去噪扩散概率模型

Jonathan Ho
加州大学伯克利分校

Ajay Jain
加州大学伯克利分校

Pieter Abbeel
加州大学伯克利分校

摘要

我们使用扩散概率模型呈现高质量的图像合成结果，扩散概率模型是一类受非平衡热力学考虑因素启发的潜在变量模型。我们的最佳结果是通过根据扩散概率模型与去噪分数与 Langevin 动力学匹配之间的新联系设计的加权变分界进行训练获得的，并且我们的模型自然地包含了渐进式有损解压缩方案，该方案可以解释为自回归解码的推广。在无条件 CIFAR10 数据集上，我们获得了 9.46 的 Inception 分数和 3.17 的最先进的 FID 分数。在 256x256 LSUN 上，我们获得了类似于 ProgressiveGAN 的样本质量。我们的实现可在 https://github.com/hojonathanho/diffusion 获得。

介绍

各种深度生成模型最近在各种数据模式中展示了高质量的样本。生成对抗网络 (GAN)、自回归模型、流和变分自动编码器 (VAE) 已经合成了引人注目的图像和音频样本 [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取得了显着进步，产生的图像可与 GAN 的图像相媲美 [11, 55]。

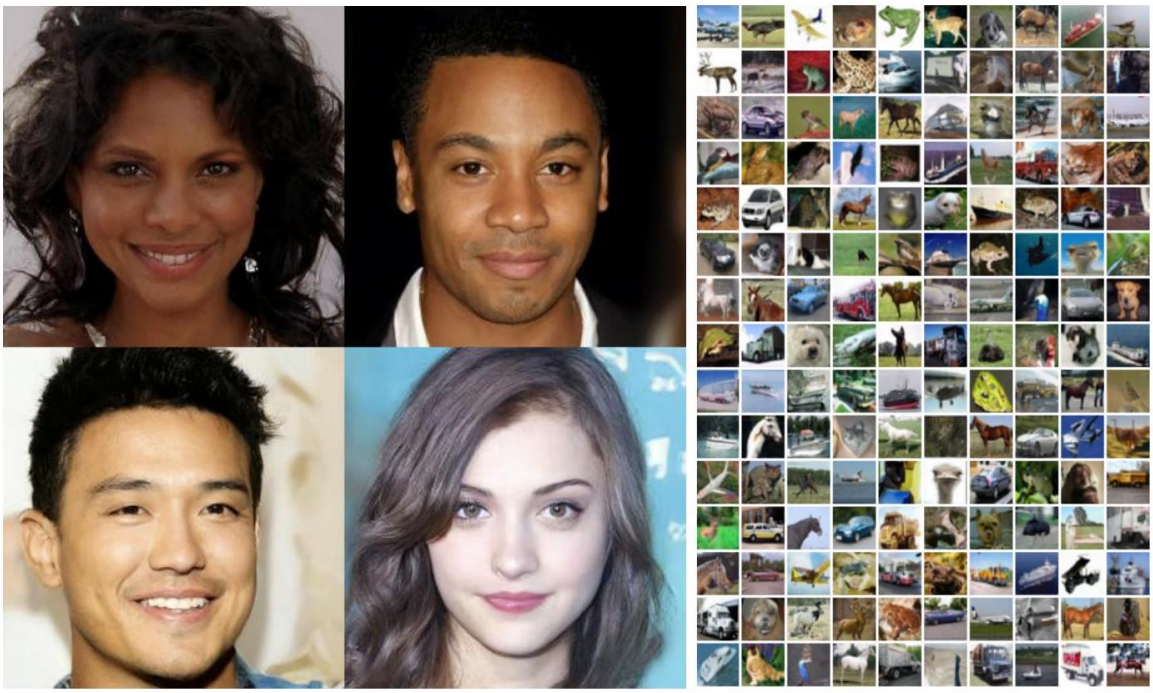

图 1：在 CelebA-HQ 256×256（左）和无条件 CIFAR10（右）上生成的样本

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

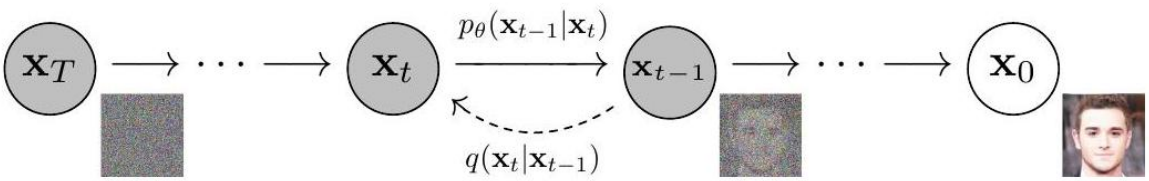

图 2：这项工作中考虑的有向图模型。

本文介绍了扩散概率模型的进展 [53]。扩散概率模型（为简洁起见，我们将其称为“扩散模型”）是使用变分推理训练的参数化马尔可夫链，以在有限时间后生成与数据匹配的样本。该链的转换被学习来逆转扩散过程，这是一个马尔可夫链，逐渐向相反采样方向的数据添加噪声，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就足够了，从而允许特别简单的神经网络参数化。

