

# Diffusion Models in Vision: A Survey

Florinel-Alin Croitoru, Vlad Hondu, Radu Tudor Ionescu, *Member, IEEE*, and Mubarak Shah, *Fellow, IEEE*

**摘要—警告：**该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址：[https://github.com/binary-husky/gpt\\_academic/](https://github.com/binary-husky/gpt_academic/)。项目在线体验地址：<https://chatpaper.org>。当前大语言模型：gpt-3.5-turbo，当前语言模型温度设定：1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址：[https://github.com/binary-husky/gpt\\_academic/](https://github.com/binary-husky/gpt_academic/)。项目在线体验地址：<https://chatpaper.org>。当前大语言模型：gpt-3.5-turbo，当前语言模型温度设定：1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

去噪扩散模型是计算机视觉领域的最新研究课题，展示出在生成建模方面的显著结果。扩散模型是一种基于两个阶段的深度生成模型，包括前向扩散阶段和逆向扩散阶段。在前向扩散阶段，通过逐步加入高斯噪声对输入数据进行逐渐扰动。在逆向阶段，模型被要求通过逐步学习逆转扩散过程，来恢复原始输入数据。尽管由于采样时涉及的步骤多，扩散模型存在已知的计算负担，即低速度，但其生成样本的质量和多样性仍广受赞赏。在本调研中，我们对应用于计算机视觉领域的去噪扩散模型的文章进行了全面回顾，包括该领域的理论和实际贡献。首先，我们确定并介绍了三种通用扩散建模框架，它们分别是基于去噪扩散概率模型、噪声条件化评分网络和随机微分方程。我们进一步讨论了扩散模型与其他深度生成模型之间的关系，包括变分自编码器、生成对抗网络、基于能量的模型、自回归模型和归一化流。然后，我们介绍了应用于计算机视觉中的扩散模型的多角度分类。最后，我们说明了扩散模型目前存在的局限性，并展望了一些有趣的未来研究方向。

**Index Terms**—diffusion models, denoising diffusion models, noise conditioned score networks, score-based models, image generation, deep generative modeling.

## 1 INTRODUCTION

扩散模型 [?], [?], [?], [?], [?], [?], [?], [?], [?], [?] 是一类深度生成模型，最近成为计算机视觉领域热门的话题之一（见图1），展现出令人印象深刻的生成能力，从高级细节到生成示例的多样性。我们甚至可以说，这些生成模型将生成建模领域提升到一个新的水平，特别是 Imag-en [?] 和潜在扩散模型 (Latent Diffusion Models, LDMs) [?] 等模型。图2中展示的图像样本由 Stable Diffusion 生成，它是 LDMs 的一种版本 [?], 根据文本提示生成图像。生成的图像几乎没有伪影，并且与文本提示非常吻合。值得注意的是，这些提示是故意选择的，目的是代表训练时从未见过的不现实的情景，从而展示了扩散模型的高泛化能力。迄今为止，扩散模型已被应用于各种生成建模任务，如图像生成 (Sohl 等, 2015 年; Song 等, 2019 年; Ho 等, 2020 年; Song 等, 2020 年; Song 等, 2021 年; Dhariwal 等, 2021 年; Nichol 等, 2021 年; Song 等, 2021 年; Sinha 等, 2021 年; Vahdat 等, 2021 年; Saharia 等, 2021 年; Nichol 等, 2021 年; Pandey 等, 2021 年; Rombach 等, 2022 年; Bao 等, 2022 年;

Dockhorn 等, 2021 年; Rombach 等, 2022 年; Liu 等, 2022 年; Jiang 等, 2022 年), 图像超分辨率 (Saharia 等, 2021 年; Batzolis 等, 2021 年; Daniels 等, 2021 年; Rombach 等, 2022 年; Chung 等, 2022 年; Kawar 等, 2022 年), 图像修复 (Sohl 等, 2015 年; Song 等, 2019 年; Song 等, 2021 年; Esser 等, 2021 年; Batzolis 等, 2021 年; Lugmayr 等, 2022 年; Rombach 等, 2022 年; Chung 等, 2022 年; Jing 等, 2022 年), 图像编辑 (Avrahami 等, 2022 年; Choi 等, 2021 年; Meng 等, 2021 年), 图像到图像的转换 (Saharia 等, 2022 年; Choi 等, 2021 年; Zhao 等, 2022 年; Wang 等, 2022 年; Li 等, 2022 年; Wolleb 等, 2022 年), 等等。此外，扩散模型学习的潜在表征还被发现对于判别任务也非常有用，例如图像分割 (Baranchuk 等, 2021 年; Graikos 等, 2022 年; Wolleb 等, 2021 年; Amit 等, 2021 年)，分类 (Zimmermann 等, 2021 年) 以及异常检测 (Pinaya 等, 2022 年; Wolleb 等, 2022 年; Wyatt 等, 2022 年)。这证实了降噪扩散模型的广泛适用性，并表明还有待发现更多的应用。此外，学习强大的潜在表征的能力，与表示学习 (Bengio 等, 2013 年; Goodfellow, 2016 年) 建立了联系，表示学习是一个综合性的领域，研究如何学习强大的数据表示，包括从设计新的神经网络架构 (Hinton, 2006 年; Kingma, 2014 年; Higgins 等, 2017 年; Goodfellow, 2014 年) 到开发学习策略 (Caron, 2020 年; Chen 等, 2020 年; Croitoru 等, 2022 年; Oord, 2018 年; Samuli 等, 2017 年; Tarvainen, 2017 年) 等多种方法。

- F.A. Croitoru, V. Hondu and R.T. Ionescu are with the Department of Computer Science, University of Bucharest, Bucharest, Romania. F.A. Croitoru and V. Hondu have contributed equally. R.T. Ionescu is the corresponding author.

E-mail: raducu.ionescu@gmail.com

- M. Shah is with the Center for Research in Computer Vision (CRCV), Department of Computer Science, University of Central Florida, Orlando, FL, 32816.

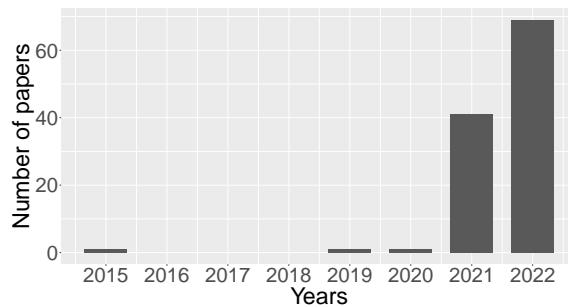


图 1. 每年关于扩散模型的论文大致数量。

根据图 1 所示的曲线，扩散模型的研究论文数量正在以极快的速度增长。为了概述这一快速发展主题的历史和现状成就，我们对计算机视觉中去噪扩散模型的研究文章进行了全面回顾。更具体地说，我们调查了下面定义的生成模型类别中的文章。扩散模型是一类基于以下两个阶段构建的深度生成模型：(i) 正向扩散阶段，即通过逐步添加高斯噪声逐渐扰动输入数据，和 (ii) 反向（反扩散）阶段，即生成模型被要求学习逐步逆转扩散过程，从扩散的（有噪声的）数据中逐渐恢复原始输入数据。

我们要强调的是，符合上述定义的扩散模型至少有三个子类别。第一个子类别包括去噪扩散概率模型 (DDPMs) [?], [?], 它们受到非平衡热力学理论的启发。DDPMs 是使用潜在变量来估计概率分布的潜变量模型。从这个角度来看，DDPMs 可以被看作是一种特殊类型的变分自动编码器 (VAEs) [?], 其中正向扩散阶段对应于 VAE 内部的编码过程，而反向扩散阶段对应于解码过程。第二个子类别由噪声条件得分网络 (NCSNs) [?] 表示，它们基于通过评估不同噪声水平下扰动数据分布的分数函数（定义为对数密度的梯度）的共享神经网络进行训练。随机微分方程 (SDEs) [?] 是一种用于建模扩散的替代方法，形成了扩散模型的第三个子类别。通过正向和反向 SDE 对扩散进行建模可以获得高效的生成策略以及强大的理论结果 [?]。基于 SDE 的后一种表述可以被看作是对 DDPMs 和 NCSNs 的泛化。

我们确定了几个定义性设计选择，并将它们综合到三个通用扩散建模框架中，分别对应上述三个子类别。为了将通用扩散建模框架放入背景中，我们进一步讨论了扩散模型与其他深度生成模型之间的关系。具体而言，我们描述了与变分自动编码器 (VAEs) [?], 生成对抗网络 (GANs) [?], 能量基模型 (EBMs) [?], [?], 自回归模型 [?] 和归一化流 [?], [?] 的关系。然后，我们介绍了在计算机视觉中应用的扩散模型的多角度分类，根据多个标准对现有模型进行分类，例如基础框架、目标任务或去噪条件。最后，我们阐述了扩散模型的现有限制，并设想了一些有趣的未来研究方向。例如，也许最大的问题之一是推理过程中的时间效率差，这是由于需要进行大量的评估步骤（例如数千步）才能生成一个样本 [?]。自然地，克服这一限制而不影响生成样本的质量，是未来研究的重要方向。

总之，我们的贡献有两个方面：

- 由于基于扩散模型的许多研究近年来在视觉领域涌现出来，我们对应用于计算机视觉中的去噪扩散模型进行了全面而及时的文献综述，旨在为读者快速理解通用的扩散建模框架。
- 我们设计了一种多角度的扩散模型分类方法，旨在帮助其他研究人员在特定领域应用扩散模型时快速找到相关的作品。

## 2 GENERIC FRAMEWORK

扩散模型是一类概率生成模型，它学习逆转逐渐降低训练数据结构的过程。因此，训练过程包括两个阶段：前向扩散过程和反向去噪过程。

前者由多个步骤组成，在每个输入图像中添加低级噪声，噪声的规模在每个步骤中都有所变化。训练数据逐渐被破坏，直到最终变为纯高斯噪声。

后者通过逆转前向扩散过程来表示。使用相同的迭代过程，但是反向进行：噪声逐步被去除，因此原始图像被重新创建。因此，在推理时，图像通过逐步重建开始于随机白噪声而生成。每个时间步骤减去的噪声是通过神经网络估计的，通常基于 U-Net 架构 [?], 以保持尺寸不变。

