度生成模型已在多种数据模态中展现出高质量的样本。生成对抗网络 (GAN)、自回归模型、流程和变分自编码器 (VAE) 已合成出令人惊叹的图像和音频样本[14, 27, 3, 58, 38, 25, 10, 32, 44, 57, 26, 33, 45]，并且在基于能量的建模和分数匹配方面也取得了显著进展，生成的图像质量已可与 GAN 媲美 [11, 55]。

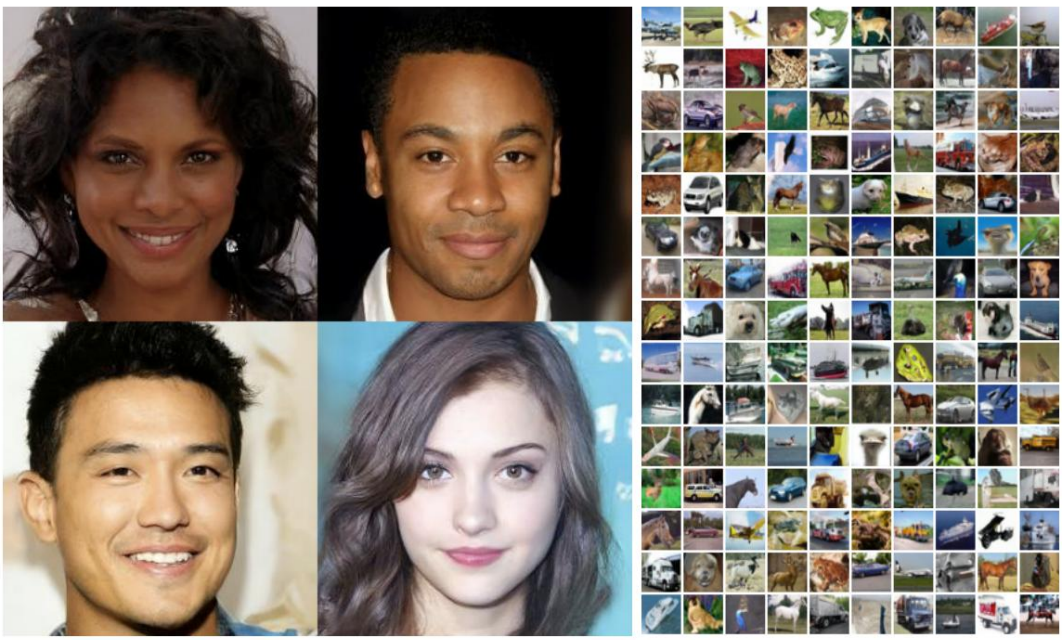


图 1：在 CelebA-HQ 256 × 256（左）和无条件 CIFAR10（右）上生成的样本

第 34 届神经信息处理系统会议（NeurIPS 2020），加拿大温哥华。

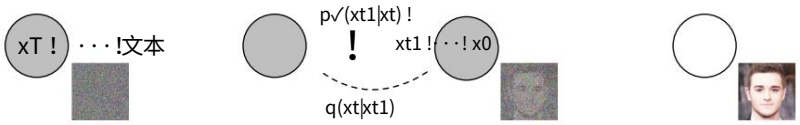


图 2:本研究中考虑的有向图模型。

本文介绍了扩散概率模型的进展[53]。扩散概率模型（简称“扩散模型”）是一个参数化的马尔可夫链,使用变分推理进行训练,在有限时间后生成与数据匹配的样本。该链的转换被学习以逆转扩散过程。扩散过程是一个马尔可夫链,它会在与采样相反的方向上逐渐向数据中添加噪声,直到信号被破坏。当扩散包含少量高斯噪声时,将采样链转换也设置为条件高斯即可,从而实现特别简单的神经网络参数化。

扩散模型定义简单,训练高效,但据我们所知,尚无证据表明它们能够生成高质量样本。我们证明了扩散模型实际上能够生成高质量样本,有时甚至优于已发表的其他类型生成模型的结果（第 4 节）。此外,我们还证明了扩散模型的某种参数化方法与训练过程中在多个噪声水平上进行去噪得分匹配以及采样过程中采用退火朗之万动力学方法具有等价性（第 3.2 节） [55, 61]。我们使用此参数化方法获得了最佳样本质量结果（第 4.2 节）,因此我们认为这种等价性是我们的主要贡献之一。

尽管我们的模型样本质量较高,但与其他基于似然的模型相比,我们的模型的对数似然值并不具有竞争力（不过,我们的模型的对数似然值确实优于已报道的基于能量的模型和分数匹配的退火重要性抽样产生的较大估计值[11, 55]）。我们发现,我们模型的大多数无损码长都用于描述难以察觉的图像细节（第 4.3 节）。我们用有损压缩的语言对这一现象进行了更精细的分析,并表明扩散模型的采样过程是一种渐进式解码,类似于沿位排序的自回归解码,这极大地概括了自回归模型通常可能实现的功能。

2 背景

扩散模型[53]是形式为pθ(x0) := pθ(x0:T) dx1:T的隐变量模型,其中xT是与数据x0 q (x0)同维的隐变量。联合分布x1, …, pθ (x0:T)被称为逆过程,它被定义为一个马尔可夫链,具有学习到的高斯转移,起始于p(xT) = N(xT ; 0, I):

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \qquad p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{1}$$

扩散模型与其他类型的潜在变量模型的区别在于,近似后验q(x1:T |x0)（称为前向过程或扩散过程)固定在马尔可夫链上,该链根据方差表β1, ..., βT逐渐将高斯噪声添加到数据中:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \qquad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

训练是通过优化负对数似然的通常变分界限来进行的:

$$\mathbb{E}[-\log p_\theta(x_0)] \leqslant \mathbb{E}_q[-\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] = \text{等式} -\log p(x_T) - \sum_{t\geqslant 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} =: L \tag{3}$$

正向过程的方差βt可以通过重新参数化[33]来学习,也可以作为超参数保持不变;逆向过程的表达能力部分由pθ(xt−1|xt)中高斯条件句的选择来保证,因为当βt较小时,两个过程具有相同的函数形式[53]。正向过程的一个显著特性是,它允许在任意时间步 t 以闭式形式对xt进行采样:使用符号αt := 1 − βt和 $\bar{\alpha}_t := q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$,有（4）

因此,通过随机梯度下降优化L的随机项,可以实现高效的训练。通过将 L (3) 重写为：

$$\text{等式} \underbrace{D_{KL}(q(x_T|x_0)\,p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)\,p_\theta(x_{t-1}|x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0} \tag{5}$$

（详情见附录 A。术语的标签在第 3 节中使用。）等式 (5) 使用 KL散度直接将pθ(xt−1|xt)与前向过程后验进行比较,当以x0 为条件时,后验是可处理的： $q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \mu_t(x_t, x_0), \beta_t I)$, $\sqrt{\alpha}_{t-1}\beta_t$其中 $\mu_t(x_t, x_0) := x_0 + 1 - \alpha_t$

(6)

$$\frac{\sqrt{\alpha_t}(1-\alpha_{t-1})}{\alpha_t 1-\alpha_t} \frac{1-\alpha_{t-1}}{} x_t \text{和} \beta_t := \beta_t \frac{1-}{} \tag{7}$$

因此,公式 (5) 中的所有 KL 散度都是高斯之间的比较,因此它们可以采用 Rao-Blackwellized 方式计算,采用封闭形式表达式,而不是高方差蒙特卡洛估计。

3 扩散模型和去噪自动编码器

扩散模型看似是一类受限的潜变量模型,但它们在实现上却拥有巨大的自由度。我们必须选择正向过程的方差βt ,以及逆向过程的模型架构和高斯分布参数化。为了指导我们的选择,我们在扩散模型和去噪分数匹配之间建立了新的明确联系（第3.2节）,从而为扩散模型提供了一个简化的加权变分边界目标函数（第3.4节）。最终,我们的模型设计通过简洁性和实证结果得到了验证（第4节）。我们的讨论可按照公式(5)进行分类。

3.1 正向过程和LT

我们忽略了前向过程方差βt可以通过重新参数化学习的事实,而是将其固定为常数（详见第 4 节）。因此,在我们的实现中,近似后验概率 q 没有可学习的参数,因此LT在训练期间是一个常数,可以忽略不计。

3.2 逆过程与L1:T −1

现在我们讨论其中$1 < t \leqslant T$时$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$的选择。首先,我们设置$\Sigma_\theta(x_t, t) = \sigma_t^2 = \beta_t$和$\sigma_t^2$ I为未训练的时间相关常数。实验上， $\sigma_t^2 = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t$的结果类似。第一个选择对于$x_0 \sim N(0, I)$ 是最优的,第二个选择对于x0确定性地设为一个点是最优的。这两个极端选择分别对应于坐标系方差为单位的数据的逆过程熵的上下界[53]。

其次,为了表示平均值$\mu_\theta(x_t, t)$,我们提出了一个特定的参数化方法,其动机是对Lt进行以下分析。其中$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$,我们可以写成：

$$L_{t-1} = \text{等式} \frac{1}{2\sigma_t^2} \left\| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \right\|^2 + C \tag{8}$$