扩散模型易于定义且训练高效，但据我们所知，还没有证据表明它们能够生成高质量的样本。我们表明扩散模型实际上能够生成高质量的样本，有时比其他类型的生成模型的已发表结果更好（第 4p 节）。此外，我们表明扩散模型的特定参数化揭示了与去噪分数匹配的等价性训练期间的多个噪声水平和采样期间的退火 Langevin 动力学（第 3.2 节）[55, 61]。我们使用此参数化（第 4.2 节）获得了最佳样本质量结果，因此我们认为这种等效性是我们的主要贡献之一。

尽管它们的样本质量很高，但与其他基于似然的模型相比，我们的模型没有具有竞争力的对数似然（但是，我们的模型确实具有比据报道为基于能量的模型和分数匹配产生的退火重要性抽样的大估计更好的对数似然[11, 55]）。我们发现我们模型的大部分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种现象进行了更精细的分析，我们表明扩散模型的采样过程是一种渐进解码，类似于沿着位排序的自回归解码，它极大地概括了自回归模型通常可能发生的情况.

背景

扩散模型 [53] 是形式为 $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$ 的潜在变量模型，其中 $\mathbf{x}_1, \ldots, \mathbf{x}_T$ 是与数据 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 具有相同维数的潜在变量。联合分布 $p_\theta(\mathbf{x}_{0:T})$ 称为逆向过程，它被定义为具有从 $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ 开始的学习高斯转换的马尔可夫链：

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

扩散模型与其他类型的潜变量模型的区别在于，近似后验分布 $\mathrm{q(x1:T\ |x0)}$，称为正向过程或扩散过程，被固定为一个马尔可夫链，根据方差调度$\beta 1, \ldots, \beta \mathrm{T}$逐渐向数据添加高斯噪声。

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \quad q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$$

训练是通过优化负对数似然的通常变分界限来进行的：

$$\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] =: L$$

正向过程方差 β_t 可以通过重新参数化 [33] 学习或作为超参数保持不变，并且通过 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 中选择高斯条件来部分确保反向过程的表现力，因为这两个过程具有相同的函数形式当 β_t 很小时 [53]。前向过程的一个显着特性是它允许在任意时间步 $\mathbf{x}_t$ 以封闭形式对 t 进行采样：使用符号 $\alpha_t := 1 - \beta_t$ 和 $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$，我们有

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

因此，通过使用随机梯度下降优化 L 的随机项，可以进行有效的训练。通过将 $L(3)$ 重写为：

$$\mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \| p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}_{L_0}]$$

（有关详细信息，请参阅附录 A。术语上的标签在第 3 节中使用） 方程式（5p 使用 KL 散度直接比较 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 与前向过程后验，在 $\mathbf{x}_0$ 条件下易于处理：

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

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

因此，方程式中的所有 KL 分歧(5) 是高斯分布之间的比较，因此可以使用封闭形式表达式而不是高方差蒙特卡罗估计以 Rao-Blackwellized 方式计算它们。

扩散模型和去噪自动编码器

扩散模型可能看起来是一类受限制的潜在变量模型，但它们在实施中允许大量的自由度。必须选择正向过程的方差 β_t 和反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（第 3.2 节）之间建立了新的显式联系，从而为扩散模型（第 3.4 节）提供了一个简化的加权变分边界目标。最终，我们的模型设计被简单性和实证结果证明是合理的（第 44 节）。我们的讨论按方程式的术语分类(5).