在接下来的三个小节中，我们介绍三种扩散模型的形式，分别是去噪扩散概率模型、噪声条件分数网络和基于随机微分方程的方法，该方法推广了前两种方法。对于每种形式，我们描述了向数据添加噪声的过程，学习逆转该过程的方法，以及推理时如何生成新样本。在图 3 中，所有三种形式都被表示为一个通用框架。

我们在最后一个章节中讨论与其他深度生成模型的关系。

### 2.1 Denoising Diffusion Probabilistic Models (DDPMs)

前向过程。DDPMs ([?], [?]) 通过使用高斯噪声缓慢破坏训练数据。设  $p(x_0)$  为数据密度，其中下标 0 表示数据为未破坏（原始）的。给定一个未破坏的训练样本  $x_0 \sim p(x_0)$ ，根据以下马尔科夫过程获得加噪版本  $x_1, x_2, \dots, x_T$ ：

$$p(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t} \cdot x_{t-1}, \beta_t \cdot \mathbf{I}\right), \forall t \in \{1, \dots, T\}, \quad (1)$$

其中  $T$  是扩散步数， $\beta_1, \dots, \beta_T \in [0, 1]$  是表示在不同扩散步中方差调度的超参数， $\mathbf{I}$  是与输入图像  $x_0$  具有相同维度的单位矩阵， $\mathcal{N}(x; \mu, \sigma)$  表示具有均值  $\mu$  和协方差  $\sigma$  的正态分布生成  $x$ 。这个递归公式的一个重要特性是，当  $t$  从均匀分布中抽取时，即  $\forall t \sim \mathcal{U}(\{1, \dots, T\})$ ，它也允许直接对  $x_t$  进行采样：

$$p(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\hat{\beta}_t} \cdot x_0, (1 - \hat{\beta}_t) \cdot \mathbf{I}\right), \quad (2)$$

其中， $\hat{\beta}_t = \prod_{i=1}^t \alpha_i$  且  $\alpha_t = 1 - \beta_t$ 。实质上，公式 (2) 表明，如果我们拥有原始图像  $x_0$  并且确定一个方差调度  $\beta_t$ ，我们可以通过单一步骤来采样任何噪声版本的  $x_t$ 。

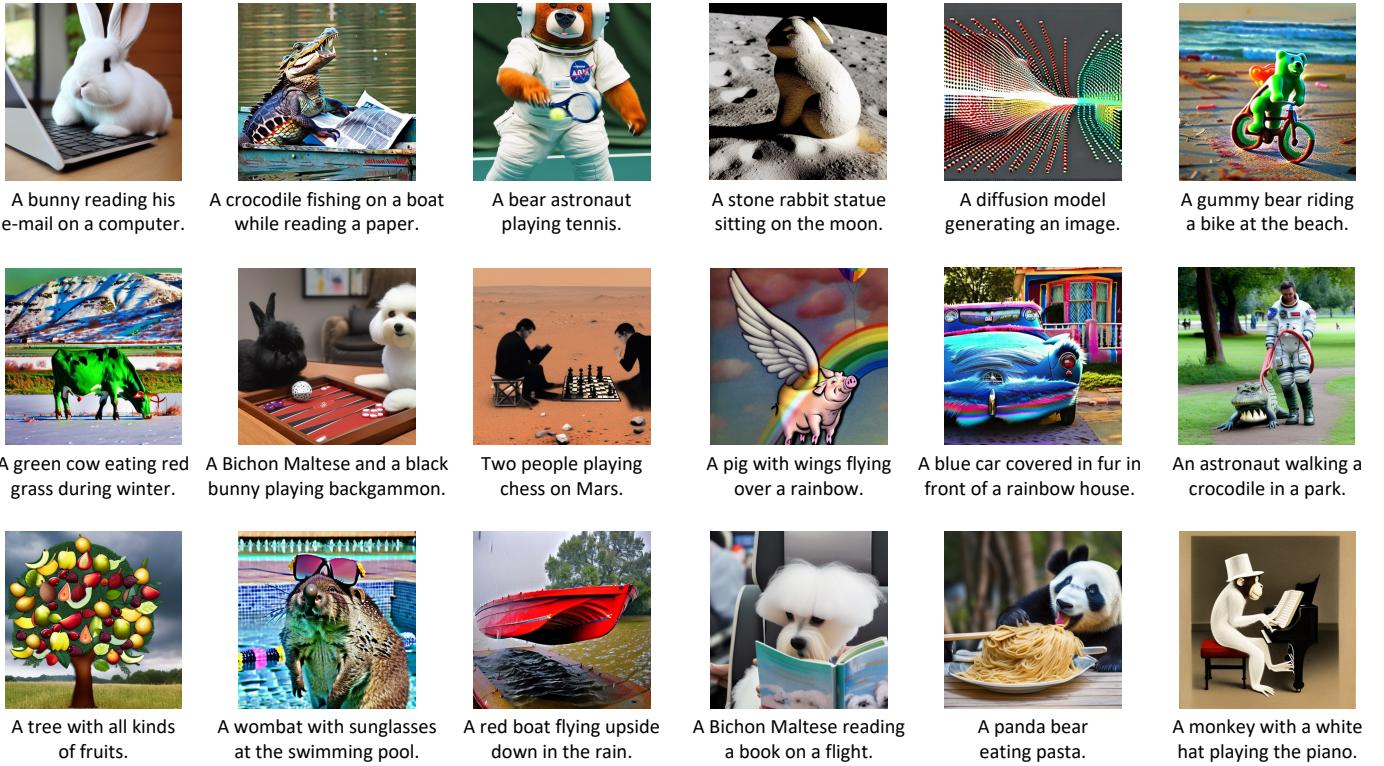


图 2. 由基于不同文本提示的稳定扩散 (Stable Diffusion) [?]生成的图像，通过<https://beta.dreamstudio.ai/dream>平台产生。

从 $p(x_t|x_0)$ 中进行采样是通过重新参数化技巧完成的。通常来说，为了对正态分布 $x \sim \mathcal{N}(\mu, \sigma^2 \cdot \mathbf{I})$ 的样本 $x$ 进行标准化，我们会减去平均值 $\mu$ 并除以标准差 $\sigma$ ，从而得到标准正态分布的样本 $z = \frac{x-\mu}{\sigma}$ ，其中 $z \sim \mathcal{N}(0, \mathbf{I})$ 。重新参数化技巧执行的是这个操作的逆过程，从 $z$ 开始，通过将 $z$ 乘以标准差 $\sigma$ 并加上平均值 $\mu$ ，得到样本 $x$ 。如果我们将这个过程应用到我们的情况中，那么 $x_t$ 将按以下方式从 $p(x_t|x_0)$ 中进行采样：

$$x_t = \sqrt{\hat{\beta}_t} \cdot x_0 + \sqrt{(1 - \hat{\beta}_t)} \cdot z_t, \quad (3)$$

### $\beta_t$ 的性质

如果方差调度 $(\beta_t)_{t=1}^T$ 被选择成 $\hat{\beta}_T \rightarrow 0$ ，那么根据公式(2)， $x_T$ 的分布应该很好地近似于标准高斯分布 $\pi(x_T) = \mathcal{N}(0, \mathbf{I})$ 。此外，如果每个 $(\beta_t)_{t=1}^T \ll 1$ ，那么逆向步骤 $p(x_{t-1}|x_t)$ 的函数形式与正向过程 $p(x_t|x_{t-1})$ 相同[?], [?]. 直观地说，当 $x_t$ 以非常小的步长生成时， $x_{t-1}$ 更有可能来自于与 $x_t$ 观察位置附近的区域，从而允许我们用一个高斯分布来建模这个区域。为了符合上述特性，Ho等人[?]将 $(\beta_t)_{t=1}^T$ 选择为在 $\beta_1 = 10^{-4}$ 和 $\beta_T = 2 \cdot 10^{-2}$ 之间线性增加的常数，其中 $T = 1000$ 。

### 逆向过程

利用上述性质，如果我们从一个样本 $x_T \sim \mathcal{N}(0, \mathbf{I})$ 开始并按照逆向步骤 $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu(x_t, t), \Sigma(x_t, t))$

来生成新样本，我们可以从 $p(x_0)$ 中生成新样本。为了近似这些步骤，我们可以训练一个神经网络 $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ ，该网络接收噪声图像 $x_t$ 和时间步长 $t$ 上的嵌入作为输入，并学习预测均值 $\mu_\theta(x_t, t)$ 和协方差 $\Sigma_\theta(x_t, t)$ 。

在理想情况下，我们会使用最大似然目标来训练神经网络，使得模型 $p_\theta(x_0)$ 对每个训练样本 $x_0$ 的分配概率尽可能大。然而，由于我们必须对所有可能的逆向轨迹进行边缘化计算， $p_\theta(x_0)$ 是不可计算的。因此，解决这个问题的方法[?], [?]是最小化负对数似然的变分下界，其形式如下：

$$\begin{aligned} \mathcal{L}_{vlb} = & -\log p_\theta(x_0|x_1) + KL(p(x_T|x_0)\|\pi(x_T)) \\ & + \sum_{t>1} KL(p(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)), \end{aligned} \quad (4)$$

其中KL表示两个概率分布之间的Kullback-Leibler散度。对这一目标函数的完整推导见附录[?]. 经过分析每个组成部分，我们可以看到第二项可以被移除，因为它不依赖于 $\theta$ 。最后一项表明神经网络在训练过程中，在每个时间步 $t$ 时， $p_\theta(x_{t-1}|x_t)$ 尽可能接近于在原始图像条件下前向过程的真实后验。此外，可以证明后验 $p(x_{t-1}|x_t, x_0)$ 是一个高斯分布，从而意味着KL散度具有闭合形式表达式。