其中C是一个与 θ 无关的常数。因此,我们看到， μ_θ最直接的参数化方法是预测前向过程后验均值$\tilde{\mu}_t$的模型。然而,我们可以进一步扩展公式 (8),将公式 (4) 重新参数化为$x_t(x_0, \epsilon) = \sqrt{\alpha}_t x_0 + \sqrt{1-\alpha_t}\epsilon$,其中 $\epsilon \sim N(0, I)$,并应用前向过程后验公式 (7)：

$$L_{t-1} - C = E_{x_0, \epsilon}\left[\frac{1}{2\sigma_t^2}\left\| \frac{1}{\sqrt{\alpha_t}}(x_t(x_0, \epsilon) - \sqrt{1-\alpha_t}\epsilon) - \mu_\theta(x_t(x_0, \epsilon), t) \right\|^2\right] \tag{9}$$

$$= E_{x_0, \epsilon}\left[\frac{1}{2\sigma_t^2}\left\| \frac{1}{\sqrt{\alpha_t}}\left(x_t(x_0, \epsilon) - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon\right) - \mu_\theta(x_t(x_0, \epsilon), t) \right\|^2\right] \tag{10}$$

算法 1训练1:重复2: x_0

$q(x_0)$ 3:t
Uniform({1, . . . ,
T}) 4: N (0, I)

5:在∇θ − θ(√ α¯tx0 + √ 1 − α¯t, t)上 2
进行梯度下降6:直到收敛

算法2采样

1: xT N (0, I) 2:对于t
= T, . . . , 1 do 3:z N (0, I) 如果
t > 1,否则 z = 0 4: xt−1 = √1 αt xt − √1−αt 1−α¯t
θ(xt, t) + σtz
5:结束6:返回
x0

等式 (10) 表明μθ必须预测模型的√ √ 1−α¯t输入,我们可以选择参数化 $\frac{}{\alpha}$ xt − $\frac{\beta t1}{}$ 给定xt。由于xt可表示为

$$\frac{1}{} \qquad \frac{1}{} \; x_t - \frac{\beta_t}{} \tag{11}$$

√ αt其中θ是一个函数逼近器,用于根据xt 进行预测。对xt−1 pθ(xt−1|xt)进行采样,即θ(xt), t) + σtz,其中z μ N(0, I),完整的采样计算xt−1 = √ αt 的过程,即算法 2,类似于朗之万动力学,其中θ是数据密度的学习梯度。此外,利用参数化 (11),等式 (10) 简化为:

$$\frac{1}{} \; x_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}$$

$$E_{x_0,} \; \frac{\beta_t^2}{2\sigma_t^2 \alpha_t(1-\bar\alpha_t)} \; - \theta(\sqrt{\bar\alpha_t} x_0 + \sqrt{1-\bar\alpha_t}, t)^2 \tag{12}$$

它类似于在以t为指标的多个噪声尺度上进行去噪分数匹配[55]。由于公式 (12)等于类朗之万逆过程 (11) 的变分界限(其中一项),由此可见,优化一个类似于去噪分数匹配的目标函数,等价于使用变分推理来拟合一个类似于朗之万动力学的采样链的有限时间边际。

总而言之,我们可以训练逆过程均值函数逼近器μθ来预测μ t ,或者通过修改其参数化,我们可以训练它来预测 。(也可以预测x0,但我们在实验初期发现这会导致样本质量下降。)我们已经证明 -预测参数化既类似于朗之万动力学,又将扩散模型的变分界限简化为类似于去噪分数匹配的目标。尽管如此,它只是pθ(xt−1|xt) 的另一种参数化,因此我们在第 4 节中通过比较预测与预测 μ t 来验证其有效性。

3.3 数据缩放、逆过程解码器和L0

我们假设图像数据由{0, 1, ..., 255}中的整数组成,并线性缩放到[-1, 1]。这确保了神经网络的逆过程能够对从标准正态先验p(xT) 开始的、经过一致缩放的输入进行操作。为了获得离散对数似然值,我们将逆过程的最后一项设置为一个独立的离散解码器,该解码器源自高斯分布 N (x0; μθ (x1, 1), σ2 1 I):

$$p_\theta(x_0|x_1) = \prod_{i=1}^{D} \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} N(x; \mu_\theta^i(x_1, 1), \sigma_1^2)\, dx$$

$$\delta_+(x) = \begin{cases} \infty & \text{如果 } x = 1, \text{则} \\ x + \frac{1}{255} & \text{如果 } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{如果 } x = -1, \text{则为} \\ x - \frac{1}{255} & \text{如果 } x > -1 \end{cases} \tag{13}$$

其中D是数据维数, i上标表示提取一个坐标。

(其实,更简单的方法是加入一个更强大的解码器,比如条件自回归模型,但这留待以后研究。)与VAE 解码器和自回归模型[34, 52] 中使用的离散化连续分布类似,我们在此的选择确保了变分界限是离散数据的无损码长,无需向数据中添加噪声,也无需将缩放操作的雅可比矩阵合并到对数似然中。在采样结束时,我们无噪声地显示μθ (x1, 1)。

3.4 简化的训练目标

通过上面定义的逆过程和解码器,由方程 (12) 和 (13) 导出的变分界限关于θ明显可微,可以用于

表 1： CIFAR10 结果。NLL 以比特/维度为单位。

模型	是	FID NLL 测试（培训）	
条件			
循证医学 [11]	8.30	37.9	
正义运动 [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	10.06	2.67	
无条件			
扩散（原始）[53]			≤5.40
门控PixelCNN[59]	4.60	65.93	3.03（2.90）
稀疏变换器[7]			2.80
PixelIQN [43] 49.46	5.29		
循证医学 [11] 38.2	6.78		
NCSNv2 [56] 31.75			
NCSN [55] 8.87±0.12 25.32			
SNGAN [39] 8.22±0.05 21.7			
SNGAN-DDLS [4] 9.09±0.10 15.42			
StyleGAN2 + ADA (v1) [29] 9.74 ± 0.05 3.26			
我们的（L，固定各向同性Σ）	7.67±0.13 13.51	≤3.70（3.69）	
我们的（Lsimple） 9.46±0.11	3.17	≤3.75 （3.72）	

表 2:无条件 CIFAR10 反向
过程参数化和训练目标消融。空白条目不稳定
训练并生成分数超出范围的差样本。

客观的	是	
μ预测（基线）		
L,学习对角线Σ 7.28±0.10 23.69		
L,固定各向同性Σ 8.06±0.09 13.22		
$\|\mu - \mu_\theta\|^2$	–	–
预测（我们的）		
L,学习对角线Σ	–	–
L,固定各向同性Σ	7.67±0.13 13.51	
$\|\tilde{\epsilon} - \epsilon_\theta\|^2$ （简单） 9.46±0.11 3.17		

训练。然而，我们发现，在
变分界限的以下变体：

$$L_{simple}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right] \quad (14)$$

其中t在1和T之间均匀分布。t = 1 的情况对应于L0 ，其中积分
离散译码器定义（13)近似于高斯概率密度函数乘以
箱宽，忽略σ_1^2 和边缘效应。t > 1 的情况对应于未加权的
等式（12)类似于 NCSN 去噪分数匹配模型[55]使用的损失加权。
（LT不会出现，因为前向过程方差βt是固定的。)算法 1 显示
通过这个简化的目标完成训练程序。

由于我们的简化目标 (14) 丢弃了方程 (12) 中的权重，因此它是一个加权变分
与标准变分相比，强调重建不同方面的界限
界限[18, 22]。特别是，我们在第 4 节中设置的扩散过程导致了简化的目标
对应于较小 t 的下行权重损失项。这些项训练网络对数据进行去噪
噪音很小，因此降低它们的权重是有益的，这样网络就可以
专注于更大t项下更困难的去噪任务。我们将在实验中看到，这
重新加权可提高样本质量。

4 实验

我们将所有实验的T = 1000设置为所需的神经网络评估次数
采样过程中的方差与先前的研究结果一致[53, 55]。我们将前向过程方差设置为常数
从β1 = 10－4线性增加到βT = 0.02。这些常数被选择得较小
相对于缩放到[－1, 1] 的数据，确保反向和正向过程大约
相同的函数形式，同时保持xT处的信噪比尽可能小（LT =
在我们的实验中， $D_{KL}(q(x_T|x_0)\|\mathcal{N}(0, I)) \approx 10$－5位/维度）。

为了表示逆向过程，我们使用了类似于未掩蔽的PixelCNN++ 的U-Net 主干[52，
48]并始终进行组归一化[66]。参数跨时间共享，这在
使用 Transformer 正弦位置嵌入[60] 到网络中。我们使用自注意力机制
16×16特征图分辨率[63, 60]。详情见附录B。

4.1 样品质量

表 1 展示了 Inception 分数、FID 分数和负对数似然值（无损码长）
CIFAR10。我们的无条件模型的 FID 得分为 3.17，其样本质量优于
文献中的大多数模型，包括类条件模型。我们的 FID 得分是使用
按照标准做法，对训练集进行计算；当我们根据测试集进行计算时，
分数为 5.24，仍然比文献中的许多训练集 FID 分数要好。

图 3： LSUN 教会样本。FID=7.89

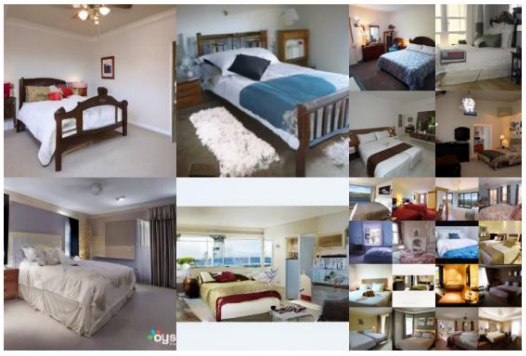


图4： LSUN卧室样品。FID=4.90

算法 3发送x0 1:使用 p(xT)发送xT q(xT |x0) 2:对于t = T − 1, . . . , 2, 1执行3:使用pθ (xt|xt+1)发送 xt q (xt|xt+1, x0) 4:结束