前向过程和 L_T

我们忽略了正向过程方差 β_t 可以通过重新参数化学习这一事实，而是将它们固定为常量（详见第 4 节）。因此，在我们的实现中，近似后验 q 没有可学习的参数，因此 L_T 在训练期间是一个常数，可以忽略。

逆向过程和 $L_{1:T-1}$

现在我们讨论我们在 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$ 中对 $1 < t \leq T$ 的选择。首先，我们将 $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$ 设置为未经训练的时间相关常数。实验上， $\sigma_t^2 = \beta_t$ 和 $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ 都有相似的结果。第一个选择是 $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 的最佳选择，第二个选择是确定性地设置为一个点的 $\mathbf{x}_0$ 的最佳选择。对于具有坐标单位方差的数据，这些是对应于逆向过程熵上限和下限的两个极端选择 [53]。

其次，为了表示平均值 $\boldsymbol{\mu}_\theta(\mathrm{x}_t, t)$，我们提出了一个特定的参数化，其动机是对 L_t 的以下分析。使用 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}\big)$，我们可以写：

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C$$

其中 C 是一个不依赖于 θ 的常数。因此，我们看到 $\boldsymbol{\mu}_\theta$ 最直接的参数化是预测 $\tilde{\boldsymbol{\mu}}_t$ 的模型，前向过程后验均值。但是，我们可以扩展方程式 (8) 进一步通过重新参数化Eq(4) 作为 $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ 的 $\boldsymbol{\epsilon}\mathcal{N}(\mathbf{0}, \mathbf{I})$ 并应用前向过程后验公式 (77 ：

$$
\begin{aligned}
L_{t-1} - C &= \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}) - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}\right)\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),t)\right\|^2\right] \\
&= \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),t)\right\|^2\right]
\end{aligned}
$$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1,\dots,T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
2: **for** $t = T,\dots,1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

等式 10 表明 $\boldsymbol{\mu}_\theta$ 必须在给定 $\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$ 的情况下预测 $\mathbf{x}_t$。由于 $\mathbf{x}_t$ 可用作模型的输入，我们可以选择参数化

$$
\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right)\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right)
$$

其中 ϵ_θ 是一个函数逼近器，用于根据 $\boldsymbol{\epsilon}$ 预测 $\mathbf{x}_t$。采样 $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 就是计算 $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right) + \sigma_t\mathbf{z}$，其中 $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$。完整的采样过程，算法 2，类似于 Langevin 动力学，其中 ϵ_θ 作为数据密度的学习梯度。此外，通过参数化111，Eq. 110 简化为：

$$
\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]
$$

这类似于 t[55] 索引的多个噪声尺度上的去噪分数匹配。作为方程式[12] 等于 Langevin 类逆过程 (11) 的变分界（的一项），我们看到优化类似于去噪分数匹配的目标等同于使用变分推理来拟合有限时间边际类似于 Langevin 动力学的采样链。

总而言之，我们可以训练反向过程均值函数逼近器 $\boldsymbol{\mu}_\theta$ 来预测 $\tilde{\boldsymbol{\mu}}_t$，或者通过修改其参数化，我们可以训练它来预测 ϵ（也有预测 $\mathbf{x}_0$ 的可能性，但我们发现这会导致我们实验早期的样本质量变差。）我们已经证明 ϵ-预测参数化既类似于 Langevin 动力学，又简化了扩散模型的变分界到类似于去噪分数匹配的目标。尽管如此，它只是 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 的另一个参数化，因此我们在第 4 节中的消融中验证了它的有效性，我们将预测 $\boldsymbol{\epsilon}$ 与预测 $\tilde{\boldsymbol{\mu}}_t$ 进行比较。

数据缩放、逆向处理解码器和 L_0

我们假设图像数据由 $\{0,1,\dots,255\}$ 中的整数组成，线性缩放到 $[-1,1]$。这确保了神经网络反向过程从标准正常先验 $p(\mathbf{x}_T)$ 开始对一致缩放的输入进行操作。为了获得离散对数似然，我们将逆向过程的最后一项设置为从高斯 $\mathcal{N}\left(\mathbf{x}_0;\boldsymbol{\mu}_\theta(\mathbf{x}_1,1),\sigma_1^2\mathbf{I}\right)$ 派生的独立离散解码器：

$$
\begin{aligned}
p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) &= \prod_{i=1}^{D}\int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}\left(x;\mu_\theta^i(\mathbf{x}_1,1),\sigma_1^2\right)dx \\
\delta_+(x) &= \begin{cases}\infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1\end{cases} \quad \delta_-(x) = \begin{cases}-\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1\end{cases}
\end{aligned}
$$

其中 D是数据维度， i上标表示提取一个坐标（直接合并一个更强大的解码器，如条件自回归模型会很简单，但我们将其留给未来的工作。）类似于 VAE 解码器和自回归模型 [34、52] 中使用的离散连