Ho等人[?]提议将协方差 $\Sigma_\theta(x_t, t)$ 固定为一个常数，并将均值 $\mu_\theta(x_t, t)$ 重写为噪声的函数，如下所示：

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \hat{\beta}_t}} \cdot z_\theta(x_t, t) \right). \quad (5)$$

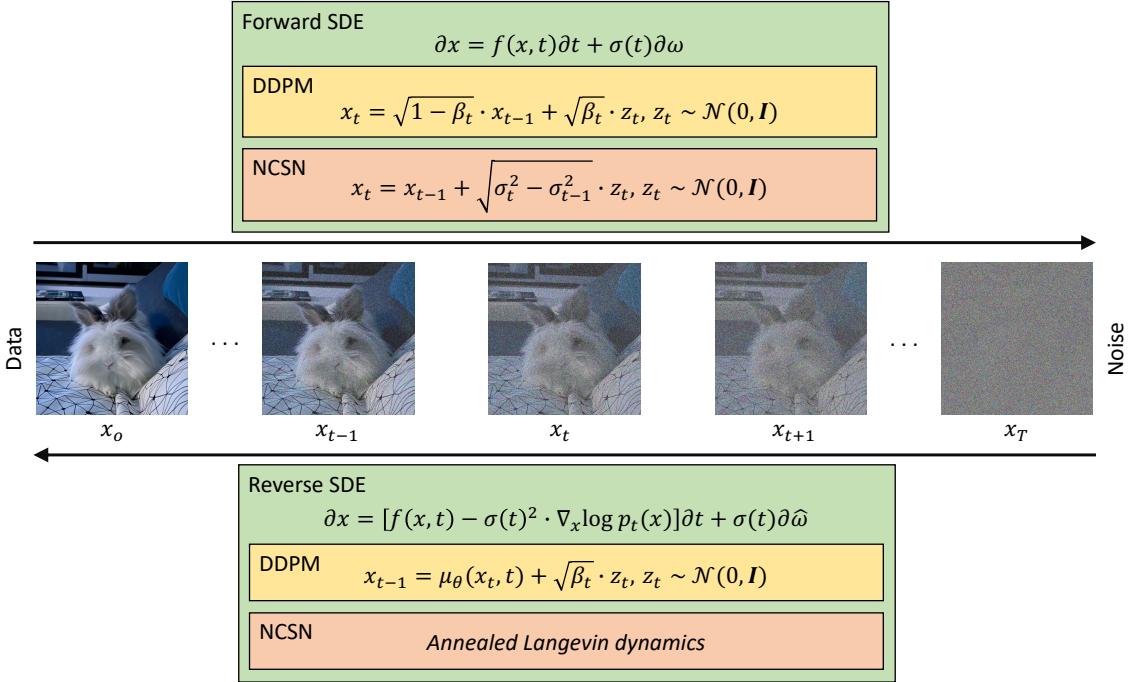


图 3. 一个通用框架，包括基于随机微分方程 (SDE)、降噪扩散概率模型 (DDPM) 和噪声条件分数网络 (NCSN) 的三种扩散模型的替代表达式。一般来说，扩散模型由两个过程组成。第一个过程称为正向过程，将数据转化为噪声，而第二个过程是一个生成过程，可以将正向过程的效果逆转。这个后续过程学会将噪声转换回数据。我们为这三种表达形式都举例说明这些过程。正向 SDE 表明， $x$  随时间的变化由一个函数  $f$  加上一个按比例缩放的随机成分  $\sigma \sim \mathcal{N}(0, \sigma_t)$ ，缩放比例为  $\sigma_t$ 。需要强调的是，不同的  $f$  和  $\sigma$  选择将导致不同的扩散过程。这就是为什么 SDE 表达式是另外两种表达式的推广。反向（生成）SDE 表明如何改变  $x$  以从纯噪声中恢复数据。我们保留随机成分，并使用对数概率  $\nabla_x \log p_t(x)$  的梯度来修改确定性成分，以便  $x$  移动到数据密度  $p(x)$  较高的区域。DDPM 在正向过程中从正态分布  $\mathcal{N}(x_t; \sqrt{1 - \beta_t} \cdot x_{t-1}, \beta_t \cdot \mathbf{I})$  中采样数据点，其中  $\beta_t \ll 1$ 。该迭代采样逐渐破坏数据中的信息，并将其替换为高斯噪声。采样过程通过重参数化技巧来说明（详见第 2.1 节）。DDPM 的反向过程也从一个正态分布中进行迭代采样，但是分布的均值  $\mu_\theta(x_t, t)$  是通过从前一步骤的图像中减去由神经网络估计得到的噪声来导出的。方差与正向过程中使用的方差相等。进入反向过程的初始图像只包含高斯噪声。NCSN 的正向过程只是将正常噪声添加到前一步的图像中。这也可看作是从一个正态分布中采样  $\mathcal{N}(x_t; x_{t-1}, (\sigma_t^2 - \sigma_{t-1}^2) \cdot \mathbf{I})$ ，其中均值是前一步的图像。NCSN 的反向过程基于第 2.2 节中描述的一种算法。最佳观看效果需使用彩色版。

这些简化（更多细节见附录 ??）解锁了目标函数  $\mathcal{L}_{vib}$  的新表达式，该表达式用于衡量正向过程的随机时间步长  $t$  上真实噪声  $z_t$  与模型的噪声估计  $z_\theta(x_t, t)$  之间的距离。

$$\mathcal{L}_{simple} = \mathbb{E}_{t \sim [1, T]} \mathbb{E}_{x_0 \sim p(x_0)} \mathbb{E}_{z_t \sim \mathcal{N}(0, \mathbf{I})} \|z_t - z_\theta(x_t, t)\|^2, \quad (6)$$

其中  $\mathbb{E}$  表示期望值， $z_\theta(x_t, t)$  是预测  $x_t$  中的噪声的神经网络。需要强调的是， $x_t$  是通过式 (3) 进行采样得到的，其中我们使用训练集中的一个随机图像  $x_0$ 。

生成过程仍然由  $p_\theta(x_{t-1}|x_t)$  定义，但神经网络不直接预测均值和协方差。相反，它被训练来预测图像中的噪声，并且均值根据式 (5) 确定，而协方差被固定为一个常数。算法 1 规范了整个生成过程。

## 2.2 Noise Conditioned Score Networks (NCSNs)

某些数据密度  $p(x)$  的得分函数被定义为相对于输入的对数密度的梯度，即  $\nabla_x \log p(x)$ 。这些梯度所指示的方向被 Langevin 动力学算法 [?] 用于从随机样本 ( $x_0$ ) 向具有高密度区域的样本 ( $x_N$ ) 移动。Langevin 动力学是受物理启发的一种迭代方法，可用于数据抽样。在物理学中，该方法用于确定分子系统中的粒子轨迹，该系统允许粒子与其他分子之间的相互作用。粒子的轨迹受系统的阻力和由分子之间的快速

---

### Algorithm 1 DDPM 抽样方法

---

#### Input:

$T$  – the number of diffusion steps.

$\sigma_1, \dots, \sigma_T$  – the standard deviations for the reverse transitions.

---

#### Output:

$x_0$  – the sampled image.

---

#### Computation:

- 1:  $x_T \sim \mathcal{N}(0, \mathbf{I})$
  - 2: **for**  $t = T, \dots, 1$  **do**
  - 3:     **if**  $t > 1$  **then**
  - 4:          $z \sim \mathcal{N}(0, \mathbf{I})$
  - 5:     **else**
  - 6:          $z = 0$
  - 7:      $\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \beta_t}} \cdot z_\theta(x_t, t) \right)$
  - 8:      $x_{t-1} = \mu_\theta + \sigma_t \cdot z$
- 

相互作用引起的随机力的影响。在我们的情况下，我们可以将对数密度的梯度视为一种将随机样本从数据空间拖入具有高数据密度  $p(x)$  的区域的力。物理学中的另一项  $\omega_i$  代表随机

力, 但对我们来说, 它有助于逃脱局部最小值。最后,  $\gamma$  的值衡量两种力的影响, 因为它代表粒子所处环境的摩擦系数。从抽样观点来看,  $\gamma$  控制更新的大小。总结起来, Langevin 动力学的迭代更新如下:

$$x_i = x_{i-1} + \frac{\gamma}{2} \nabla_x \log p(x) + \sqrt{\gamma} \cdot \omega_i, \quad (7)$$

其中  $i \in \{1, \dots, N\}$ ,  $\gamma$  控制更新方向得分的幅度,  $x_0$  是从先验分布中采样得到的, 噪声  $\omega_i \sim \mathcal{N}(0, \mathbf{I})$  解决了陷入局部最小值的问题, 并且该方法递归地应用了  $N \rightarrow \infty$  步。因此, 生成模型可以利用上述方法, 在用神经网络  $s_\theta(x) \approx \nabla_x \log p(x)$  估计得分之后, 从  $p(x)$  中进行采样。可以通过评分匹配方法训练此网络, 该方法需要最优化以下目标函数:

$$\mathcal{L}_{sm} = \mathbb{E}_{x \sim p(x)} \|s_\theta(x) - \nabla_x \log p(x)\|_2^2. \quad (8)$$

在实践中, 直接最小化这个目标函数是不可能的, 因为  $\nabla_x \log p(x)$  是未知的。然而, 还有其他方法, 如去噪得分匹配 [?] 和切片得分匹配 [?] 可以克服这个问题。

虽然上述方法可用于数据生成, 但 Song 等人 [?] 强调了在实际数据上应用该方法时存在的一些问题。大部分问题都与流形假设有关。例如, 当数据存在于低维流形上时, 得分估计  $s_\theta(x)$  是不一致的, 而且可能导致 Langevin 动力学无法收敛到高密度区域。在同一篇工作中 [?], 作者证明了这些问题可以通过在不同尺度上用高斯噪声扰动数据来解决。此外, 他们提出通过单个噪声条件的得分网络(NCSN)来学习生成噪声分布的得分估计。关于采样, 他们采用了等式(7)中的策略, 并使用与每个噪声尺度相关的得分估计。