5:使用pθ(x0|x1)发送x0

算法4接收

1:使用 p(xT)接收xT 2:对于t = T − 1, . . . , 1, 0执行3:使用pθ (xt|xt+1)接收xt 4:结束5:返回x0

我们发现,与预期一致,在真实变分界限上训练模型比在简化目标上训练模型能获得更好的代码长度,但后者的样本质量最佳。图 1 为CIFAR10 和 CelebA-HQ 256 × 256样本,图 3 和图 4 为 LSUN 256 × 256样本[71],更多信息请参见附录 D。

4.2 逆向过程参数化与训练目标消融

在表 2 中,我们展示了逆过程参数化和训练目标对样本质量的影响（第 3.2 节）。我们发现,预测μ的基线选项仅在使用真实变分界限而非未加权均方误差进行训练时效果良好,这是一个类似于公式 (14) 的简化目标。我们还看到,学习逆过程方差（通过将参数化的对角线Σθ(xt)纳入变分界限)会导致训练不稳定,并且与固定方差相比样本质量更差。正如我们所提出的,在使用固定方差的变分界限进行训练时,预测的表现大致与预测μ一样好,但在使用我们的简化目标进行训练时,效果要好得多。

4.3 渐进式编码

表 1 还展示了我们 CIFAR10 模型的码长。训练集和测试集之间的差距最多为每维 0.03 位,这与其他基于似然度的模型报告的差距相当,表明我们的扩散模型并未过拟合（最近邻可视化结果见附录 D）。尽管我们的无损码长优于基于能量的模型和使用退火重要性采样的得分匹配[11] 报告的较大估计值,但它们与其他类型的基于似然度的生成模型 [7] 相比仍不具竞争力。

尽管如此,由于我们的样本质量很高,我们得出结论:扩散模型具有归纳偏差,这使得它们成为优秀的有损压缩器。将变分界限项L1 +· · ·+LT视为速率,将L0视为失真,我们具有最高质量样本的 CIFAR10 模型的速率为1.78位/维,失真为1.97位/维,这相当于在 0 到 255 的范围内均方根误差为0.95 。超过一半的无损码长描述了难以察觉的失真。

渐进式有损压缩我们可以通过引入与公式 (5) 形式相似的渐进式有损编码来进一步探究我们模型的率失真行为:参见算法 3 和算法 4,它们假设可以访问诸如最小随机编码[19, 20] 之类的程序,该程序可以平均使用大约DKL(q(x) p(x))位来传输样本x q(x) ,对于任何分布p和q,其中只有p事先可供接收方使用。当应用于x0 q(x0) 时,算法 3 和算法 4 传输xT

, ..., x0依次采用等于公式 (5) 的总预期码长。接收机

在任何时间 t,都有部分信息xt完全可用,并且可以逐步估计:

$$x0 \approx x \quad 0 = xt - \sqrt{1 - \bar{\alpha}t}\theta(xt) / \sqrt{\bar{\alpha}t} \tag{15}$$

根据等式 (4) 计算得出。(随机重构x0 pθ(x0|xt)也成立,但我们在此不予考虑,因为它会使失真更难评估。)图 5 展示了 CIFAR10 测试集上得到的率失真图。在每个时间t,失真计算为均方根误差x0 − x 0 2/D,率计算为迄今为止在时间t接收到的累计比特数。在率失真图的低率区域,失真急剧下降,表明大多数比特确实分配给了不可察觉的失真。

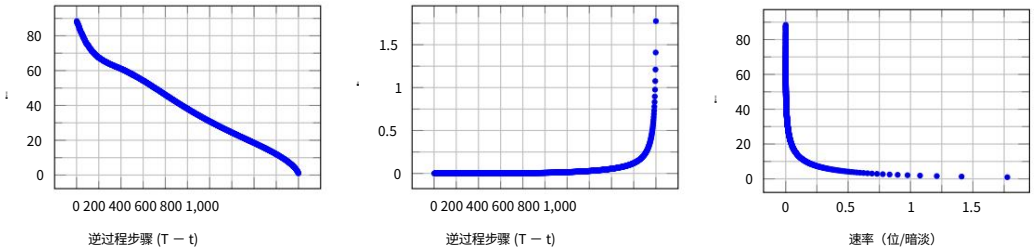


图 5:无条件 CIFAR10 测试集率失真与时间的关系。失真以 [0, 255] 范围内的均方根误差来衡量。详情见表 4。

渐进式生成我们还运行一个渐进式无条件生成过程,该过程由随机位的渐进式解压缩给出。换句话说,我们预测逆过程的结果 x 0,同时使用算法 2 从逆过程中进行采样。图 6 和图 10 显示了逆过程过程中x 0的样本质量。大规模图像特征首先出现,细节最后出现。图 7 显示了随机预测x0 pθ(x0|xt),其中xt在不同的t 下保持不变。当t较小时,除了精细细节之外的所有细节都会被保留,而当t较大时,只有大规模特征会被保留。也许这些是概念压缩的暗示 [18]。

图 6:无条件 CIFAR10 渐进式生成（x 0随时间变化,从左到右）。附录中提供了随时间变化的扩展样本和样本质量指标（图 10 和图 14）。

图 7:当以相同潜在值为条件时,CelebA-HQ 256×256样本共享高级属性。
右下象限是xt,其他象限是来自pθ(x0|xt) 的样本。

与自回归解码的联系请注意,变分界限 (5) 可以重写为:

$$L = DKL(q(xT) \| p(xT)) + \text{等式} \quad \sum_{t \geqslant 1} DKL(q(xt-1|xt) \| p\theta(xt-1|xt)) + H(x0) \tag{16}$$

（参见附录 A 的推导。)现在考虑将扩散过程长度T设置为数据的维数,定义正向过程,以便q(xt|x0)将所有概率质量放在x0上,并将前t 个坐标屏蔽掉（即q(xt|xt−1)屏蔽掉t坐标）,设置p(xT)将所有质量放在空白图像上,并且,为了论证的目的,取pθ(xt−1|xt)为

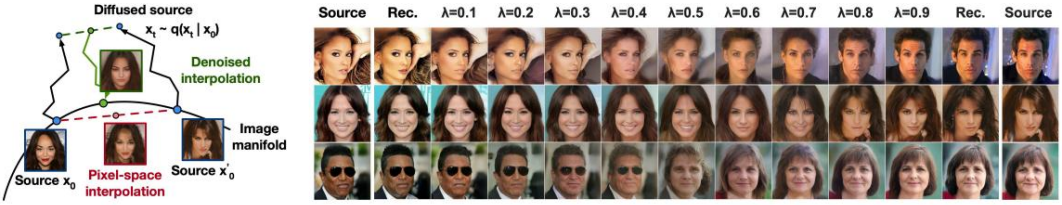


图 8:具有 500 个时间步长的扩散的 CelebA-HQ 256x256 图像的插值。

是一个完全可表达的条件分布。在这些选择下， DKL(q(xT) p(xT)) = 0,并且最小化DKL(q(xt−1|xt) pθ(xt−1|xt))训练pθ复制坐标t + 1, ..., T不变,并预测给定t + 1, ..., T 的t坐标。因此,使用这种特定的扩散训练pθ就是训练一个自回归模型。

因此,我们可以将高斯扩散模型 (2) 解释为一种具有广义位序的自回归模型,而这种位序无法通过重新排序数据坐标来表达。先前的研究表明,此类重新排序会引入归纳偏差,从而影响样本质量[38],因此我们推测高斯扩散也能达到类似的目的,甚至可能效果更佳,因为与掩蔽噪声相比,将高斯噪声添加到图像中可能更为自然。此外,高斯扩散的长度不限于等于数据维数;例如,我们使用T = 1000,这小于我们实验中32 × 32 × 3或256 × 256 × 3图像的维数。

高斯扩散可以做得更短以实现快速采样,或者做得更长以实现模型表达。

4.4 插值

p(x0|x¯t)。实际上,我们使用逆过程从线性插值损坏的源图像版本中 0 q(x0) ,使用q作为随机编码器,我们可以通过逆过程将源图像x0, x插值到图像xt, x空间中， x¯0 HQ 256 q(xt|x0),然后解码线性插值的潜在x¯t 去得到x0 插值。我们固定了不同λ值的噪声,因此xt和x保持不变。图 8（右）展示了原始 CelebA-× 256图像（t = 500)的插值和重建。逆过程产生高质量的重建和合理的插值,可以平滑地改变姿势、肤色、发型、表情和背景等属性,但不能平滑地改变眼镜。t 越大,插值越粗糙、变化越多,在 t = 1000 时会出现新的样本（附录图 9）。