续分布，我们在这里的选择确保变分界是离散数据的无损码长，不需要向数据添加噪声或将缩放操作的雅可比行列式合并到对数似然中。在采样结束时，我们无声地显示 $\boldsymbol{\mu}_\theta(\mathbf{x}_1, 1)$。

简化训练目标

通过上面定义的反向过程和解码器，变分界由从方程式导出的项组成(12) 和 (13)，相对于 θ 是明显可微的，并且可以用于训练。然而，我们发现在变分界的以下变体上进行训练有利于样本质量（并且更易于实现）：

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
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	
Ours (L, fixed isotropic $\boldsymbol{\Sigma}$)	7.67±0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46±0.11	**3.17**	≤ 3.75 (3.72)

Table 2: Unconditional CIFAR10 reverse process parameterization and training objective ablation. Blank entries were unstable to train and generated poor samples with out-of-range scores.

Objective	IS	FID
$\tilde{\boldsymbol{\mu}}$ prediction (baseline)		
L, learned diagonal $\boldsymbol{\Sigma}$	7.28±0.10	23.69
L, fixed isotropic $\boldsymbol{\Sigma}$	8.06±0.09	13.22
$\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_\theta\|^2$	–	–
$\boldsymbol{\epsilon}$ prediction (ours)		
L, learned diagonal $\boldsymbol{\Sigma}$	–	–
L, fixed isotropic $\boldsymbol{\Sigma}$	7.67±0.13	13.51
$\|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_\theta\|^2$ (L_{simple})	**9.46±0.11**	**3.17**

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]$$

其中 t 在 1 和 T 之间是统一的。 $t = 1$ 情况对应于 L_0，其中离散解码器定义 (13 中的积分近似于高斯概率密度函数乘以 bin 宽度，忽略 σ_1^2 和边缘效应。 $t > 1$ 案例对应于方程式的未加权版本（12），类似于 NCSN 去噪分数匹配模型 [55] 使用的损失加权（ L_T 没有出现，因为前向过程方差 β_t 是固定的。）算法 1 显示了具有这个简化目标的完整训练过程。

由于我们的简化目标（14）丢弃了等式中的权重(12)，它是一个加权变分界，与标准变分界相比，它强调重建的不同方面 [18, 22]。特别是，我们在第 4 节中的扩散过程设置导致简化目标以降低对应于小 t 的权重损失项。这些术语训练网络去噪具有极少量噪声的数据，因此降低它们的权重是有益的，这样网络可以专注于更大 t 术语的更困难的去噪任务。我们将在我们的实验中看到，这种重新加权会带来更好的样本质量。

实验

我们为所有实验设置 $T = 1000$，以便采样期间所需的神经网络评估数量与之前的工作相匹配 [53、55]。我们将前向过程方差设置为从 $\beta_1 = 10^{-4}$ 到 $\beta_T = 0.02$ 线性增加的常数。这些常数被选择为相对于缩放到 $[-1, 1]$ 的数据较小，确保反向和正向过程具有大致相同的函数形式，同时保持 $\mathbf{x}_T$ 的信噪比尽可能小 ($L_T = D_{\text{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0)\|\mathcal{N}(\mathbf{0}, \mathbf{I})) \approx 10^{-5}$ | 在我们的实验中每个维度的位)。

为了表示逆向过程，我们使用类似于 unbackbone PixelCNN++ [52] 48] 的 U-Net masked，并在整个 [66] 中进行组归一化。参数跨时间共享，使用 Transformer 正弦位置 embedding [60] 指定给网络。我们在 16 × 16 特征图分辨率 [63、60] 上使用自注意力。详情见附录 B。

样本质量

表 1 显示了 CIFAR10 上的 Inception 分数、FID 分数和负对数似然（无损代码长度）。我们的 FID 得分为 3.17，我们的无条件模型比文献中的大多数模型（包括类条件模型）实现了更好的样本质量。我们的 FID 分数是根据训练集计算的，这是标准做法；当我们针对测试集计算它时，得分为 5.24 ，这仍然优于文献中的许多训练集 FID 分数。