形式上, 给定一个高斯噪声尺度的序列  $\sigma_1 < \sigma_2 < \dots < \sigma_T$ , 满足  $p_{\sigma_1}(x) \approx p(x_0)$  和  $p_{\sigma_T}(x) \approx \mathcal{N}(0, \mathbf{I})$ , 我们可以通过去噪得分匹配训练一个 NCSN  $s_\theta(x, \sigma_t)$ , 使得  $s_\theta(x, \sigma_t) \approx \nabla_x \log(p_{\sigma_t}(x))$ , 对于所有  $t \in \{1, \dots, T\}$ 。我们可以按以下方式推导  $\nabla_x \log(p_{\sigma_t}(x))$ :

$$\nabla_{x_t} \log p_{\sigma_t}(x_t | x) = -\frac{x_t - x}{\sigma_t^2}, \quad (9)$$

given that:

$$\begin{aligned} p_{\sigma_t}(x_t | x) &= \mathcal{N}(x_t; x, \sigma_t^2 \cdot \mathbf{I}) \\ &= \frac{1}{\sigma_t \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{x_t - x}{\sigma_t}\right)^2\right), \end{aligned} \quad (10)$$

其中,  $x_t$  是  $x$  的加噪版本,  $\exp$  是指数函数。因此, 将公式(8)推广到所有  $(\sigma_t)_{t=1}^T$ , 并使用公式(9)中的形式替换梯度, 通过最小化以下目标来训练  $s_\theta(x_t, \sigma_t)$ , 对于所有  $t \in \{1, \dots, T\}$ :

$$\mathcal{L}_{dsm} = \frac{1}{T} \sum_{t=1}^T \lambda(\sigma_t) \mathbb{E}_{p(x)} \mathbb{E}_{x_t \sim p_{\sigma_t}(x_t | x)} \left\| s_\theta(x_t, \sigma_t) + \frac{x_t - x}{\sigma_t^2} \right\|_2^2, \quad (11)$$

其中  $\lambda(\sigma_t)$  是一个加权函数。训练完成后, 神经网络  $s_\theta(x_t, \sigma_t)$  将返回在给定噪声图像  $x_t$  和对应时间步  $t$  的条件下估计得到的得分  $\nabla_{x_t} \log(p_{\sigma_t}(x_t))$ 。

在推理阶段, Song 等人 [?] 引入了退火 Langevin 动力学,

---

**Algorithm 2** 退火朗之万动力学

---

**Input:**

$\sigma_1, \dots, \sigma_T$  – a sequence of Gaussian noise scales.

$N$  – the number of Langevin dynamics iterations.

$\gamma_1, \dots, \gamma_T$  – the update magnitudes for each noise scale.

---

**Output:**

$x_0^0$  – the sampled image.

---

**Computation:**

- 1:  $x_T^0 \sim \mathcal{N}(0, \mathbf{I})$
  - 2: **for**  $t = T, \dots, 1$  **do**
  - 3:     **for**  $i = 1, \dots, N$  **do**
  - 4:          $\omega \sim \mathcal{N}(0, \mathbf{I})$
  - 5:          $x_t^i = x_t^{i-1} + \frac{\gamma_t}{2} \cdot s_\theta(x_t^{i-1}, \sigma_t) + \sqrt{\gamma_t} \cdot \omega$
  - 6:      $x_{t-1}^0 = x_t^N$
- 

其形式化描述如算法2所示。他们的方法从白噪声开始, 并对固定的迭代次数应用公式(7)。训练得到的神经网络根据时间步  $T$  得到所需的梯度(得分)。该过程继续进行下一个时间步, 将一步的输出作为下一步的输入。最终样本是针对  $t = 0$  返回的输出。

### 2.3 Stochastic Differential Equations (SDEs)

与之前两种方法类似, [?] 中提出的方法逐渐将数据分布  $p(x_0)$  转化为噪声。然而, 它在之前两种方法的基础上进行了推广, 因为在该方法中, 扩散过程被认为是连续的, 从而成为随机微分方程(SDE)的解。正如 [?] 所示, 这种扩散的逆向过程可以用逆时间 SDE 进行建模, 该 SDE 在每个时间步骤中需要密度的得分函数。因此, Song 等人 [?] 的生成模型采用神经网络来估计得分函数, 并通过数值 SDE 求解器从  $p(x_0)$  中生成样本。与 NCSNs 的情况类似, 神经网络接收扰动数据和时间步长作为输入, 并生成得分函数的估计。

正向扩散过程  $(x_t)_{t=0}^T$  的 SDE, 其中  $t \in [0, T]$ , 具有以下形式:

$$\frac{\partial x}{\partial t} = f(x, t) + \sigma(t) \cdot \omega_t \iff \partial x = f(x, t) \cdot \partial t + \sigma(t) \cdot \partial \omega, \quad (12)$$

其中  $\omega_t$  是高斯噪声,  $f$  是一个关于  $x$  和  $t$  的函数, 用于计算漂移系数,  $\sigma$  是一个关于时间的函数, 用于计算扩散系数。为了使扩散过程成为这个随机微分方程的解, 漂移系数应该被设计成逐渐消除数据  $x_0$ , 而扩散系数控制添加了多少高斯噪声。相关的逆时间随机微分方程(Reverse-time SDE) [?] 定义如下:

$$\partial x = \left[ f(x, t) - \sigma(t)^2 \cdot \nabla_x \log p_t(x) \right] \cdot \partial t + \sigma(t) \cdot \partial \hat{\omega}, \quad (13)$$

其中,  $\hat{\omega}$  表示时间反转时的布朗运动, 从  $T$  到  $0$ 。反向时间 SDE 表明, 如果我们从纯噪声开始, 通过消除导致数据破坏的漂移, 我们可以恢复数据。通过减去  $\sigma(t)^2 \cdot \nabla_x \log p_t(x)$  来执行去除操作。

**Algorithm 3** 欧拉-丸山采样方法**Input:**

$\Delta t < 0$  – a negative step close to 0.

$f$  – a function of  $x$  and  $t$  that computes the drift coefficient.

$\sigma$  – a time-dependent function that computes the diffusion coefficient.

$\nabla_x \log p_t(x)$  – the (approximated) score function.

$T$  – the final time step of the forward SDE.

**Output:**

$x$  – the sampled image.

**Computation:**

- 1:  $t = T$
- 2: **while**  $t > 0$  **do**
- 3:    $\Delta x = [f(x, t) - \sigma(t)^2 \cdot \nabla_x \log p_t(x)] \cdot \Delta t + \sigma(t) \cdot \Delta \hat{\omega}$
- 4:    $x = x + \Delta x$
- 5:    $t = t + \Delta t$

我们可以通过优化与Eq. (11)中相同的目标来训练神经网络  $s_\theta(x, t) \approx \nabla_x \log p_t(x)$ , 但针对连续情况进行适应, 如下所示:

$$\begin{aligned} \mathcal{L}_{dsm}^* = \\ = \mathbb{E}_t [\lambda(t) \mathbb{E}_{p(x_0)} \mathbb{E}_{p_t(x_t|x_0)} \|s_\theta(x_t, t) - \nabla_x \log p_t(x_t|x_0)\|_2^2], \end{aligned} \quad (14)$$

其中  $\lambda$  是一个加权函数,  $t \sim \mathcal{U}([0, T])$ 。需要强调的是, 当漂移系数  $f$  是仿射时,  $p_t(x_t|x_0)$  是一个高斯分布。当  $f$  不符合这个性质时, 我们不能使用去噪得分匹配, 但可以退而求其次使用切片得分匹配 [?].

对于这种方法, 可以使用在Eq. (13)定义的随机微分方程组上应用任何数值方法进行采样。在实践中, 求解器不能使用连续的形式。例如, 欧拉-马鲁雅马方法固定一个微小的负步长  $\Delta t$  并执行算法 3 直到初始时间步长  $t = T$  变为  $t = 0$ 。在步骤3中, 布朗运动由  $\Delta \hat{\omega} = \sqrt{|\Delta t|} \cdot z$  给出, 其中  $z \sim \mathcal{N}(0, \mathbf{I})$ 。

Song等人 (Song *et al.*, 2021年) 在采样技术方面提出了几个贡献。他们引入了预测-校正采样器, 该采样器生成更好的示例。这个算法首先采用数值方法从反向时间SDE中采样, 然后使用基于得分的方法作为校正器, 例如在前一小节中描述的退火朗之万动力学。此外, 他们还展示了常微分方程 (ODEs) 也可以用来建模反向过程。因此, 由SDE解释提供的另一种采样策略是基于应用于ODE的数值方法。后一种策略的主要优势是其效率。

## 2.4 Relation to Other Generative Models

我们在下面讨论了扩散模型与其他类型的生成模型之间的联系。我们从基于似然的方法开始, 然后介绍生成对抗网络。

扩散模型与VAEs [?] 在许多方面具有共同之处。例如, 在这两种情况下, 数据被映射到潜变量空间中, 生成过程学习将潜变量表示转化为数据。此外, 在这两种情况下, 目标函数都可以推导为数据似然的下界。然而, 这两种方法之间存在重要的差异, 接下来我们将提到其中的一些差异。VAE的潜变量表示包含有关原始图像的压缩信息, 而扩散模型则在正向过程的最后一步之后完全破坏数据。扩散模型的潜变量表示与原始数据具有相同的维度, 而在维度减少时, VAE的效果更好。最终, VAE的潜变量空间映射是可训练的, 而扩散模型的正向过程并非如此, 因为如前所述, 潜变量是通过逐渐向原始图像添加高斯噪声获得的。上述相似性和差异可能成为未来两种方法发展的关键。例如, 已经有一些研究将扩散模型应用于VAE的潜变量空间以构建更高效的扩散模型 [?], [?].

