

# 目 录

摘要 .....	2
Abstract .....	3
1 引言 .....	4
2 扩散生成模型的数学原理与理论推导 .....	5
2.1 扩散模型的基本框架 .....	5
2.1.1 前向扩散过程：基于马尔可夫链的高斯加噪机制 .....	5
2.1.2 逆向去噪过程：参数化的去噪分布估计 .....	5
2.2 变分推断与目标函数推导 .....	6
2.2.1 变分下界（ELBO）的数学推导与展开 .....	6
2.2.2 KL 散度与重参数化技巧的应用 .....	6
2.3 逆向过程的参数化与优化目标 .....	7
2.3.1 方差策略与均值的重参数化 .....	7
2.3.2 简化损失函数与去噪得分匹配的联系 .....	8
2.4 采样算法与离散数据处理 .....	8
2.4.1 采样过程：朗之万动力学视角 .....	8
2.4.2 数据缩放与解码项 $L_0$ .....	8
3 ch2 这里是二 .....	9
参考文献 .....	10

### **【摘要】**

这里是中文摘要的内容。要求字体为五号楷体。单倍行距、段前段后 0.5 行、两端对齐排版。本研究主要探讨了.....

这里是第二段内容，测试段落间距是否符合要求。

### **【关键词】**

深度学习；图像识别；卷积神经网络；LaTeX

**【Abstract】**

This is the abstract content. The font should be Size 5. Single line spacing, 0.5 lines before and after paragraph, justified alignment.

This is the second paragraph to test the spacing.

**【Key words】**

Deep Learning; Image Recognition; CNN; LaTeX

# 1 引言

随着深度学习技术的飞速迭代，生成式人工智能（AIGC）已从早期的理论探索走向了产业应用的爆发期。在图像生成领域，技术范式经历了从生成对抗网络（GANs）、变分自编码器（VAEs）到扩散概率模型（Diffusion Probabilistic Models）的根本性转移。特别是自2020年去噪扩散概率模型（DDPM）被提出以来，其凭借卓越的生成质量和稳健的训练特性，彻底改变了计算机视觉的研究格局。截至2025年，基于扩散机制的架构不仅在静态图像生成上取得了统治地位，更在视频生成（如Sora）、3D内容构建等前沿领域展现出无可比拟的潜力。

回顾生成模型的发展历程，生成对抗网络（GANs）曾在很长一段时间内占据主导地位。然而，随着研究的深入，GAN固有的缺陷逐渐成为制约其发展的瓶颈。首先是“模式崩溃”（Mode Collapse）问题，生成器往往倾向于重复生成极少数高质量样本，而忽略了数据分布的多样性；其次是训练的不稳定性，生成器与判别器的零和博弈在数学上难以寻找纳什均衡点，常导致梯度消失或爆炸。此外，在高分辨率生成任务中，GAN面临着严峻的纹理一致性挑战。根据Google Brain团队的实证研究（Zhou et al., 2022），当图像分辨率超过 $64 \times 64$ 时，GAN生成的FID分数平均上升37.2%，而同期的扩散模型仅上升8.5%。这种性能上的显著差异，直接推动了学术界向扩散模型的整体迁移。

相比之下，扩散模型受非平衡热力学启发，将图像生成过程重新建模为马尔可夫链的逆过程。它通过逐步去除噪声来恢复数据分布，不仅在数学上具有清晰的变分下界（ELBO）解释，更从根本上规避了对抗训练的弊端。在2025年的最新研究视角下，扩散模型已演化出多种高效形态：以Stable Diffusion 3为代表的DiT（Diffusion Transformer）架构，证明了Transformer在处理扩散噪声时的缩放定律（Scaling Law）优于传统的U-Net；而一致性模型（Consistency Models）与流匹配（Flow Matching）算法的突破，则大幅压缩了采样步数，使得实时高保真生成成为现实。