5 相关工作

虽然扩散模型可能类似于流[9, 46, 10, 32, 5, 16, 23]和VAE [33, 47, 37],但扩散模型的设计使得q没有参数,并且顶层潜在xT与数据x0的互信息几乎为零。我们的-预测逆过程参数化建立了扩散模型与在多个噪声水平上使用退火朗之万动力学进行采样的去噪分数匹配之间的联系[55, 56]。然而,扩散模型允许直接进行对数似然评估,并且训练过程使用变分推断明确地训练朗之万动力学采样器（详见附录C）。这种联系也具有相反的含义,即某种加权形式的去噪分数匹配与训练类朗之万采样器的变分推断相同。学习马尔可夫链转换算子的其他方法包括注入训练[2]、变分回溯[15]、生成随机网络[1]等[50, 54, 36, 42, 35, 65]。

鉴于分数匹配与基于能量的建模之间的已知联系,我们的工作可能对其他近期基于能量的模型研究产生影响[67–69, 12, 70, 13, 11, 41, 17, 8]。我们的率失真曲线是在一次变分界限求值中随时间推移计算出来的,这让人联想到如何在一次退火重要性采样中计算出随失真惩罚而变化的率失真曲线[24]。我们的渐进式解码论证可以在卷积DRAW及其相关模型[18, 40]中得到体现,并且可能为自回归模型的子尺度排序或采样策略带来更通用的设计[38, 64]。

6 结论

我们已经使用扩散模型展示了高质量的图像样本，并发现了扩散模型与变分推理在训练马尔可夫链、去噪分数匹配和退火朗之万动力学（以及基于能量的模型）、自回归模型和渐进式有损压缩方面的联系。由于扩散模型似乎对图像数据具有出色的归纳偏差，我们期待研究它们在其他数据模态中的效用，以及作为其他类型生成模型和机器学习系统的组成部分。

更广泛的影响

我们对扩散模型的研究与现有其他类型深度生成模型的研究范围相似，例如致力于提升 GAN、流、自回归模型等样本质量。我们的论文代表了在使扩散模型成为此类技术中普遍适用的工具方面取得的进展，因此它或许能够放大生成模型对更广阔世界已经（以及未来）产生的影响。

遗憾的是，生成模型有许多众所周知的恶意用途。样本生成技术可用于制作知名人物的虚假图像和视频，以用于政治目的。虽然早在软件工具出现之前，虚假图像就已经是手动创建的，但像我们这样的生成模型使这一过程变得更容易。幸运的是，CNN 生成的图像目前存在一些可以被检测到的细微缺陷[62]，但生成模型的改进可能会使这变得更加困难。生成模型还会反映它们所训练的数据集中的偏见。由于许多大型数据集都是由自动化系统从互联网上收集的，因此很难消除这些偏见，尤其是在图像未标记的情况下。如果在这些数据集上训练的生成模型的样本在整个互联网上激增，那么这些偏见只会进一步加强。

另一方面，扩散模型可能有助于数据压缩。随着数据分辨率的提高和全球互联网流量的增加，数据压缩对于确保互联网对广大受众的可访问性至关重要。我们的工作或许有助于对未标记原始数据进行表征学习，以用于从图像分类到强化学习等一系列下游任务，扩散模型也可能在艺术、摄影和音乐领域的创意应用方面变得可行。

致谢和资金披露

这项工作得到了 ONR PECASE 和 NSF 研究生奖学金的支持，资助编号为 DGE-1752814。Google 的 TensorFlow 研究云 (TFRC) 提供了云 TPU。

参考

[1] Guillaume Alain、Yoshua Bengio、Li Yao、Jason Yosinski、Eric Thibodeau-Laufer、Saizheng Zhang 和Pascal Vincent。GSN：生成随机网络。《信息与推理：IMA 杂志》， 5(2):210–249，2016 年。

[2] Florian Bordes、Sina Honari 和 Pascal Vincent。学习通过注入从噪声中生成样本训练。在2017年国际学习表征会议上。

[3] Andrew Brock、Jeff Donahue 和 Karen Simonyan。大规模 GAN 训练，实现高保真自然图像合成。2019 年国际学习表征会议。

[4] Tong Che、Ruixiang Zhang、Jascha Sohl-Dickstein、Hugo Larochelle、Liam Paull、Yuan Cao 和 Yoshua Bengio。你的 GAN 本质上是一个基于能量的模型，你应该使用判别器驱动的潜在采样。arXiv 预印本 arXiv:2003.06060，2020 年。

[5] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, 和 David K Duvenaud. 神经常微分方程. 载于《神经信息处理系统进展》第 6571–6583 页，2018 年。

[6] Xi Chen、Nikhil Mishra、Mostafa Rohaninejad 和 Pieter Abbeel。PixelSNAIL：一种改进的自回归生成模型。国际机器学习会议，第 863-871 页，2018 年。

[7] Rewon Child、Scott Gray、Alec Radford 和 Ilya Sutskever。利用稀疏变压器。arXiv 预印本 arXiv:1904.10509，2019 年。

[8] Yuntian Deng、Anton Bakhtin、Myle Ott、Arthur Szlam 和 Marc Aurelio Ranzato。基于残差能量的文本生成模型。arXiv 预印本 arXiv:2004.11714,2020 年。

[9] Laurent Dinh、David Krueger 和 Yoshua Bengio。NICE:非线性独立成分估计。 arXiv 预印本 arXiv:1410.8516, 2014。

[10] Laurent Dinh、Jascha Sohl-Dickstein 和 Samy Bengio。使用 Real NVP 进行密度估计。 arXiv 预印本 arXiv:1605.08803,2016 年。

[11] Yilun Du 和 Igor Mordatch。基于能量的模型的隐式生成与建模。载于《神经信息处理系统进展》,第 3603-3613 页,2019 年。

[12] Ruiqi Gao、Yang Lu、Junpei Zhou、Song-Chun Zhu 和 Ying Nian Wu。通过多网格建模和采样学习生成式卷积神经网络。载于《IEEE 计算机视觉与模式识别会议论文集》 ,第 9155–9164 页,2018 年。

[13] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, 和 Ying Nian Wu. 基于能量的模型的光流对比估计。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》 ,第 7518–7528 页,2020 年。

[14] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。载于《神经信息处理系统进展》,第 2672-2680 页,2014 年。

[15] Anirudh Goyal、Nan Rosemary Ke、Surya Ganguli 和 Yoshua Bengio。变分回溯:学习过渡算子作为随机循环网络。载于《神经信息处理系统进展》,第 4392-4402 页,2017 年。

[16] Will Grathwohl、Ricky TQ Chen、Jesse Bettencourt 和 David Duvenaud。FFJORD:可扩展可逆生成模型的自由形式连续动力学。2019 年国际学习表征会议。

[17] Will Grathwohl、Kuan-Chieh Wang、Joern-Henrik Jacobsen、David Duvenaud、Mohammad Norouzi 和Kevin Swersky。你的分类器本质上是一个基于能量的模型,你应该像对待一个分类器一样对待它。国际学习表征会议,2020 年。

[18] Karol Gregor、Frederic Besse、Danilo Jimenez Rezende、Ivo Danihelka 和 Daan Wierstra。《迈向概念压缩》。载于《神经信息处理系统进展》,第 3549-3557 页,2016 年。

[19] Prahladh Harsha、Rahul Jain、David McAllester 和 Jaikumar Radhakrishnan。相关性的通信复杂性。第二十二届 IEEE 计算复杂性年会 (CCC 07),第 10-23 页。IEEE,2007 年。

[20] Marton Havasi、Robert Peharz 和 José Miguel Hernández-Lobato。最小随机码学习:从压缩模型参数中恢复比特。2019年国际学习表征大会。

[21] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。通过双时间尺度更新规则训练的 GAN 收敛于局部纳什均衡。载于《神经信息处理系统进展》,第 6626-6637 页,2017 年。

[22] Irina Higgins、Loic Matthey、Arka Pal、Christopher Burgess、Xavier Glorot、Matthew Botvinick 、Shakir Mohamed 和 Alexander Lerchner。beta-VAE:使用约束变分框架学习基本视觉概念。2017 年国际学习表征会议。

[23] Jonathan Ho、Xi Chen、Aravind Srinivas、Yan Duan 和 Pieter Abbeel。Flow++:通过变分反量化和架构设计改进基于流的生成模型。2019 年国际机器学习大会。

[24] Sicong Huang、Alireza Makhzani、Yanshuai Cao 和 Roger Grosse。评估深度生成模型的有损压缩率。2020 年国际机器学习大会。

[25] Nal Kalchbrenner、Aaron van den Oord、Karen Simonyan、Ivo Danihelka、Oriol Vinyals、Alex Graves 和Koray Kavukcuoglu。视频像素网络。国际机器学习会议,第 1771-1779 页,2017 年。

[26] Nal Kalchbrenner、Erich Elsen、Karen Simonyan、Seb Noury、Norman Casagrande、Edward Lockhart、 Florian Stimberg、Aaron van den Oord、Sander Dieleman 和 Koray Kavukcuoglu。高效的神经音频合成。在国际机器学习会议上,第 2410-2419 页,2018 年。

[27] Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。《逐步发展 GAN,以提升质量、稳定性和多样性》。2018 年国际学习表征大会。

[28] Tero Karras、Samuli Laine 和 Timo Aila。一种基于风格的生成对抗网络生成器架构。《IEEE 计算机视觉与模式识别会议论文集》,第 179-185 页。

4401–4410,2019年。

[29] Tero Karras、Miika Aittala、Janne Hellsten、Samuli Laine、Jaakko Lehtinen 和 Timo Aila。使用有限的数据训练生成对抗网络。 arXiv 预印本 arXiv:2006.06676v1,2020 年。