自回归模型 [?], [?] 将图像表示为像素序列。它们的生成过程通过在先前生成的像素的条件下逐像素生成图像的新样本。这种方法意味着存在单向偏差, 明显是这类生成模型的一个限制。Esser等人 [?] 将扩散模型和自回归模型视为互补, 并解决了上述问题。他们的方法通过马尔可夫链学习倒置多项式扩散过程, 其中每个转移被实现为自回归模型。自回归模型接收的全局信息由马尔可夫链的上一步提供。

归一化流 [?], [?] 是一类将简单的高斯分布转化为复杂数据分布的生成模型。这种转化通过一组可逆函数完成, 这些函数具有易于计算的雅可比行列式。这些条件在实践中转化为架构限制。这种模型的一个重要特点是似然函数是易于计算的。因此, 训练的目标是负对数似然。与扩散模型相比, 这两种模型共同将数据分布映射到高斯噪声中。然而, 这两种方法之间的相似性就此结束, 因为归一化流通过学习一个可逆且可微的函数以确定性方式执行映射。与扩散模型相比, 这些特性意味着网络架构上的额外约束和可学习的正向过程。一个连接这两种生成算法的方法是DiffFlow。在 [?] 中引入的DiffFlow扩展了扩散模型和归一化流, 使得正向和反向过程都是可训练和随机的。基于能量的模型 (Energy-based models, EBMs) [?], [?], [?], [?] 专注于提供密度函数的非标准化版本, 即能量函数的估计。得益于这一特性, 与之前的基于似然的方法相比, 这种类型的模型可以用任何回归神经网络表示。然而, 由于这种灵活性, EBMs的训练是困难的。实践中常用的一种流行的训练策略是分数匹配 (score matching) [?], [?]. 关于抽样, 除了其他策略, 还有基于得分函数的马尔可夫链蒙特卡洛 (Markov Chain Monte Carlo, MCMC) 方法。因此, 扩散模型 (diffusion models) 的第2.2小节中的表述可以被认为是能量模型框架的一个特例, 准确地说, 这是当训练和抽样仅需要得分函数时的情况。

GAN (生成对抗网络) [?] 被许多人视为在生成样本的质量方面的最新技术, 直到最近才出现了扩散模型的崛起 [?]. 由于其对抗性目标, GAN的训练被认为是困难的 [?], 并且经常出现模式崩溃的问题。相比之下, 扩散模型具有稳定的

训练过程，并且因为基于似然，提供更多的多样性。尽管这些优点，与GAN相比，扩散模型仍然效率较低，在推理过程中需要进行多次网络评估。GAN和扩散模型之间的一个关键方面是它们的潜空间。GAN具有低维潜空间，而扩散模型保留了图像的原始大小。此外，扩散模型的潜空间通常被建模为随机高斯分布，与变分自编码器（VAEs）类似。在语义属性方面，研究发现GAN的潜空间中包含与视觉属性相关的子空间 [?]. 由于这一特性，可以通过改变潜空间来操作属性 [?], [?]. 相比之下，当希望对扩散模型进行这种转换时，首选的技术是引导技术 [?], [?], 它不利用潜空间的任何语义属性。然而，Song等人 [?] 证明了扩散模型的潜空间具有明确定义的结构，并且插值在该空间中导致图像空间的插值。总结来说，从语义角度来看，扩散模型的潜空间探索的比GAN要少得多，但这可能是未来研究方向之一。

### 3 A CATEGORIZATION OF DIFFUSION MODELS

我们根据多个角度的划分标准将扩散模型分类。分离模型最重要的标准可以通过以下几点来定义：(i) 它们应用于的任务，以及(ii) 它们所需的输入信号。此外，由于制定扩散模型的方法存在多种途径，(iii) 底层框架是分类扩散模型的另一个关键因素。最后，(iv) 在训练和评估过程中使用的数据集也非常重要，因为它们提供了对同一任务下不同模型进行比较的手段。我们根据上述标准对扩散模型进行了分类，分类结果见表格1。

在本节的剩余部分，我们按照目标任务作为主要标准对扩散模型进行了几个贡献的介绍。我们选择了这个分类标准，因为它相当平衡且代表了扩散模型研究中的相关工作，可以帮助专注于特定任务的读者迅速了解相关工作。尽管主要任务通常与图像生成有关，但相当多的工作已经在其他主题上进行，甚至在超分辨率、修补、图像编辑、图像到图像的转换或分割等方面超过了GAN的性能。

#### 3.1 Unconditional Image Generation

下面介绍的扩散模型用于在无条件环境下生成样本。这些模型不需要监督信号，完全是无监督的。我们认为这是图像生成最基本和通用的设置。

##### 3.1.1 Denoising Diffusion Probabilistic Models

Sohl-Dickstein等人的工作 [?] 在第2.1节中形式化了扩散模型。所提出的神经网络基于包含多尺度卷积的卷积结构。

Austin等人的工作 [?] 将Sohl-Dickstein等人 [?] 的方法扩展到离散扩散模型，研究了在前进过程中使用的转移矩阵的不同选择。他们的结果在图像生成任务上与先前的连续扩散模型竞争力相当。

Ho等人扩展了 [?] 中提出的工作，提出通过估计每个步骤中图像中的噪声来学习逆过程。这个改变导致了一个类似

于 [?] 中应用的去噪评分匹配的目标函数。为了预测图像中的噪声，作者使用了PixelCNN++结构，该结构在 [?] 中引入。

在Ho等人的工作基础上，Nichol等人 [?] 引入了一些改进，观察到对于低分辨率，线性噪声调度是次优的。他们提出了一个新的选项，以避免在前向过程末尾的快速信息破坏。此外，他们还表明，为了改善扩散模型在对数似然性能方面，需要学习方差。这个最后的改变使采样更快，大约需要50步。

Song等人 [?] 将 [?] 中使用的马尔可夫前进过程替换为非马尔可夫过程。生成过程发生改变，模型首先预测正常样本，然后用于估计链中的下一步。这个改变导致采样过程更快，对生成样本的质量影响较小。所得到的框架被称为去噪扩散隐式模型（DDIM）。

Sinha等人的工作 [?] 提出了具有对比表示的扩散解码模型（D2C），这是一种基于编码器生成的潜在表示训练扩散模型的生成方法。这个框架基于 [?] 中提出的DDPM结构，通过将潜在表示映射到图像来生成图像。

在 [?] 中，作者提出了一种在推理时估计噪声参数的方法。他们的改变在改善Fréchet Inception Distance (FID) 的同时，需要更少的步骤。作者使用VGG-11来估计噪声参数，并使用DDPM [?] 生成图像。

Nachmani等人的工作 [?] 建议用两个其他分布替换扩散过程中的高斯噪声分布，即两个高斯混合分布和Gamma分布。结果表明，由于Gamma分布具有更高的建模能力，因此具有更好的FID值和更快的收敛速度。Lam等人 [?] 学习了采样的噪声调度。训练的噪声调度保持线性不变，训练得到的分数网络假设与最优值接近，以便用于噪声调度的训练。推断分为两个步骤。首先，通过固定两个初始超参数确定调度。第二步是使用确定的调度进行的正常反向过程。

Bond-Taylor等人 [?] 提出了一个两阶段的过程，他们对图像应用向量量化以获得离散表示，并使用transformer [?] 来反转离散扩散过程，其中元素在每个步骤中都被随机屏蔽。采样过程更快，因为扩散应用于高度压缩的表示，可减少降噪步骤（50-256）。

Watson等人 [?] 提出了一个动态规划算法，用于找到最优的推断调度，其时间复杂度为  $\mathcal{O}(T)$ ，其中  $T$  是步骤数。他们在CIFAR-10和ImageNet上进行了图像生成实验，使用了DDPM架构。

在另一项的工作中，Watson等人 [?] 首先展示了如何将一个再参数化技巧集成到扩散模型的反向过程中，以优化一系列快速采样器。他们使用核感知距离作为损失函数，展示了如何使用随机梯度下降进行优化。接下来，他们提出了一种特殊的参数化采样器族，使用与之前相同的过程，可以使用更少的采样步骤得到具有竞争力的结果。使用FID和Inception Score (IS) 作为度量标准，该方法似乎胜过了一些扩散模型基准。

类似于Bond-Taylor等人 [?] 和Watson等人 [?], [?],

Xiao等人 [?] 尝试提高采样速度的同时保持样本的质量、覆盖率和多样性。他们的方法是在降噪过程中整合生成对抗网络 (GAN)，以区分真实样本 (正向过程) 和伪造样本 (来自生成器的降噪样本)，目标是最小化软化反向KL散度 [?]. 然而，通过直接生成一个干净 (完全降噪) 的样本，并将伪造样本与其条件化，修改了模型。使用NCSN++架构和自适应分组归一化层的GAN生成器，他们在图像合成和基于笔画的图像生成上实现了相似的FID值，采样速率比其他扩散模型快20到2000倍。

Kingma等人 [?] 介绍了一类扩散模型，对图像密度估计获得了最先进的似然度。他们在网络的输入中增加了傅里叶特征以预测噪声，并调查了观察到的改进是否特定于这类模型。他们的结果证实了这个假设，即之前的最先进模型没有从这个改变中受益。作为一个理论贡献，他们证明了扩散损失只通过极端值受到信噪比函数的影响。

在 [?] 中提到的工作之后，Bao等人 [?] 提出了一个推断框架，不需要使用非马尔科夫扩散过程进行训练。通过首先推导出关于得分函数的最优均值和方差的解析估计，并使用预训练的基于得分的模型获得得分值，他们展示了更好的结果，同时节省了20到40倍的时间。得分通过蒙特卡洛采样来近似。然而，为了减小预训练DDPM模型的任何偏差，得分被剪切在一些预计计算的边界内。郑等人 [?] 提出在任意步骤截断该过程，并提出了一种通过放松正向扩散的最终输出为高斯随机噪声的约束来逆转扩散的方法。为了解决从不可计算的分布开始进行逆向过程的问题，使用了一个隐式生成分布来匹配扩散数据的分布。代理分布通过生成对抗网络 (GAN) 或条件转移进行拟合。值得注意的是，生成器使用与扩散模型采样器相同的U-Net模型，因此不会增加额外的需要训练的参数。