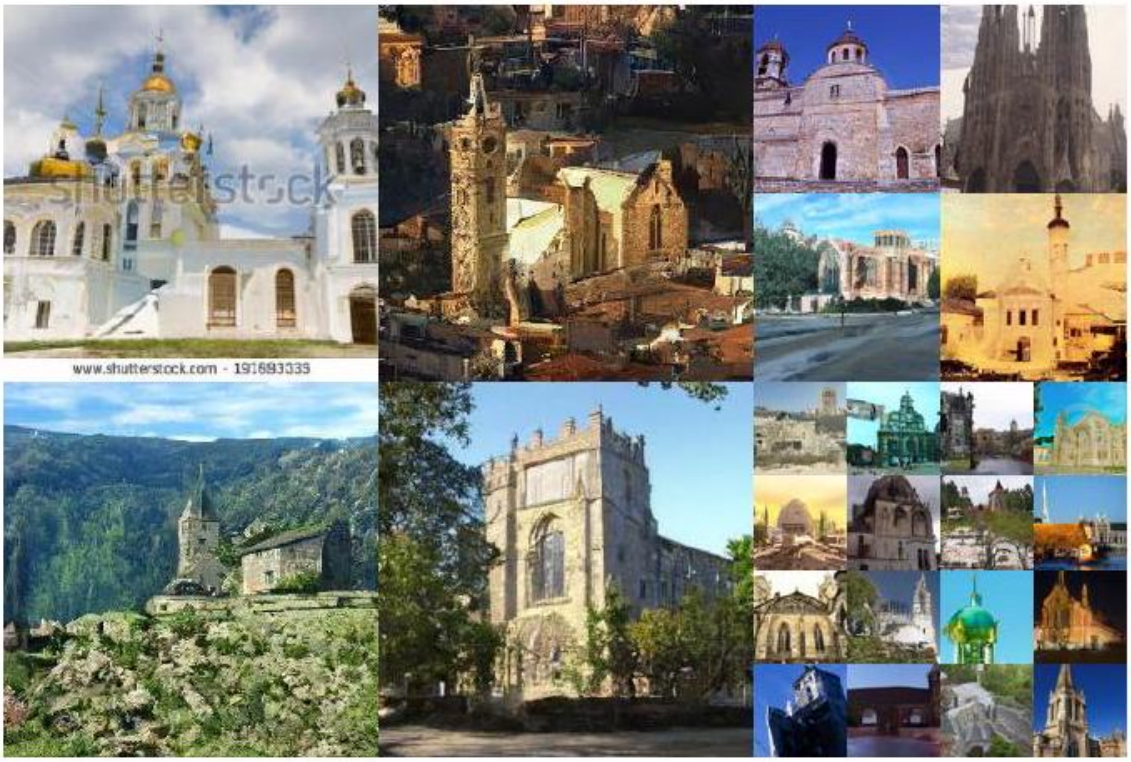

图 3：LSUN 教会样本。 FID = 7.89

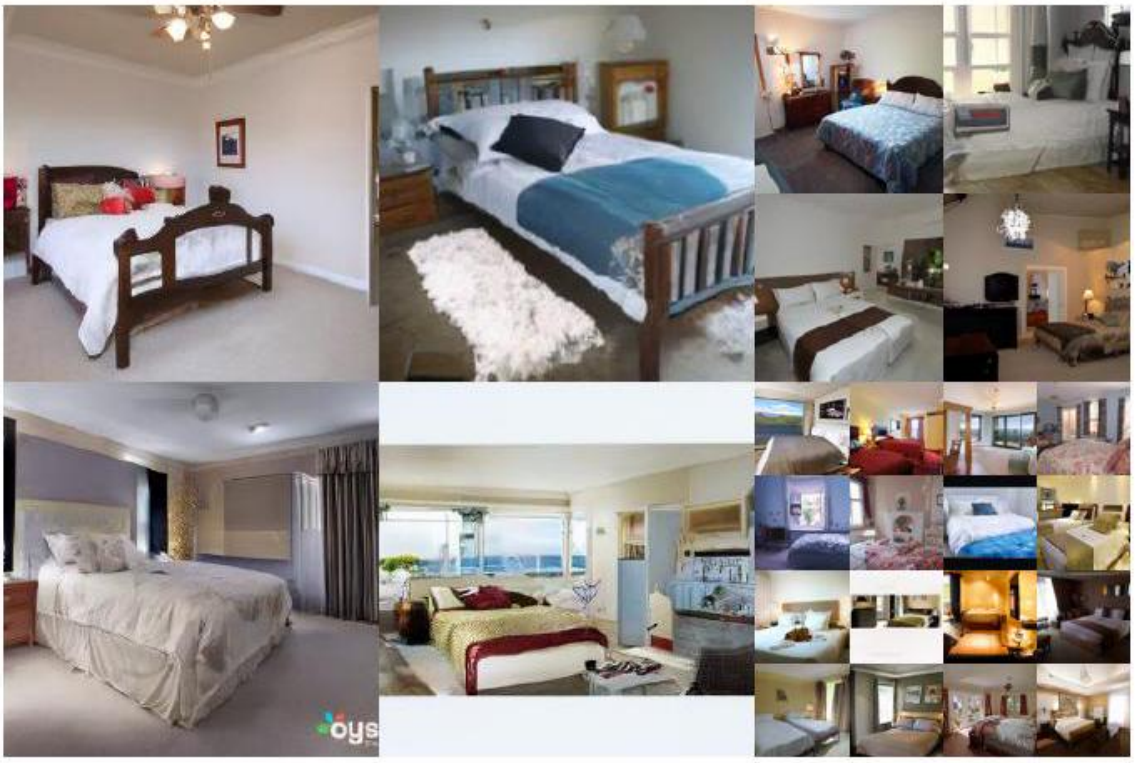

图 4：LSUN 卧室样本。 FID = 4.90

Algorithm 3 Sending $\mathbf{x}_0$

1: Send $\mathbf{x}_T \sim q(\mathbf{x}_T|\mathbf{x}_0)$ using $p(\mathbf{x}_T)$
2: **for** $t = T-1, \ldots, 2, 1$ **do**
3: Send $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0)$ using $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$
4: **end for**
5: Send $\mathbf{x}_0$ using $p_\theta(\mathbf{x}_0|\mathbf{x}_1)$

Algorithm 4 Receiving

1: Receive $\mathbf{x}_T$ using $p(\mathbf{x}_T)$
2: **for** $t = T-1, \ldots, 1, 0$ **do**
3: Receive $\mathbf{x}_t$ using $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$
4: **end for**
5: **return** $\mathbf{x}_0$

我们发现，正如预期的那样，在真正的变分边界上训练我们的模型比在简化目标上训练产生更好的代码长度，但后者产生最好的样本质量。 CIFAR10 和 CelebA-HQ 256×256 样本参见图 1]，LSUN 256×256 样本 [71] 参见图 3 和图 4，更多信息参见附录 D。

逆过程参数化和训练目标消融

在表 2 中，我们展示了逆向过程参数化和训练目标（第 3.2 节）对样本质量的影响。我们发现预测 $\tilde{\boldsymbol{\mu}}$ 的baseline选项只有在真正的变分界而不是未加权的均方误差上训练时才有效，这是一个类似于方程式的简化目标。 14. 我们还看到，与固定方差相比，学习逆向过程方差（通过将参数化对角线 $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t)$ 纳入变分界）会导致训练不稳定和样本质量较差。正如我们提出的那样，预测 ϵ 在具有固定方差的变分界上训练时的表现与预测 $\tilde{\boldsymbol{\mu}}$ 大致相同，但在使用我们的简化目标进行训练时要好得多。