[30] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。分析并改进 StyleGAN 的图像质量。在 IEEE/CVF 计算机视觉与模式识别会议论文集,第 8110-8119 页,2020 年。

[31] Diederik P Kingma 和 Jimmy Ba. Adam：一种随机优化方法。2015年国际学习表征会议。

[32] Diederik P Kingma 和 Prafulla Dhariwal。Glow：基于可逆 1x1 卷积的生成流。在神经信息处理系统进展,第 10215-10224 页,2018 年。

[33] Diederik P Kingma 和 Max Welling。自编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114，2013年。

[34] Diederik P Kingma、Tim Salimans、Rafal Jozefowicz、Xi Chen、Ilya Sutskever 和 Max Welling。利用逆自回归流改进变分推理。载于《神经信息处理系统进展》,第 4743-4751 页,2016 年。

[35] John Lawson、George Tucker、Bo Dai 和 Rajesh Ranganath。能量启发模型：利用采样器诱导分布进行学习。载于《神经信息处理系统进展》,第 8501-8513 页，2019 年。

[36] Daniel Levy、Matt D. Hoffman 和 Jascha Sohl-Dickstein。利用神经网络推广哈密顿蒙特卡洛方法。2018 年国际学习表征会议。

[37] Lars Maaløe、Marco Fraccaro、Valentin Liévin 和 Ole Winther。BIVA：用于生成模型的超深层潜在变量层次结构。载于《神经信息处理系统进展》,第 6548-6558 页,2019 年。

[38] Jacob Menick 和 Nal Kalchbrenner。利用子尺度像素网络和多维上采样生成高保真图像。2019 年国际学习表征会议。

[39] Takeru Miyato、Toshiki Kataoka、Masanori Koyama 和 Yuichi Yoshida。生成对抗网络的谱归一化。2018 年国际学习表征会议。

[40] Alex Nichol。VQ-DRAW：一种顺序离散 VAE。arXiv 预印本 arXiv:2003.01599,2020 年。

[41] Erik Nijkamp、Mitch Hill、Tian Han、Song-Chun Zhu 和 Ying Nian Wu。基于 MCMC 的能量模型最大似然学习剖析。arXiv 预印本 arXiv:1903.12370,2019 年。

[42] Erik Nijkamp、Mitch Hill、Song-Chun Zhu 和 Ying Nian Wu。学习面向能量基模型的非收敛非持久短期 MCMC 算法。载于《神经信息处理系统进展》,第 5233-5243 页,2019 年。

[43] Georg Ostrovski、Will Dabney 和 Remi Munos。用于生成建模的自回归分位数网络。在国际机器学习会议上,第 3936-3945 页,2018 年。

[44] Ryan Prenger、Rafael Valle 和 Bryan Catanzaro。WaveGlow：基于流的语音合成生成网络。载于 ICASSP 2019-2019 IEEE 国际声学、语音和信号处理会议（ICASSP）,第 3617-3621 页。IEEE,2019 年。

[45] Ali Razavi、Aaron van den Oord 和 Oriol Vinyals。使用 VQ-VAE-2 生成多样化高保真图像。载于《神经信息处理系统进展》,第 14837–14847 页,2019 年。

[46] Danilo Rezende 和 Shakir Mohamed。基于正则化流的变分推断。国际机器学习会议,第 1530-1538 页,2015 年。

[47] Danilo Jimenez Rezende、Shakir Mohamed 和 Daan Wierstra。深度生成模型中的随机反向传播和近似推理。国际机器学习会议,第 1278-1286 页,2014 年。

[48] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。国际医学图像计算与计算机辅助干预会议,第 234-241 页。Springer,2015 年。

[49] Tim Salimans 和 Durk P Kingma。权重归一化：一种简单的重新参数化方法,可加速深度神经网络的训练。载于《神经信息处理系统进展》,第 901-909 页， 2016 年。

[50] Tim Salimans、Diederik Kingma 和 Max Welling。马尔可夫链蒙特卡罗与变分推理：弥合差距。国际机器学习会议,第 1218-1226 页,2015 年。

[51] Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford 和 Xi Chen。改进的GAN 训练技术。载于《神经信息处理系统进展》，第 2234-2242 页， 2016 年。

[52] Tim Salimans、Andrej Karpathy、Xi Chen 和 Diederik P Kingma。PixelCNN++：通过离散化逻辑混合似然及其他修改改进 PixelCNN。2017年国际学习表征会议。

[53] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。基于非平衡热力学的深度无监督学习。国际机器学习会议，第 2256-2265 页，2015 年。

[54] Jiaming Song, Shengjia Zhao, 和 Stefano Ermon. A-NICE-MC：MCMC 的对抗性训练。载于《神经信息处理系统进展》，第 5140–5150 页，2017 年。

[55] Yang Song 和 Stefano Ermon。通过估计数据分布梯度进行生成建模。在神经信息处理系统进展，第 11895-11907 页，2019 年。

[56] Yang Song 和 Stefano Ermon。改进的基于分数的生成模型训练技术。arXiv 预印本 arXiv:2006.09011，2020 年。

[57] Aaron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior 和 Koray Kavukcuoglu。 WaveNet：原始音频的生成模型。 arXiv 预印本 arXiv:1609.03499, 2016。

[58] Aaron van den Oord、Nal Kalchbrenner 和 Koray Kavukcuoglu。像素循环神经网络。2016 年国际机器学习会议。

[59] Aaron van den Oord、Nal Kalchbrenner、Oriol Vinyals、Lasse Espeholt、Alex Graves 和 Koray Kavukcuoglu。《使用 PixelCNN 解码器进行条件图像生成》。载于《神经信息处理系统进展》 ，第 4790-4798 页，2016 年。

[60] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jacob Uszkoreit、Llion Jones、Aidan N Gomez、 ukasz Kaiser 和 Illia Polosukhin。你所需要的只是关注。在神经信息处理系统进展中，第 5998-6008 页，2017 年。

[61] Pascal Vincent. 分数匹配与去噪自编码器之间的联系. 神经计算，23(7):1661–1674, 2011年。

[62] Sheng-Yu Wang、Oliver Wang、Richard Zhang、Andrew Owens 和 Alexei A Efros。CNN 生成的图像目前出奇地容易识别……。载于 2020 年 IEEE 计算机视觉与模式识别会议论文集。

[63] Xiaolong Wang, Ross Girshick, Abhinav Gupta, 和 Kaiming He. 非局部神经网络。载于《IEEE 计算机视觉与模式识别会议论文集》，第 7794-7803 页， 2018 年。

[64]奥克·J·威格斯和埃米尔·胡格博姆。使用预测自回归模型进行预测抽样。arXiv 预印本 arXiv:2002.09928, 2020。

[65] Hao Wu, Jonas Köhler 和 Frank Noé。随机正则化流。arXiv 预印本 arXiv:2002.06707，2020年。

[66]吴宇欣和何开明，群规范化，欧洲计算机应用会议论文集Vision（ECCV），第 3-19 页，2018 年。

[67] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In International 机器学习会议，第 2635-2644 页，2016 年。

[68]谢建文，朱松春，吴英年。通过时空生成卷积网络合成动态模式。载于《IEEE 计算机视觉与模式识别会议论文集》，第 7093-7101 页，2017 年。

[69]谢建文，郑子龙，高瑞琪，王文冠，朱松春，吴英年。用于三维形状合成与分析的学习描述符网络。载于《IEEE计算机视觉与模式识别会议论文集》 ，第 8629–8638 页，2018 年。

[70]谢建文，朱松春，吴英年。学习基于能量的时空生成卷积网络以实现动态模式。《IEEE 模式分析与机器智能学报》，2019 年。

[71] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 和 Jianxiong Xiao. LSUN：利用深度学习与人类参与构建大规模图像数据集. arXiv 预印本 arXiv:1506.03365, 2015 年。

[72]谢尔盖·扎戈鲁伊科和尼科斯·科莫达基斯。宽残差网络。 arXiv 预印本 arXiv:1605.07146, 2016年。

额外信息

表 3 中包含了 LSUN 数据集的LSUN FID 分数。标记为 StyleGAN2 的分数为基线,其他分数由各自的作者报告。 被报道

表 3:LSUN 256 × 256 数据集的 FID 分数

模型	LSUN 卧室	LSUN 教堂	LSUN 猫
ProgressiveGAN[27]	8.34	6.42	37.52
StyleGAN[28]	2.65	4.21	8.53
StyleGAN2 [30]	-	3.86	6.93
我们的（Lsimple）	6.36	7.89	19.75
我们的（简单,大）	4.90	-	-

渐进式压缩我们在第 4.3 节中提出的有损压缩论证只是一个概念证明，
因为算法 3 和算法 4 依赖于诸如最小随机编码[20]之类的过程,而该过程
对于高维数据来说,这些算法是压缩解释
Sohl-Dickstein 等人 [53] 的变分界限（5)尚未作为实用的压缩系统。

表 4:无条件 CIFAR10 测试集率失真值（附图 5）

逆过程时间（T − t + 1）	速率（位/维）	失真度（RMSE [0, 255]）
1000	1.77581	0.95136
900	0.11994	12.02277
800	0.05415	18.47482
700	0.02866	24.43656
600	0.01507	30.80948
500	0.00716	38.03236
400	0.00282	46.12765
300	0.00081	54.18826
200	0.00013	60.97170
100	0.00000	67.60125