Deja等人 [?] 通过分析扩散模型的逆向过程，假设它由两个模型组成，一个生成器和一个去噪器。因此，他们建议将过程明确地分为两个组件：通过自编码器实现去噪器，通过扩散模型实现生成器。两个模型都使用相同的U-Net架构。

Wang等人 [?] 基于Arjovsky等人 [?] 和 Sønderby等人 [?] 提出的想法，通过向鉴别器的输入数据添加噪声来增强鉴别能力。在 [?] 中，通过从多个时间步骤的原始图像中加权扩散样本的高斯混合分布中注入噪声来实现这一目标。噪声注入机制应用于真实图像和伪造图像。实验在多个分辨率和高度多样化的数据集上进行。

### 3.1.2 Score-Based Generative Models

从之前的研究中开始 [?], Song等人 [?] 在理论和实证分析的基础上提出了几项改进。他们旨在解决训练和采样阶段的问题。在训练方面，作者展示了选择噪声尺度和将噪声条件纳入NCSNs [?] 的新策略。对于采样，他们提出将参数应用指数移动平均，并选择兰维因动力学的超参数，以使步长满足

某个方程。所提出的改变拓展了NCSNs在高分辨率图像上的应用。

Jolicoeur-Martineau等人 [?] 引入了对抗目标以及去噪分数匹配来训练基于分数的模型。此外，他们提出了一种新的采样过程，称为一致退火采样，并证明了其比退火兰维方法更加稳定。他们的图像生成实验表明，新的目标返回了更高质量的示例，并且对多样性没有影响。所建议的修改在 [?], [?], [?] 中提出的架构上进行了测试。

Song等人 [?] 通过新的加权函数改进了基于分数的扩散模型的似然性。他们通过组合分数匹配损失来实现这一点。对于他们的图像生成实验，他们采用了在 [?] 中介绍的DDPM++架构。

在 [?] 中，作者将基于分数的生成模型作为迭代比例拟合 (IPF) 的实现来介绍，IPF是一种用于解决Schrödinger桥问题的技术。这种新颖的方法被应用于图像生成和数据集插值，这是可能的，因为先验可以是任意分布。

Vahdat等人 [?] 在潜在表示上训练扩散模型。他们使用一个VAE对潜在空间进行编码和解码。这项工作实现了最多56倍的快速采样。对于图像生成实验，作者采用了在 [?] 中提出的NCSN++架构。

### 3.1.3 Stochastic Differential Equations

DiffFlow在 [?] 中作为一种将归一化流和扩散概率模型结合的新的生成建模方法被介绍。从扩散模型的角度来看，该方法具有一个采样过程，由于可学习的向前过程跳过了不需要的噪声区域，因此效率提高了多达20倍。作者使用与 [?] 中相同的架构进行实验。

Jolicoeur-Martineau等人 [?] 引入了一种新的SDE求解器，其速度比欧拉-马鲁雅马(SDE)快2到5倍，且不影响生成图像的质量。该求解器在一组图像生成实验中对来自 [?] 的预训练模型进行了评估。

Wang等人 [?] 提出了一种基于Schrödinger桥的新的深度生成模型。这是一个两阶段的方法，第一阶段学习目标分布的平滑版本，第二阶段得出实际的目标分布。

Dockhorn等人 [?] 针对基于评分的模型，通过在数据中添加另一个变量(速度)来利用临界阻尼Langevin扩散过程，这是过程中唯一的噪声来源。鉴于新的扩散空间，得到的评分函数被证明更容易学习。作者通过开发一种更合适的混合评分匹配(score matching)目标和通过积分求解SDE的采样方法来扩展他们的工作。作者将NCSN++和DDPM++架构调整为同时接受数据和速度，并在无条件图像生成方面进行评估，在性能上优于类似的基于评分的扩散模型。

Deasy等人 [?] 受到高维基于评分的扩散模型的限制，由于高斯噪声分布的影响，将去噪评分匹配扩展到了一般的噪声分布。通过添加尾部更重的分布，他们在几个数据集上的实验显示出有希望的结果，因为在某些情况下(取决于分布的

形状), 生成性能得到了改善。该方法在类别不平衡的数据集上表现出色。

Jing等人 [?]试图通过缩小扩散实现的空间, 即扩散过程中时间步长越大, 子空间越小, 来缩短扩散模型的采样过程的持续时间。数据被投影到一组有限的子空间上, 在特定的时间点处, 每个子空间与一个评分模型相关联。这样可以降低计算成本, 同时提高性能。该工作仅限于自然图像合成。通过在无条件图像生成中评估该方法, 作者在推理时间较短的同时实现了与最先进模型相似或更好的性能。该方法也被证明在修复图像任务中有效。

Kim等人 [?]提出将扩散过程改变为非线性过程。这是通过使用可训练的归一化流模型对图像进行编码, 使其可以线性地扩散到噪声分布。然后, 在去噪过程中应用类似的逻辑。这个方法应用于NCSN++和DDPM++框架, 而归一化流模型基于ResNet。

Ma等人 [?]旨在使反向扩散过程更加高效, 同时保持综合性能。在基于评分的扩散模型家族中, 他们开始在频率域中分析反向扩散, 随后在采样过程中应用空间频率滤波器, 将目标分布的信息整合到初始噪声采样中。作者在NCSN [?]和NCSN++ [?]上进行了实验证明, 该方法在图像合成方面能够明显提高速度(采样步骤减少了多达20倍), 同时对于低分辨率和高分辨率图像的生成质量保持相同的满意度。

### 3.2 Conditional Image Generation

我们接下来展示应用于条件图像合成的扩散模型。条件通常基于各种源信号, 大多数情况下使用一些类别标签。一些方法同时进行无条件和条件生成, 这些也在本文中讨论。

#### 3.2.1 Denoising Diffusion Probabilistic Models

Dhariwal等人 [?]提出了一些架构改进方法来改善扩散模型的FID。他们还提出了分类器引导策略, 该策略使用分类器的梯度来指导采样过程中的扩散。他们进行了无条件和有条件的图像生成实验。

Bordes等人 [?]通过可视化和比较自监督任务生成的表示与原始图像来检查所得到的表示。他们还比较了来自不同源的表示。因此, 扩散模型用于在这些表示的条件下生成样本。作者对Dhariwal等人 [?]提出的U-Net架构进行了一些修改, 例如添加有条件批归一化层, 并通过全连接层映射矢量表示。

[?]中提出的方法使扩散模型能够从数据流形的低密度区域产生图像。他们使用两个新的损失来引导逆过程。第一个损失将扩散引导到低密度区域, 而第二个损失则强制扩散保持在流形上。此外, 他们证明了他们的扩散模型不会记住低密度邻域的示例, 从而生成新的图像。作者采用了与Dhariwal等人 [?]类似的架构。

Kong等人 [?]定义了连续扩散步骤和噪声水平之间的双射关系。通过定义的双射关系, 他们能够构建一个需要更

少步骤的近似扩散过程。该方法在图像生成的先前DDIM [?]和DDPM [?]架构上进行了测试。

Pandey等人 [?]构建了一个生成器-细化器框架, 其中生成器是一个VAE, 细化器是由VAE的输出条件化的DDPM。VAE的潜空间可以用来控制生成图像的内容, 因为DDPM只添加细节。在训练框架后, 所得到的DDPM能够适应不同的噪声类型。更具体地说, 如果逆过程不是在VAE的输出上进行条件, 而是在不同的噪声类型上进行条件, 那么DDPM能够重构初始图像。

Ho等人 [?]提出了级联扩散模型 (CDM), 一种用于在ImageNet类条件下生成高分辨率图像的方法。他们的框架包含多个扩散模型, 其中流水线的第一个模型生成以图像类为条件的低分辨率图像。随后的模型负责生成分辨率越来越高的图像。这些模型既受到类别的条件限制, 也受到低分辨率图像的条件限制。

Benny等人 [?]研究了在逆过程中预测图像而不是噪声的优势和劣势。他们得出结论, 通过插值这两种类型的输出可以解决一些发现的问题。他们修改了先前的架构, 以返回噪声和图像以及控制插值过程中噪声重要性的值。该策略在DDPM和DDIM架构的基础上进行了评估。

Choi等人 [?]研究了噪声水平对扩散模型学到的视觉概念的影响。他们将传统的目标函数加权方案修改为一种新的方案, 以促使扩散模型学习丰富的视觉概念。该方法根据信噪比将噪声水平分为三个类别 (粗糙、内容和清除), 即小信噪比为粗糙, 中等信噪比为内容, 大信噪比为清除。加权函数对最后一组分配较低的权重。

Singh等人 [?]提出了一种新的有条件的图像生成方法。他们不是在整个采样过程中对信号进行条件, 并提出了一种将噪声信号 (采样起始点) 进行条件的方法。使用Inverting Gradients [?], 噪声注入了与条件类别的定位和方向相关的信息, 同时保持相同的随机高斯分布。

Liu等人 [?]描述了扩散模型和能量模型的相似功能, 并利用后者模型的组合结构, 提出了将多个扩散模型用于有条件的图像合成的方法。在逆过程中, 通过合取或否定可以实现多个与不同条件相关联的扩散模型的组合。

#### 3.2.2 Score-Based Generative Models

Song等人 (2021) 和Dhariwal等人 (2021) 关于基于分类器引导的得分型条件扩散模型的研究启发了Chao等人 (2022) 提出了一种新的训练目标, 该目标减小了得分模型与真实得分之间的潜在差异。分类器的损失函数被修改为添加了一个经过缩放的交叉熵项, 并结合了一种修改后的得分匹配损失。

#### 3.2.3 Stochastic Differential Equations

Ho等人 (2021) 介绍了一种不需要分类器的引导方法。它只需要一个条件扩散模型和一个无条件版本, 但他们使用同一

个模型来学习这两种情况。无条件模型的训练中，类别标识符被设为0。这个想法基于从贝叶斯规则推导出的隐式分类器。

Liu等人（2022）研究了使用传统数值方法解决反向过程的ODE形式。他们发现与之前的方法相比，这些方法返回的样本质量较低。因此，他们引入了扩散模型的伪数值方法。他们的想法将数值方法分为两部分，梯度部分和转移部分。将转移部分（标准方法具有线性转移部分）替换为尽可能接近目标流形的结果。最后一步是展示了这个改变如何解决使用传统方法时遇到的问题。