渐进式编码

表 1 还显示了我们的 CIFAR10 模型的代码长度。训练和测试之间的差距最多为每个维度 0.03 位，这与其他基于可能性的模型报告的差距相当，表明我们的扩散模型没有过度拟合（最近邻可视化请参见附录 D）。尽管如此，尽管我们的无损代码长度优于报告的基于能量的模型和使用退火重要性采样的得分匹配的大估计 [11]，但它们与其他类型的基于似然的生成模型 [7] 相比没有竞争力。

由于我们的样本质量仍然很高，因此我们得出结论，扩散模型具有使它们成为出色的有损压缩器的归纳偏差。将变分约束项 $L_1 + \cdots + L_T$ 视为速率，将 L_0 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/暗淡，失真为 **1.97** 位/暗淡，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。

其中 t 在 1 和 T 之间是统一的。 $t = 1$ 的情况对应于 L_0，离散解码器定义中的积分 (13 近似为高斯概率密度函数乘以 bin 宽度，忽略 σ_1^2 和边缘效应。 $t>1$ 的情况对应于方程 (12) 的加权版本，类似于 $NCSN$ 去噪扩散模型 [55] 使用的损失加权。（L_T 不出现是因为正向过程方差 β_t 是固定的。）算法 1 显示了具有这个简化目标的完整训练过程。

$$\mathbf{x}_0 \approx \hat{\mathbf{x}}_0 = \left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t)\right)/\sqrt{\bar{\alpha}_t}$$

由于等式(4). （随机重建 $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t)$ 也是有效的，但我们在这里不考虑它，因为它使失真更难以评估。）图 5 显示了 CIFAR10 测试集上的结果率失真图。在每个时间 t，失真被计算为均方根误差 $\sqrt{\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2/D}$，速率被计算为在时间 t 到目前为止接收到的累积比特数。在速率失真图的低速率区域，失真急剧下降，表明大部分比特确实分配给了难以察觉的失真。