A 扩展推导

下面是方程 (5) 的推导,即扩散模型的约化方差变分界限。这
材料来自Sohl-Dickstein等人[53];我们将其包含在这里只是为了完整性。

$$L = \text{等式} - \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \tag{17}$$

$$= \text{等式} \quad \text{对数} p(x_T) - \sum_{t \geqslant 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \tag{18}$$

$$= \text{等式} - \log p(x_T) - \sum_{t>1} \text{对数} \frac{p_\theta(x_{t-1}|x_t)}{\text{对数} q(x_t|x_{t-1})} \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \tag{19}$$

$$= \text{等式} - \log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \tag{20}$$

$$= \text{等式} - \log \frac{p(x_T)}{q(x_T|x_0)} - \sum_{t>1} \text{对数} \frac{p_\theta(x_{t-1}|x_t)}{\text{对数} q(x_{t-1}|x_t, x_0)} \text{对数} p_\theta(x_0|x_1) \tag{21}$$

$$=\text{等式}\, D_{KL}(q(x_T|x_0)\,p(x_T)) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_t, x_0)\,p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \tag{22}$$

以下是L 的替代版本。它不易估计,但对我们在第 4.3 节的讨论很有用。

$$L = \text{等式}\, \text{对数} p(x_T) - \sum_{t\geqslant 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \tag{23}$$

$$=\text{等式}\, \text{对数} p(x_T) - \sum_{t\geqslant 1} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} \cdot \log \frac{q(x_{t-1})}{q(x_t)} \tag{24}$$

$$=\text{等式}\, \text{对数} \frac{p(x_T)}{q(x_T)} - \sum_{t\geqslant 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_0)\,q(x_{t-1}|x_t)} \tag{25}$$

$$= D_{KL}(q(x_T)\,p(x_T)) + \text{等式} \sum_{t\geqslant 1} D_{KL}(q(x_{t-1}|x_t)\,p_\theta(x_{t-1}|x_t)) + H(x_0) \tag{26}$$

B 实验细节

我们的神经网络架构以 PixelCNN++ [52]为骨干,这是一个基于 Wide ResNet [72] 的U-Net [48] 。我们将权重正则化[49]替换为组正则化[66],以简化实现。我们的32×32模型使用四种特征图分辨率（32×32到4× 4）， 256×256模型使用六种。所有模型每个分辨率级别都有两个卷积残差块,在卷积块之间有16×16分辨率的自注意力块[6]。扩散时间t通过将 Transformer 正弦位置嵌入[60]添加到每个残差块来指定。我们的 CIFAR10 模型有 3570 万个参数,LSUN 和 CelebA-HQ 模型有 1.14 亿个参数。我们还通过增加滤波器数量,训练了一个更大的 LSUN Bedroom 模型变体,该模型具有约 2.56 亿个参数。

我们在所有实验中均使用了 TPU v3-8（相当于 8 个 V100 GPU）。我们的 CIFAR 模型以每秒 21 步的速度训练,批量大小为 128（以 80 万步的速度训练完成需要 10.6 小时）,采样一批 256 张图片需要 17 秒。我们的 CelebA-HQ/LSUN (2562) 模型以每秒 2.2 步的速度训练,批量大小为 64,采样一批 128 张图片需要 300 秒。我们在CelebA-HQ 上训练了 0.5 M 步,在 LSUN Bedroom 上训练了 2.4 M 步,在 LSUN Cat 上训练了 1.8 M 步,在 LSUN Church 上训练了 1.2 M 步。更大的 LSUN Bedroom 模型训练了 1.15 M 步。

除了早期选择超参数以使网络大小适应内存限制之外,我们还执行了大部分超参数搜索以优化 CIFAR10 样本质量,然后将结果设置转移到其他数据集：

- 我们从一组常数、线性和二次调度方案中选择了βt调度方案,所有方案均受到约束,使得LT≈0 。我们设置T = 1000而不进行扫描,并选择从β1 = 10−4到βT = 0.02的线性调度方案。·我们通过扫描值{0.1,0.2,0.3,0.4}将CIFAR10上的dropout率设置为0.1 。

  在 CIFAR10 数据集上,未使用 dropout 时,我们得到的样本质量较差,类似于未正则化的 PixelCNN++ [52] 中的过拟合伪影。我们将其他数据集上的 dropout 率设置为零,未进行扫描。

- 我们在 CIFAR10 数据集的训练过程中使用了随机水平翻转;我们尝试了使用翻转和不使用翻转两种训练方式,发现翻转可以略微提升样本质量。除 LSUN Bedroom 数据集外,我们对所有其他数据集也使用了随机水平翻转。
- 我们在实验早期尝试了 Adam [31]和 RMSProp,最终选择了Adam。我们将超参数保留为标准值。我们将学习率设置为2 × 10−4 （不进行任何扫描）,对于256 × 256 的图像,我们将学习率降低到2 × 10−5 ,因为较大的学习率似乎会导致训练不稳定。

·我们将 CIFAR10 的批次大小设置为 128,将较大图像的批次大小设置为 64。我们没有扫描
这些价值观。

·我们对衰减因子为 0.9999 的模型参数使用了 EMA。我们没有扫描
这个值。

最终实验仅训练一次,并在整个训练过程中评估样本质量。样本质量得分和对数似然值以训练过程中的最小 FID 值作为报告依据。

在 CIFAR10 数据集上,我们分别使用 OpenAI [51]和 TTUR [21]代码库中的原始代码,计算了 50,000 个样本的 Inception 和 FID 得分。在 LSUN 数据集上,我们使用 StyleGAN2 [30]代码库中的代码,计算了 50,000 个样本的 FID 得分。CIFAR10 和 CelebA-HQ 数据集均由 TensorFlow 数据集(https://www.tensorflow.org/datasets)加载。 LSUN 是使用 StyleGAN 的代码编写的。数据集分割(或不分割)是引入生成式建模应用的论文中的标准做法。所有详细信息均可在源代码发布中找到。

C 相关工作讨论

我们的模型架构、前向过程定义和先验与 NCSN [55, 56]有一些微妙但重要的不同,这些不同提升了样本质量。值得注意的是,我们直接将采样器训练为潜变量模型,而不是在事后训练后添加。更详细地说:

1.我们使用带有自注意力机制的 U-Net;NCSN 使用带有空洞卷积的 RefineNet。我们通过添加 Transformer 正弦位置嵌入来调节t上的所有层,而不仅仅是在规范化层(NCSNv1)或仅在输出层(v2)进行调节。

2.扩散模型在每次正向处理过程中都会按比例缩小数据(按$\sqrt{1-\beta_t}$因子缩小),这样在添加噪声时方差就不会增大,从而为神经网络的反向过程提供一致缩放的输入。NCSN 省略了这个缩放因子。

3.与 NCSN 不同,我们的前向过程破坏了信号($D_{KL}(q(x_T|x_0)\,\|\,\mathcal{N}(0, I)) \approx 0$),确保$x_T$的先验和聚合后验紧密匹配。
与 NCSN 不同的是,我们的β_t非常小,这确保了正向过程可以通过具有条件高斯分布的马尔可夫链进行逆向。这两个因素都能够防止采样时出现分布偏移。

4.我们的类朗之万采样器在前向过程中,其系数(学习率、噪声尺度等)严格地从β_t导出。因此,我们的训练程序会在T步后直接训练采样器以匹配数据分布:它使用变分推断将采样器训练为潜变量模型。相比之下,NCSN 的采样器系数是事后手动设置的,并且其训练程序无法保证直接优化采样器的质量指标。

D 样本

附加样本图 11、13、16、17、18 和 19 展示了在 CelebA-HQ、CIFAR10 和 LSUN 数据集上训练的扩散模型的未经整理的样本。

潜在结构和逆过程随机性在采样过程中,先验x_T
$\mathcal{N}(0, I)$和 Langevin 动力学是随机的。为了理解第二个噪声源的重要性,我们对 CelebA 256 × 256数据集采样了多幅以相同中间潜在向量为条件的图像。图 7 显示了从逆过程$x_0 \sim p_\theta(x_0|x_t)$中抽取的多个样本,它们共享 $t \in \{1000, 750, 500, 250\}$ 的潜在向量x_t。为了实现这一点,我们从先前的初始抽取中运行一个反向链。在中间时间步,链被拆分以采样多幅图像。当链在x_T =1000 处的先前抽取之后拆分时,样本差异很大。

然而,当经过更多步骤后,链条被拆分时,样本会共享一些高级属性,例如性别、发色、眼镜、饱和度、姿势和面部表情。这表明,像x_{750}这样的中间潜在变量编码了这些属性,尽管它们难以察觉。

从粗到细的插值图 9 显示了当我们改变潜在空间插值之前的扩散步骤数时,一对源 CelebA 256 × 256图像之间的插值。

增加扩散步骤的数量会破坏源图像中的更多结构,

模型在逆过程中完成。这使我们能够以细粒度和粗粒度进行插值。在扩散步骤为0的极限情况下,插值会在像素空间中混合源图像。另一方面,经过1000 个扩散步骤后,源信息会丢失,插值结果为新的样本。