Tachibana等人（2021）解决了DDPMs的慢采样问题。他们提议通过增加随机微分方程求解器（去噪部分）的阶数（从一到二）来减少采样步骤的数量。在保留网络架构和得分匹配函数的同时，他们采用了Itô-Taylor展开方案进行采样，并替代了一些导数项以简化计算。他们减少了反向步骤的数量，同时保持性能。此外，他们的另一个贡献是新的噪声时间表。

Karras等人（2022）试图将基于扩散得分的模型分离为相互独立的组件。这种分离允许修改单个组件而不影响其他单元，从而方便改进扩散模型。利用这个框架，作者首先提出了一种采样过程，该过程将Heun方法作为ODE求解器，减少了神经函数的评估，同时保持了FID得分。他们进一步展示了随机采样过程带来的巨大性能优势。第二个贡献与通过在输入和相应目标上对神经网络进行预处理以及使用图像增强来训练基于得分的模型相关。

在无条件和有条件图像生成的背景下，Salimans等人（2022）提出了一种减少采样步骤数量的技术。他们将训练有的教师模型的知识（表示为确定性DDIM）提炼到具有相同架构但减半采样步骤的学生模型中。换句话说，学生的目标是执行教师的连续两步。此外，可以重复此过程直到达到所需的采样步骤数，同时保持相同的图像合成质量。最后，为了促进提炼过程并减少采样步骤的数量（从8192减少到4），探索了三个模型版本和两个损失函数。

Campbell等人（2022）展示了一种连续时间形式的去噪扩散模型，可以处理离散数据。该工作通过转移速率矩阵对前向连续时间马尔可夫链扩散过程进行建模，并通过逆转移速率矩阵的参数化逼近对后向去噪过程进行建模。进一步的贡献与训练目标、矩阵构造和优化采样器相关。

Song等人（2021）提出的将扩散模型解释为ODE的观点被Lu等人（2022）以能够使用指数积分器求解的形式重新表述。Lu等人（2022）的其他贡献是使用Taylor展开（从一阶到三阶）近似新形式的积分项的ODE求解器，以及自适应时间步调度的算法，速度是原来的4到16倍。

### 3.3 Image-to-Image Translation

Saharia等人（?）提出了一个使用扩散模型进行图像到图像翻译的框架，着重研究了四个任务：上色、修复、取消裁剪

和JPEG恢复。所提出的框架在这四个任务中是相同的，也就是说它不需要为每个任务进行定制化的改变。作者首先对 $L_1$ 和 $L_2$ 损失进行了比较，认为 $L_2$ 更好，因为它可以获得更高的样本多样性。最后，他们再次确认了自注意力层在条件图像合成中的重要性。

Sasaki等人（?）提出了一种涉及两个联合训练的扩散模型的方法来翻译一组不成对的图像。在逆向降噪过程中，每个模型在每一步都会根据另一个模型的中间样本进行条件训练。此外，扩散模型的损失函数使用循环一致性损失（?）进行了规范化。

Zhao等人（?）的目标是利用与源域等重要的数据来改进当前基于分数的图像到图像翻译扩散模型。他们使用在源域和目标域上进行训练的基于能量的函数来指导SDE求解器。这样可以生成保留无关域特征的图像，同时将源域特定的特征转换成目标域特定的特征。能量函数基于两个特定于域的特征提取器。

Wang等人（?）利用预训练的强大能力，采用GLIDE模型（?）并对其进行训练以获得丰富的语义潜空间。从预训练版本开始，将头部替换以适应任何条件输入，然后在特定的图像生成下游任务上进行微调。这分为两个步骤进行，第一步是冻结解码器，只训练新的编码器，第二步是同时训练它们。最后，作者采用对抗训练，并对无分类器指导进行归一化，以提高生成质量。

Li等人（?）引入了一种基于布朗桥和GAN的图像到图像翻译扩散模型。所提出的过程首先利用VQ-GAN（?）对图像进行编码。在产生的量化潜空间内，扩散过程被形式化为布朗桥，将源域和目标域的潜在表示之间进行映射。最后，另一个VQ-GAN对量化向量进行解码，以在新的域中合成图像。这两个GAN模型分别在各自的域上进行独立训练。

Wolleb等人（?）在其先前工作（?）的基础上，通过将分类器替换为与任务特定的另一个模型来扩展扩散模型。因此，在采样过程的每一步中，注入了任务特定网络的梯度。这种方法可以使用回归器（基于编码器）或分割模型（使用U-Net架构）进行演示，而扩散模型基于现有的框架（?），（?）。这种设置的优点是除了任务特定模型外，不需要重新训练整个扩散模型。

### 3.4 Text-to-Image Synthesis

或许弥散模型最令人印象深刻的成果是在文本到图像合成方面取得的，通过将对象、形状和纹理等无关概念组合在一起生成不寻常的示例的能力得以展现。为了确认这一观点，我们使用了稳定弥散模型（?）根据各种文本提示生成图像，结果如图2所示。

Imagen是一种文本到图像合成方法，在（?）中介绍。它包含一个用于文本序列的编码器和一系列用于生成高分辨率图像的弥散模型级联。这些模型还受编码器返回的文本嵌入的影响。此外，作者还引入了一组新的标题（DrawBench）

用于文本到图像评估。对于架构方面，作者开发了高效U-Net以提高效率，并将此架构应用于其文本到图像生成实验中。

Gu等人 [?]引入了VQ-Diffusion模型，这是一种无单向偏差的文本到图像合成方法。通过其遮蔽机制，所提出的方法在推理过程中避免了误差的积累。该模型分为两个阶段，第一阶段基于VQ-VAE，通过离散令牌学习表示图像；第二阶段是在VQ-VAE的离散潜空间上运行的离散弥散模型。弥散模型的训练是基于标题嵌入的条件的。受到遮蔽语言建模的启发，一些令牌被替换为 $[mask]$ 令牌。

Avrahami等人 [?]提出了一种以CLIP [?]图像和文本嵌入为条件的文本条件弥散模型。这是一个两阶段的方法，第一阶段生成图像嵌入，第二阶段（解码器）根据图像嵌入和文本标题产生最终图像。为了生成图像嵌入，作者使用了在潜空间中的一个弥散模型。他们进行了主观人工评估来评估他们的生成结果。

针对弥散模型的慢速采样不便，Zhang等人 [?]将他们的工作聚焦在一种新的离散化方案上，该方案减小了误差并允许更大的步长，即更少的采样步骤。通过在评分函数中使用高阶多项式外推和指数积分器来解决反向随机微分方程，网络评估的数量大大减少，同时保持了生成能力。

Shi等人 [?]结合了VQ-VAE [?]和弥散模型来生成图像。从VQ-VAE开始，编码功能保留，而解码器被弥散模型替换。作者使用了 [?]中的U-Net架构，将图像令牌注入到中间块中。

在 [?]的基础上，Rombach等人 [?]介绍了一种修改方法，使用相同的过程创建艺术图像：从数据库中提取与CLIP [?]潜空间中的图像的k个最近邻，然后通过这些嵌入来引导逆去噪过程生成新图像。由于CLIP潜空间被文本和图像共享，弥散过程也可以由文本提示引导。然而，在推理时，数据库被另一个包含艺术图像的数据库替换。因此，模型在新数据库的风格中生成图像。

Jiang等人 [?]提出了一个框架，给定三个输入（人体姿势、衣服形状的文本描述和服装纹理的另一个文本），生成带有丰富服装表示的全身人体图像。方法的第一阶段将前一个文本提示编码为嵌入向量，并将其融入到生成形状图的模块（基于编码器-解码器）中。第二阶段，基于弥散的变换器从多个多级码书（每个码书特定于纹理）中采样后者文本提示的嵌入表示，该机制在VQ-VAE [?]中提出。最初，会对较粗粒度级别的码书索引进行采样，然后使用前馈网络预测较细粒度级别的索引。文本使用Sentence-BERT [?]进行编码。

### 3.5 Image Super-Resolution

Saharia等人 [?]将扩散模型应用于超分辨率。他们的逆过程学习生成基于低分辨率版本条件的高质量图像。本研究采用了 [?], [?]中提出的架构和以下数据集：CelebA-HQ、FFHQ和ImageNet。

Daniels等人 [?]使用基于分数的模型从两个分布的Sinkhorn耦合中采样。他们的方法利用神经网络建模了对偶变量，然后解决了最佳传输的问题。在训练完神经网络之后，可以通过兰格朗日动态和基于分数的模型进行采样。他们在图像超分辨率上运行了使用U-Net架构的实验。

### 3.6 Image Editing

Meng等人 [?]在各种引导图像生成任务中使用了扩散模型，例如绘画或基于笔画的编辑和图像合成。从包含某种形式引导的图像开始，图像的属性（如形状和颜色）会被保留下，而变形会通过逐渐添加噪声来平滑（扩散模型的正向过程）。然后，结果经过去噪处理（反向过程），根据引导创建出逼真的图像。图像是通过求解反向SDE，使用通用的扩散模型合成的，不需要任何自定义数据集或训练的修改。

[?]中介绍了一种基于自然语言描述编辑图像特定区域的方法。用户通过蒙版指定要修改的区域。该方法依赖于CLIP引导，根据文本输入生成图像，但是作者观察到，在最后将输出与原始图像结合起来时，并不能产生整体连贯的图像。因此，他们修改了去噪过程以解决这个问题。更确切地说，每一步之后，作者将蒙版应用于潜在图像，同时添加原始图像的带噪版本。

在 [?]中介绍的工作上进行扩展，Avrahami等人 [?]应用了潜在扩散模型来局部编辑图像，使用文本。VAE将图像和自适应时间蒙版（编辑区域）编码到潜在空间中，扩散过程在此发生。每个样本在被引导的兴趣区域内进行迭代去噪。然而，受到Blended Diffusion [?]的启发，图像与当前时间步的加噪掩膜区域在潜在空间中结合。最后，样本经过VAE解码生成新图像。该方法在性能上表现出优越性，同时具有较快的速度。