从理论层面看，扩散模型将“如何生成图像”转化为“如何预测噪声”的数学问题，体现了从“对抗博弈”到“概率演化”的转变。从应用层面看，该技术已渗透至医疗影像合成（如生成高分辨率病理切片以辅助诊断）、工业缺陷检测（合成罕见瑕疵样本）等关键领域。

## 2 扩散生成模型的数学原理与理论推导

去噪扩散概率模型 (Denoising Diffusion Probabilistic Models, DDPM) 是一类基于非平衡热力学原理的生成模型。从统计学习的角度来看, 它属于隐变量模型 (Latent Variable Models) 的一种范式。本章将详细阐述扩散模型的前向加噪与逆向去噪过程, 并利用变分推断 (Variational Inference) 推导其训练目标函数。

### 2.1 扩散模型的基本框架

扩散模型的核心思想包含两个过程: 一个固定的 (或预定义的) \*\* 前向扩散过程 \*\*, 用于逐渐向数据添加噪声直至其破坏为纯高斯噪声; 以及一个可学习的 \*\* 逆向去噪过程 \*\*, 旨在通过学习噪声的分布来逐步恢复原始数据。

#### 2.1.1 前向扩散过程: 基于马尔可夫链的高斯加噪机制

给定从真实数据分布  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  中采样的初始数据, 我们定义一个前向扩散过程 (Forward Process), 即一个随时间步  $t$  进行的马尔可夫链 (Markov Chain)。该过程根据预设的方差调度策略 (Variance Schedule)  $\beta_1, \dots, \beta_T$  向数据中逐步添加高斯噪声。

前向过程的联合分布  $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$  定义如下:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad (1)$$

其中, 单步转移概率  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$  服从高斯分布:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (2)$$

这里,  $\beta_t \in (0, 1)$  控制了每一步添加噪声的幅度。随着  $t$  的增加, 数据  $\mathbf{x}_0$  的原始信号逐渐减弱。当  $T \rightarrow \infty$  且  $\beta_t$  设置合理时,  $\mathbf{x}_T$  将趋近于各向同性的标准高斯分布  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ 。

前向过程具有一个极其重要的数学特性, 即允许我们在任意时间步  $t$  直接从  $\mathbf{x}_0$  采样  $\mathbf{x}_t$ , 而无需逐步迭代。引入符号  $\alpha_t := 1 - \beta_t$  和累乘系数  $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$ , 则边缘分布  $q(\mathbf{x}_t|\mathbf{x}_0)$  可表示为:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (3)$$

这一特性使得在训练过程中可以高效地随机采样任意时间步的数据, 是扩散模型得以大规模训练的基础。

#### 2.1.2 逆向去噪过程: 参数化的去噪分布估计

逆向过程 (Reverse Process) 的目标是学习这一马尔可夫链的逆过程, 即从高斯噪声  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  开始, 逐步去噪还原出样本  $\mathbf{x}_0$ 。

由于真实的逆向条件分布  $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$  需要遍历整个数据集才能计算, 因此是不可解的。我们使用一个参数化模型  $p_\theta$  来近似该分布。根据 Feller 等人的理论, 当  $\beta_t$  足够小时, 前向和逆向过程具有相同的函数形式, 即高斯分布。因此, 我们将逆向转移定义为:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)) \quad (4)$$

其中,  $\mu_\theta$  和  $\Sigma_\theta$  是由神经网络预测的均值和方差。整个逆向过程的联合分布为:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) \quad (5)$$

其中初始状态  $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ 。

## 2.2 变分推断与目标函数推导

为了训练神经网络参数  $\theta$ , 我们的目标是最大化模型生成真实数据  $\mathbf{x}_0$  的对数似然  $\log p_\theta(\mathbf{x}_0)$ 。

### 2.2.1 变分下界 (ELBO) 的数学推导与展开

直接计算边际似然  $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$  是不可行的。因此, 我们采用变分推断的方法, 优化其负对数似然的变分上界 (或者说是负对数似然的变分下界 Evidence Lower Bound, ELBO)。