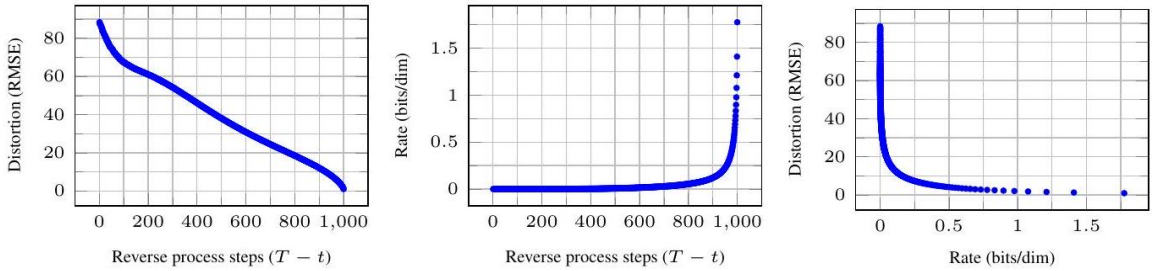

图 5：无条件 CIFAR10 测试集率失真 vs. 时间。失真以 $[0, 255]$ 等级的均方根误差来衡量。详情见表 4。

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，我们预测反向过程的结果 $\hat{\mathbf{x}}_0$，同时使用算法 2 从反向过程中采样 图 6 和 10 显示了反向过程中 $\hat{\mathbf{x}}_0$ 的最终样本质量。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 $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t)$，其中 $\mathbf{x}_t$ 冻结用于各种 t。当 t 较小时，除精细细节外的所有细节都会被保留，而当 t 较大时，只会保留大尺度特征。也许这些是概念压缩的暗示 [18]。

图 6：无条件 CIFAR10 渐进式生成（ $\hat{\mathbf{x}}_0$ 随着时间的推移，从左到右）。附录中随时间推移扩展的样本和样本质量指标（图 10 和 14。

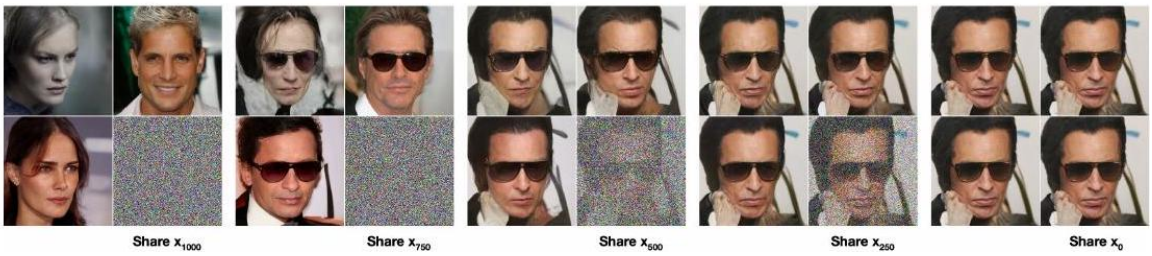

图 7：当以相同的潜伏条件为条件时，CelebA-HQ 256×256 样本共享高级属性。右下象限是 $\mathbf{x}_t$，其他象限是来自 $p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t)$ 的样本。

与自回归解码的连接请注意，变分界 (5) 可以重写为：

$$L = D_{\mathrm{KL}}(q(\mathbf{x}_T) \| p(\mathbf{x}_T)) + \mathbb{E}_q \left[\sum_{t \geq 1} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) \right] + H(\mathbf{x}_0)$$

（有关推导，请参见附录 A。）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 $q(\mathbf{x}_t \mid \mathbf{x}_0)$ 将所有概率质量放在 $\mathbf{x}_0$ 上，第一个 t 坐标masked 出（即 $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ mask 出 t^{th} 坐标），设置 $p(\mathbf{x}_T)$ 将所有质量放在空白图像上，并且为了论证，将 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 设为

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

是一个完全表达的条件分布。通过这些选择，$D_{\mathrm{KL}}(q(\mathbf{x}_T) \| p(\mathbf{x}_T)) = 0$ 和最小化 $D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t))$ 训练 p_θ 复制坐标 $t+1, \ldots, T$ 不变并预测给定 t^{th} 的 $t+1, \ldots, T$ 坐标。因此，使用这种特定的扩散训练 p_θ 就是在训练自回归模型。