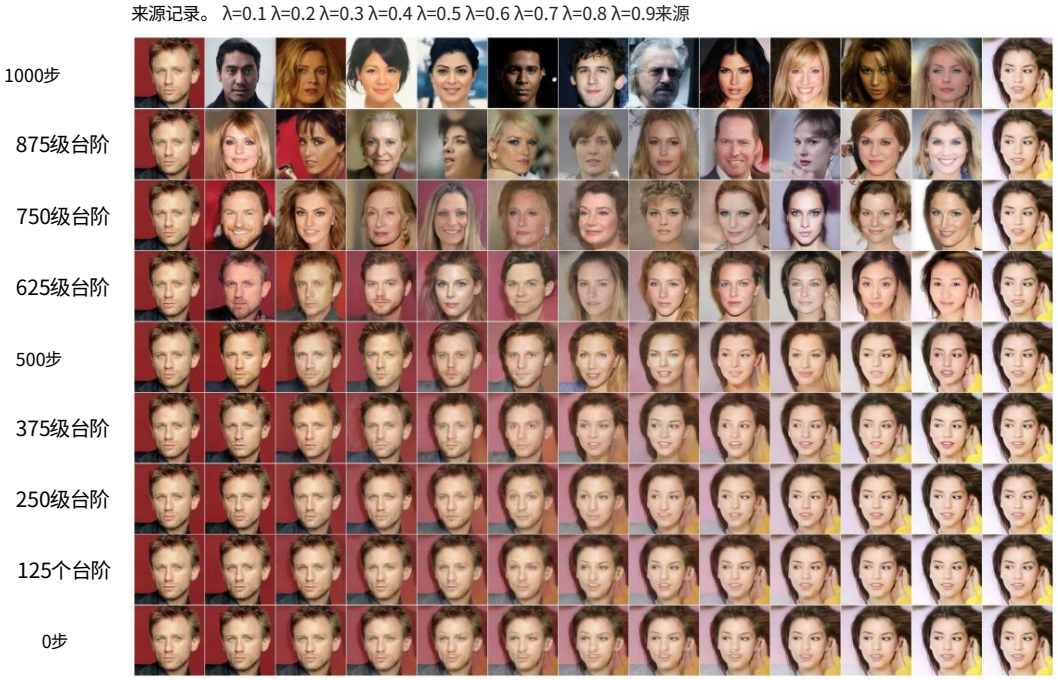


图 9:由粗到细的插值,改变潜在混合之前的扩散步骤数。

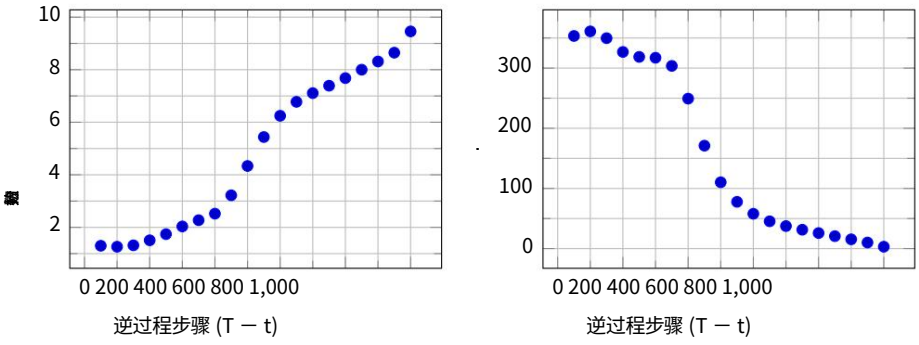
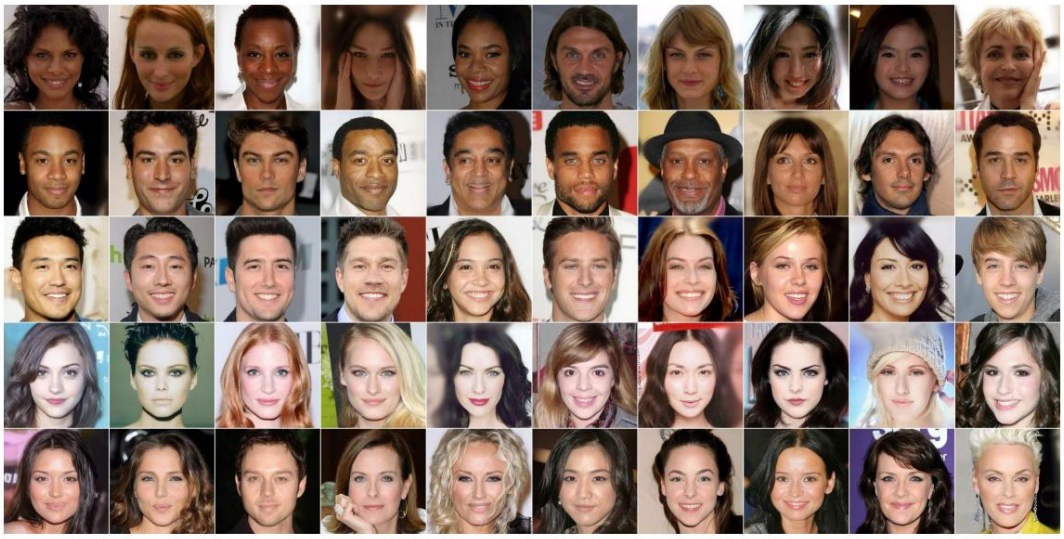