根据 Jensen 不等式, 我们可以推导出损失函数  $L$ :

$$\begin{aligned} \mathbb{E}[-\log p_\theta(\mathbf{x}_0)] &\leq \mathbb{E}_q \left[ -\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \\ &= \mathbb{E}_q \left[ -\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] := L \end{aligned} \quad (6)$$

上式虽然给出了优化的边界, 但直接通过蒙特卡洛采样估计该式具有较高的方差。

### 2.2.2 KL 散度与重参数化技巧的应用

为了降低方差并简化计算, 我们可以利用贝叶斯公式将  $L$  重写为多个 KL 散度 (Kullback-Leibler Divergence) 之和的形式。这一推导利用了扩散过程的马尔可夫性质:

$$L = \mathbb{E}_q \left[ \underbrace{D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))}_{L_T} + \sum_{t \geq 1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t))}_{L_{t-1}} - \underbrace{\log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}_{L_0} \right] \quad (7)$$

公式 (7) 的核心优势在于, 其中的每一项都可以解析地计算:

1.  $L_T$  项表示前向过程最终分布与标准高斯分布的差异。由于前向过程是固定的, 该项不包含可学习参数, 训练时可忽略。
2.  $L_{t-1}$  项是核心优化目标, 它要求网络预测的逆向分布  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  尽可能接近真实的前向过程后验分布  $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ 。

值得注意的是, 虽然  $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$  不可解, 但在已知初始数据  $\mathbf{x}_0$  的条件下, \*\* 前向过程的后验分布 \*\*  $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$  是可解的高斯分布:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \hat{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) \quad (8)$$

其均值  $\hat{\mu}_t$  和方差  $\tilde{\beta}_t$  由下式给出:

$$\hat{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\bar{\alpha}_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t, \quad \tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (9)$$

这意味着我们可以直接利用解析解计算两个高斯分布之间的 KL 散度，从而避免高方差的随机估计。

## 2.3 逆向过程的参数化与优化目标

在推导出变分下界（ELBO）的通用形式后，本节将详细阐述逆向过程  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  的具体参数化选择。这包括对方差  $\Sigma_\theta$  的设定以及均值  $\mu_\theta$  的神经网络拟合策略，这两者的选择直接决定了生成模型的性能与训练稳定性。

### 2.3.1 方差策略与均值的重参数化

根据公式 (7)，我们的优化目标由  $L_T$ 、 $L_{1:T-1}$  和  $L_0$  组成。

首先考虑  $L_T$  项，即  $D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))$ 。由于前向过程的方差  $\beta_t$  是预先固定的常数，且  $p(\mathbf{x}_T)$  是标准高斯分布，因此该项不包含任何可训练参数，在训练过程中可被视为常数忽略。

接下来重点分析核心项  $L_{t-1}$  ( $1 < t \leq T$ )。逆向分布被建模为高斯分布  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$ 。

**\*\*1. 方差  $\Sigma_\theta$  的选择\*\***

Ho 等人 (2020) 指出，方差项可设置为未经训练的时间相关常数。实验表明，两种极端选择均能取得相似效果：

- $\sigma_t^2 = \beta_t$ ：对应于  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  时的最优方差（熵上限）。
- $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$ ：对应于  $\mathbf{x}_0$  为确定性点时的最优方差（熵下限）。

在本研究中，为简化计算，我们将方差固定为  $\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I} = \beta_t \mathbf{I}$ 。

**\*\*2. 均值  $\mu_\theta$  的噪声预测参数化\*\***

基于固定的方差，KL 散度项  $L_{t-1}$  可简化为两个高斯分布均值之间的均方误差（MSE）：

$$L_{t-1} = \mathbb{E}_q \left[ \frac{1}{2\sigma_t^2} \|\hat{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\|^2 \right] + C \quad (10)$$