因此，我们可以将高斯扩散模型 (2) 解释为一种具有广义位排序的自回归模型，不能通过重新排序数据坐标来表示。先前的工作表明，这种重新排序会引入对样本质量有影响的归纳偏差 [38]，因此我们推测高斯扩散具有类似的目的，可能会产生更大的影响，因为与相比之下，高斯噪声可能更自然地添加到图像中mask 噪音。此外，高斯扩散长度不限于等于数据维度；例如，我们使用 $T = 1000$，它小于我们实验中 $32 \times 32 \times 3$ 或 $256 \times 256 \times 3$ 图像的维度。高斯扩散可以缩短以实现快速采样，或延长以提高模型表达能力。

插值

我们可以使用 $\mathbf{x}_0, \mathbf{x}'_0 \sim q(\mathbf{x}_0)$ 作为随机编码器 q 在潜在空间中插入源图像 $\mathbf{x}_t, \mathbf{x}'_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$，然后通过反向过程 $\overline{\mathbf{x}}_t = (1-\lambda)\mathbf{x}_0 + \lambda \mathbf{x}'_0$ 将线性插入的潜在 $\overline{\mathbf{x}}_0 \sim p(\mathbf{x}_0 \mid \overline{\mathbf{x}}_t)$ 解码到图像空间中。实际上，我们使用反向过程从源图像的线性插值损坏版本中移除伪影，如图 8（左）所示。我们固定了 λ 不同值的噪声，因此 $\mathbf{x}_t$ 和 $\mathbf{x}'_t$ 保持不变。图 8（右）显示了原始 CelebA-HQ 256×256 图像 ($t = 500$) 的插值和重建。反向过程产生高质量的重建和合理的插值，平滑地改变姿势、肤色、发型、表情和背景等属性，但不包括眼镜。较大的 t 会导致更粗糙和更多样化的插值，在 $t = 1000$ 处具有新样本（附录图 9）。

相关工作

通过分数匹配和基于能量的建模之间已知的联系，我们的工作可能对其他最近关于基于能量的模型的工作产生影响 [67,-69, 12, 70, 13, 11, 41, 17, 8]。我们的率失真曲线是在一次对变分边界的评估中随时间计算的，这让人想起如何在一次退火重要性采样 [24] 中通过失真惩罚计算率失真曲线。我们的渐进式解码论点可以在卷积 DRAW 和相关模型 [18、40] 中看到，也可能导致更通用的子尺度排序设计或自回归模型的采样策略 [38、64]。

结论

我们已经使用扩散模型展示了高质量的图像样本，并且我们发现了扩散模型与训练马尔可夫链的变分推理、去噪分数匹配和退火Langevin动力学（以及基于能量的扩展模型）、自回归模型和渐进有损模型之间的联系压缩。由于扩散模型似乎对图像数据具有极好的归纳偏差，我们期待研究它们在其他数据模式中的效用以及作为其他类型的生成模型和机器学习系统的组件。

更广泛的影响

我们在扩散模型方面的工作与其他类型的深度生成模型的现有工作具有相似的范围，例如努力提高GAN、流、自回归模型等的样本质量。我们的论文代表了在使扩散模型成为此类技术中普遍有用的工具方面取得的进展，因此它可能有助于扩大生成模型对更广阔世界已经产生（并将产生）的任何影响。

不幸的是，有许多众所周知的恶意使用生成模型。样本生成技术可用于制作出于政治目的的知名人物的虚假图像和视频。虽然假图像是在软件工具可用之前很久手动创建的，但像我们这样的生成模型使这个过程变得更容易。幸运的是，CNN 生成的图像目前存在允许检测的细微缺陷 [62]，但生成模型的改进可能会使检测变得更加困难。生成模型还反映了训练它们的数据集中的偏差。由于许多大型数据集是通过自动化系统从互联网上收集的，因此很难消除这些偏差，尤其是当图像未标记时。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，扩散模型可能对数据压缩有用，随着数据变得更高分辨率和全球互联网流量增加，这可能对于确保互联网对广大受众的可访问性至关重要。我们的工作可能有助于从图像分类到强化学习的大量下游任务的未标记原始数据的表示学习，扩散模型也可能在艺术、摄影和音乐的创造性用途中变得可行。

致谢和资金披露

这项工作得到了 ONR PECASE 和 NSF 研究生研究奖学金的支持，资助号为 DGE-1752814。谷歌的 TensorFlow Research Cloud (TFRC) 提供了云 TPU