图 10:无条件 CIFAR10 随时间渐进采样的质量

图 11:CelebA-HQ 256×256 生成的样本



(a)像素空间最近邻

(b)Inception 特征空间最近邻

图 12:CelebA-HQ 256 × 256最近邻,基于人脸周围100 × 100 的裁剪图像计算得出。生成的样本位于最左列，训练集最近邻位于其余列。

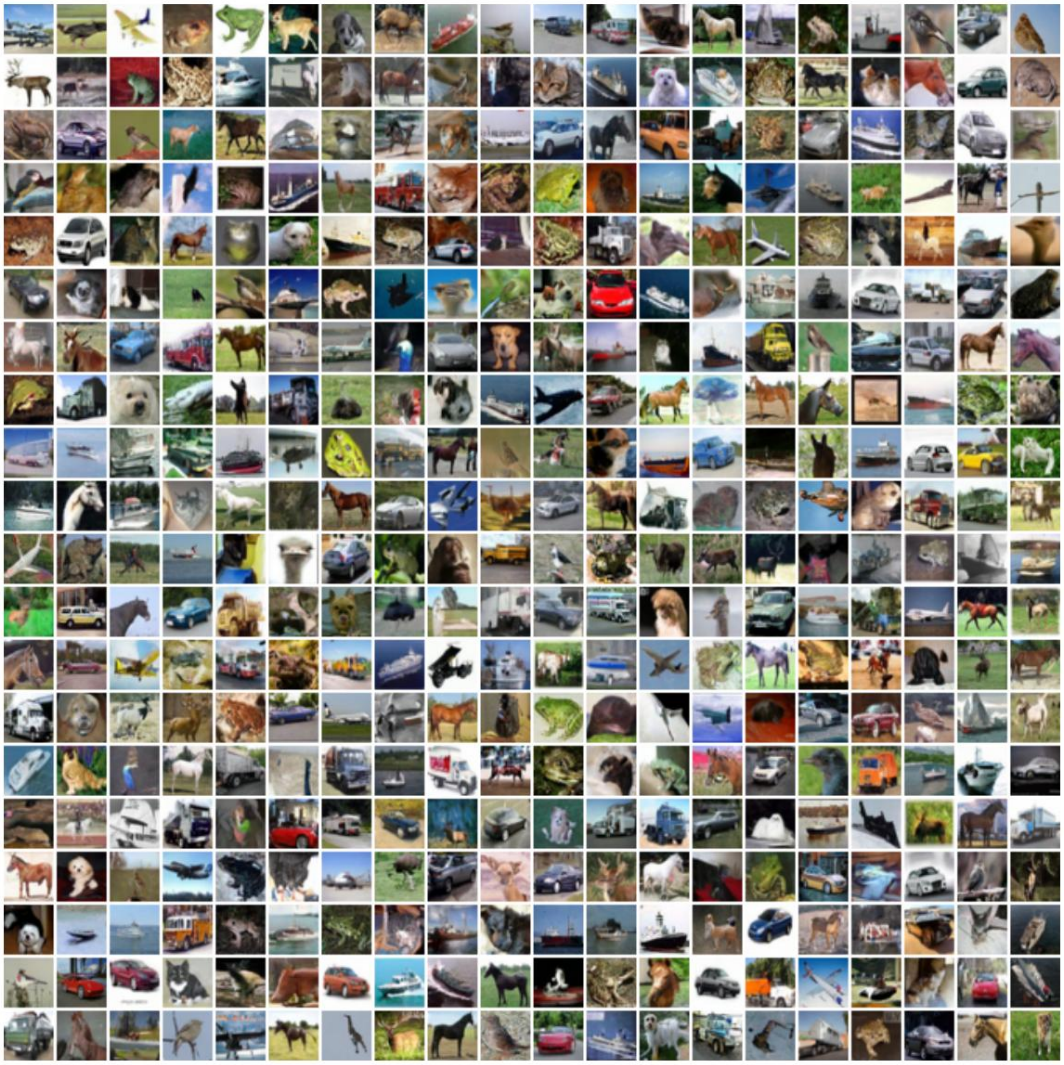


图 13:无条件 CIFAR10 生成的样本

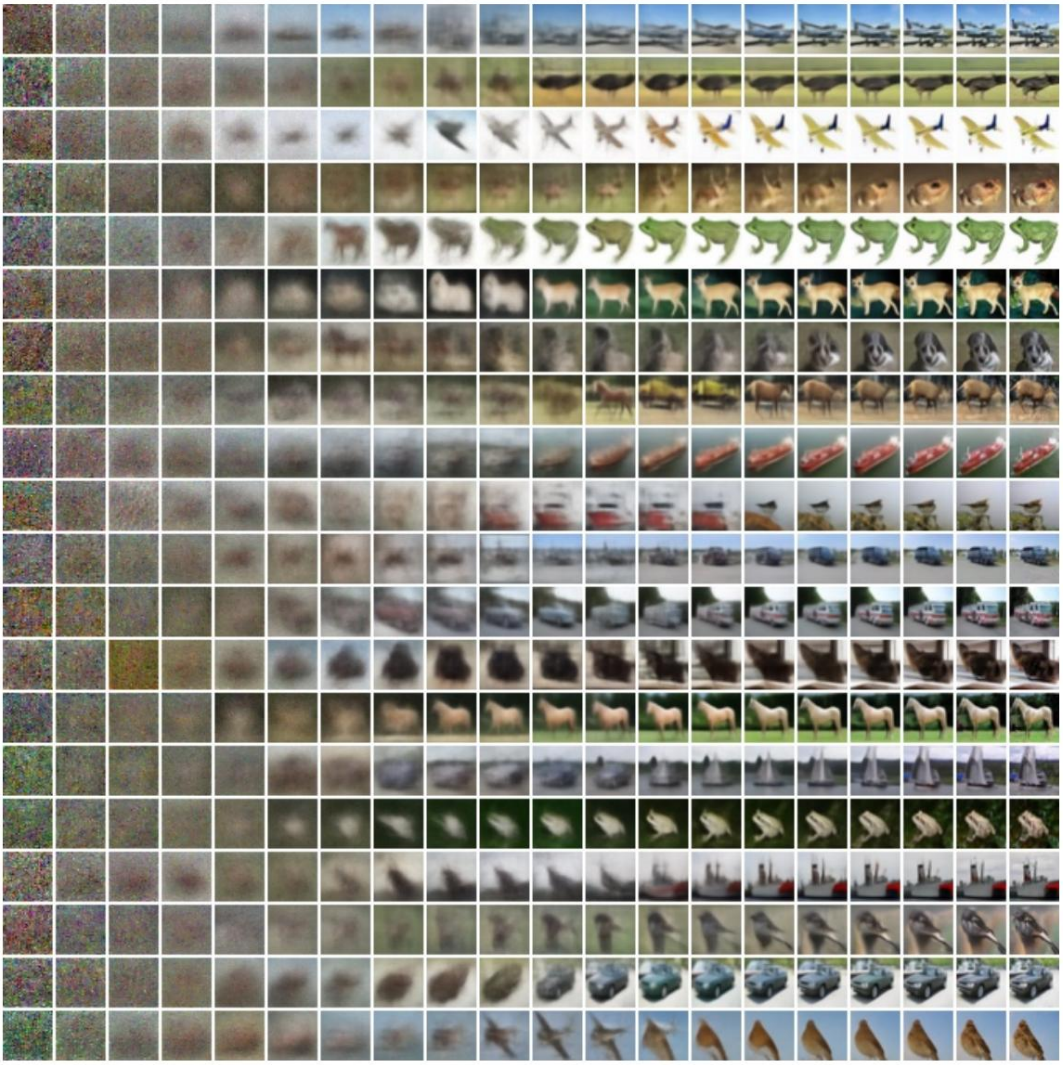
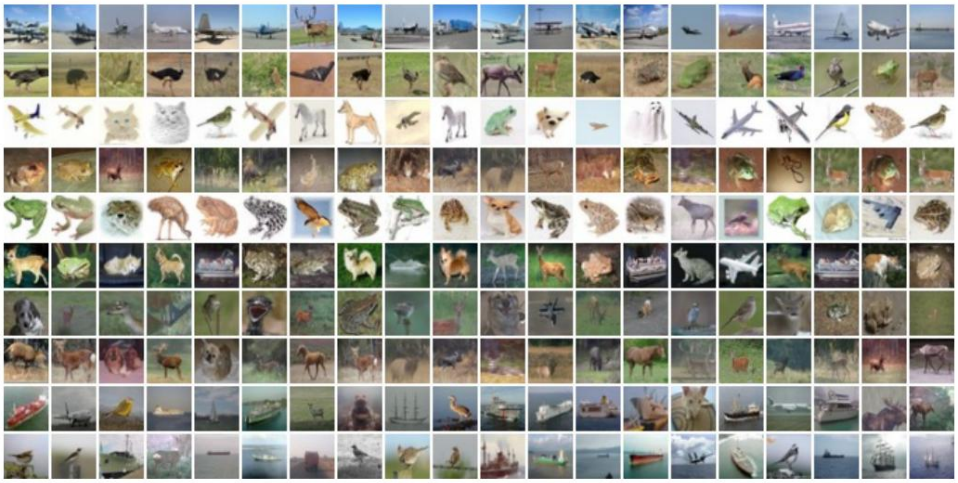


图 14:无条件 CIFAR10 渐进式生成



(a)像素空间最近邻

(b)Inception 特征空间最近邻

图 15:无条件 CIFAR10 最近邻。生成的样本位于最左侧列,训练集最近邻位于其余列。

图 16:LSUN 教会生成的样本。FID=7.89

图 17:LSUN 卧室生成的样本,大型模型。FID=4.90

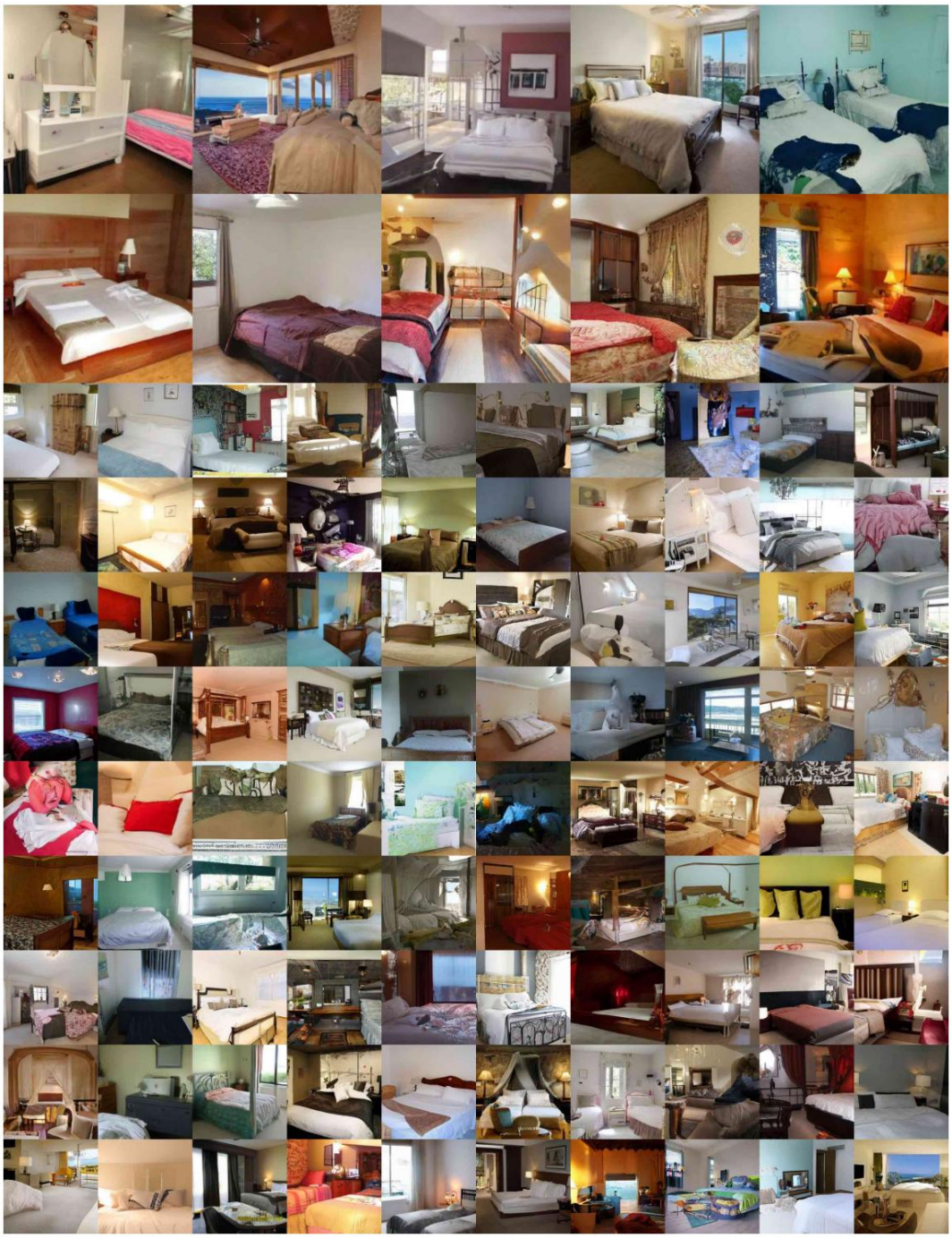


图 18:LSUN 卧室生成的样本,小模型。FID=6.36

图19:LSUN Cat 生成的样品。FID=19.75