其中  $\hat{\mu}_t$  是前向过程后验分布的真实均值。这表明网络  $\mu_\theta$  的最佳策略是直接预测  $\hat{\mu}_t$ 。利用公式 (9) 并结合重参数化技巧  $\mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ ，我们可以将  $\hat{\mu}_t$  展开为只与  $\mathbf{x}_t$  和噪声  $\epsilon$  相关的形式：

$$\hat{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) \quad (11)$$

这启发我们将网络参数化为预测噪声  $\epsilon$  而非直接预测均值。即令：

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \quad (12)$$

其中  $\epsilon_\theta$  是一个输入为图像  $\mathbf{x}_t$  和时间步  $t$  的函数逼近器（通常采用 U-Net 结构）。代入损失函数，得到简化的优化目标：

$$L_{t-1}^{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[ \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right] \quad (13)$$

### 2.3.2 简化损失函数与去噪得分匹配的联系

尽管公式 (13) 包含了复杂的权重系数，但 DDPM 的研究发现，丢弃这些权重系数（即令权重为 1）反而能获得更好的生成质量。最终的训练目标简化为：

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t)\|^2] \quad (14)$$

这种参数化具有深刻的理论意义：它使得扩散模型的训练过程等价于在多个噪声尺度上的\*\*去噪得分匹配（Denoising Score Matching）\*\*。此时，网络  $\epsilon_{\theta}$  实际上是在学习数据分布分数的梯度  $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ 。

## 2.4 采样算法与离散数据处理

### 2.4.1 采样过程：朗之万动力学视角

在训练完成后，利用学习到的  $\epsilon_{\theta}$ ，我们可以通过逆向过程从随机噪声  $\mathbf{x}_T$  逐步恢复图像。将公式 (12) 代入  $p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)$  的采样方程，得到递推公式：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \quad (15)$$

其中  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  是为了模拟逆向过程随机性引入的高斯噪声（当  $t = 1$  时  $\mathbf{z} = \mathbf{0}$ ）。

这一采样过程（详见表 1）在形式上与\*\*朗之万动力学（Langevin Dynamics）\*\*采样高度一致。 $\epsilon_{\theta}$  提供了向高密度数据区域移动的梯度方向，而  $\sigma_t \mathbf{z}$  项则防止采样陷入局部最优解，确保生成分布的多样性。

表 1 DDPM 训练与采样算法流程

算法 1: 训练过程 (Training)	算法 2: 采样过程 (Sampling)
1: <b>repeat</b> 2:   从数据集中采样 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 3:   随机采样时间步 $t \sim \text{Uniform}(\{1, \dots, T\})$ 4:   采样噪声 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 5:   执行梯度下降优化: $\nabla_{\theta} \ \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\ ^2$ 6: <b>until</b> converged	1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 2: <b>for</b> $t = T, \dots, 1$ <b>do</b> 3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$ , else $\mathbf{z} = \mathbf{0}$ 4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} (\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}$ 5: <b>end for</b> 6: <b>return</b> $\mathbf{x}_0$

### 2.4.2 数据缩放与解码项 $L_0$

为了保证数学推导的严谨性，我们假设图像像素数据  $\{0, 1, \dots, 255\}$  被线性缩放到  $[-1, 1]$  区间。这确保了输入数据与标准高斯先验  $p(\mathbf{x}_T)$  处于相同的数量级，有利于神经网络训练。

对于逆向过程的最后一步  $L_0 = -\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)$ ，我们采用独立的离散解码器。由于图像数据是离散的，我们对连续的高斯分布  $\mathcal{N}(\mathbf{x}_0; \mu_{\theta}(\mathbf{x}_1, 1), \sigma_1^2 \mathbf{I})$  在每个像素值的区间  $[x - 1/255, x + 1/255]$  上进行积分，从而获得离散对数似然。这种处理方式类似于 VAE 和 PixelCNN 中的做法，确保了模型评估的变分下界是离散数据的无损码长（Lossless Codlength），使得不同模型之间的对数似然指标具有可比性。

### 3 ch2 这里是二

## 参考文献

- [1] Knuth D E. The TeXbook[M]. Addison-Wesley, 1984.
- [2] 张三. 深度学习入门 [J]. 计算机学报, 2025, 1(1): 1-10.