### 3.7 Image Inpainting

Nichol等人（2021）训练了一个基于文本描述的扩散模型，并研究了无分类器和基于CLIP的引导方法的效果。他们发现第一个选项获得了更好的结果。此外，他们对模型进行了微调，用于基于文本输入的图像修复。

Lugmay等人（2022）提出了一种与遮罩形式无关的修复方法。他们使用了一个无条件的扩散模型，并修改了其逆过程。他们通过从遮罩图像中采样已知区域，并对在第t步获得的图像应用去噪来生成第t-1步的图像。通过这个过程，作者观察到未知区域具有正确的结构，但语义不正确。为了解决这个问题，他们反复执行所提出的步骤多次，并在每次迭代中将从第t-1步生成的去噪版本中获得的新样本替换为第t步的先前图像。

### 3.8 Image Segmentation

Baranchuk等人 [?]展示了扩散模型在语义分割中的应用。他们从U-Net解码器中以不同尺度获取特征图（中间块），并将

它们串联起来（上采样特征图以使尺寸相同），然后通过附加多层感知机的集合来对每个像素进行分类。作者表明，提取的这些特征图在去噪过程的后期步骤中包含丰富的表示信息。实验证明，基于扩散模型的分割方法优于大部分基准方法。

Amit等人 [?] 提出了在扩展U-Net编码器的架构中使用扩散概率模型进行图像分割。输入图像和当前估计图像经过两个不同的编码器，并通过求和进行组合。然后将结果提供给U-Net的编码器-解码器。由于在每个时间步骤中注入了随机噪声，会生成多个样本用于单个输入图像的平均分割图计算。U-Net架构基于以前的工作 [?], 而输入图像生成器是使用残差密集块构建的 [?]. 去噪样本生成器是一个简单的2D卷积层。

### 3.9 Multi-Task Approaches

应用了一系列的扩散模型于多个任务，展示了跨任务的良好泛化能力。下面我们讨论这些贡献。

Song等人 [?] 提出了噪声条件评分网络 (NCSN)，该方法在不同噪声尺度下估计了评分函数。为了进行采样，他们引入了退火版本的 Langevin 动力学，并用它来报告了图像生成和修复方面的结果。NCSN 架构主要基于 [?] 中的工作，只是做了一些小的修改，如将批归一化替换为实例归一化。

Kadkhodaie等人 [?] 训练了一个神经网络，用于恢复受高斯噪声污染的图像，生成的噪声使用受限于特定范围的随机标准差生成。训练后，神经网络的输出与输入的噪声图像之间的差异与噪声数据的对数密度的梯度成正比。这个性质基于 [?] 中的先前工作。对于图像生成，作者使用上述差异作为梯度（评分）估计，并使用类似退火 Langevin 动力学的迭代方法从网络的隐式数据先验中采样 [?]. 但是，这两种采样方法存在一些不同之处，例如在迭代更新中注入的噪声遵循不同的策略。在 [?] 中，注入的噪声根据网络的估计进行调整，而在 [?] 中则是固定的。此外，[?] 中的梯度估计是通过得分匹配进行学习的，而 Kadkhodaie等人 [?] 利用前面提到的性质来计算梯度。Kadkhodaie等人 [?] 的贡献进一步发展的是将该算法适用于线性反问题，如去模糊和超分辨率。

扩散模型在 [?] 中引入的 SDE 公式推广了几种先前的方法 [?], [?], [?]. Song等人 [?] 将正向和反向扩散过程作为 SDE 的解来呈现。这种技术解锁了新的采样方法，例如预测-校正采样器或基于 ODE 的确定性采样器。作者对图像生成、修复和上色进行了实验。

Batzolis等人 [?] 引入了扩散模型中的一种新的正向过程，称为非均匀扩散。这是由每个像素使用不同的 SDE 进行扩散决定的。在这个过程中需要使用多个网络，每个网络对应一个不同的扩散尺度。该论文进一步演示了一种新颖的条件采样器，通过插值两种基于去噪评分的采样方法之间的差异。该模型的架构基于 [?] 和 [?], 在无条件合成、超分辨率、修复和从边缘到图像的转换方面进行了评估。

Esser等人 [?] 提出了 ImageBART，一种生成模型，它学习在紧凑图像表示上恢复多项式扩散过程。Transformer 用于自回归地建模反向步骤，其中编码器的表示是通过前一步的输出得到的。ImageBART 在无条件、类别条件和文本条件下的图像生成以及局部编辑方面进行了评估。

Gao等人 [?] 提出了扩散恢复似然，一种新的能量模型训练过程。他们学习了一系列能量模型，用于扩散过程的边缘分布。因此，他们不是用正态分布来近似反向过程，而是从边缘能量模型中推导出条件分布。作者在图像生成和修复两个任务上进行了实验。Batzolis等人 [1] 分析了先前基于分数的扩散模型在条件图像生成上的应用。此外，他们提出了一种新的条件图像生成方法，称为条件多速扩散估计器 (CMDE)。该方法基于这样的观察：以相同速率扩散目标图像和条件图像可能是次优的。因此，他们建议使用SDE对具有相同漂移但不同扩散速率的两幅图像进行扩散。该方法在修补、超分辨率和边缘到图像合成方面进行了评估。

Liu等人 [2] 引入了一个框架，允许从参考图像中获取文本、内容和样式指导。核心思想是使用最大化图像和文本表示之间相似性的方向。图像和文本嵌入是由CLIP模型[3]生成的。为了解决对噪声图像进行训练CLIP的需求，作者提出了一种不需要文本说明的自监督过程。该过程使用正常图像和噪声图像对的方式，通过最大化正对和最小化负对的相似性来进行对比目标。

Choi等人 [4] 提出了一种新颖的方法，该方法不需要进一步的训练，即可使用无条件扩散模型进行条件图像合成。给定一幅参考图像（即条件），通过消除低频内容并用参考图像的内容替换，将每个样本靠近参考图像。低通滤波器由下采样操作表示，其后是相同因子的上采样滤波器。作者展示了这种方法如何应用于各种图像到图像的转换任务，例如绘画到图像和用涂鸦进行编辑。

Hu等人 [5] 提出了在离散表示上应用扩散模型的方法，该离散表示由离散VAE给出。他们通过在CelebA-HQ和LSUN Church数据集上进行图像生成和修补实验来评估这一想法。

Rombach等人 [6] 引入了潜在扩散模型，其中前向和后向过程发生在由自动编码器学习的潜在空间上。他们还在架构中包含了交叉注意力，从而进一步提高了条件图像合成的性能。该方法在超分辨率、图像生成和修补方面进行了测试。

Preechakul等人 [7] 提出的方法包括一个语义编码器，用于学习描述性潜在空间。该编码器的输出用于对DDIM的实例进行条件约束。所提出的方法使得DDPM在插值或属性操作等任务上表现良好。

Chung等人 [8] 引入了一种用于采样的算法，该算法减少了条件情况下所需的步骤数。与标准情况相比，在标准情况下，反向过程从高斯噪声开始，他们的方法首先执行一个前向步骤以获得一个中间噪声图像，并从此点继续采样。该方法在修补、超分辨率和核磁共振成像 (MRI) 重建方面进行

了测试。

在[9]中，作者对预训练的DDIM进行微调，以根据文本描述生成图像。他们提出了一种局部定向CLIP损失，该损失基本上强制生成图像和原始图像之间的方向尽可能接近参考（原始域）和目标文本（目标域）之间的方向。在评估中考虑的任务是不同领域之间的图像转换和多属性转移。

[1] Batzolis等人，《扩散模型在条件图像生成上的分析》，arXiv, 2021。[2] Liu等人，《基于参考图像的文本-图像生成框架》，arXiv, 2021。[3] Radford,《CLIP模型》，ICML, 2021。[4] Choi等人，《无条件扩散模型的条件图像合成方法》，arXiv, 2021。[5] Hu等人，《离散表示上的扩散模型在图像生成和修补中的应用》，CVPR, 2022。[6] Rombach等人，《基于潜在扩散模型的条件图像合成》，CVPR, 2022。[7] Preechakul等人，《具有描述性潜在空间的DDIM方法》，CVPR, 2022。[8] Chung等人，《减少条件情况下的采样步骤数的算法》，CVPR, 2022。[9] Kim等人，《基于文本描述的预训练DDIM生成图像的方法》，CVPR, 2022。从Meng等人的论文中提出的SDE扩散模型的公式开始，Khrulkov等人研究了潜在空间和由此产生的编码器映射。根据蒙热公式，这些编码器映射被证明是最优传输映射，但这仅针对多元正态分布进行了证明。作者进一步通过数值实验以及Dhariwal等人模型实现的实际实验来支持这一观点。

Shi等人首先观察到无条件的基于评分的扩散模型可以被规约为Schrödinger桥，可以使用修改过的迭代比例拟合法来求解。然后，将先前的方法重构为可接受条件的形式，从而实现有条件合成。为了优化收敛所需的时间，对迭代算法进行了进一步调整。该方法首先在Kovachki等人的合成数据上进行验证，显示出估计真实值的能力有所提高。作者还在超分辨率、修复以及生化需氧量方面进行了实验，后者的任务受到Marzouk等人的启发。

受检索变压器的启发，Blattmann等人提出了一种用于训练扩散模型的新方法。首先，使用最近邻算法从数据库中获取一组相似图像。然后，使用具有固定参数的编码器对图像进行进一步编码，并投影到CLIP特征空间中。最后，扩散模型的逆过程在这个潜在空间上进行条件操作。该方法可以进一步扩展，通过增强潜在空间中的信号编码表示，例如文本，以使用其他条件信号。

Lyu等人提出了一种减少扩散模型采样步骤数量的新技术，同时提高性能。其想法是在较早的阶段停止扩散过程。由于采样不能从随机高斯噪声开始，所以使用GAN或VAE模型将最后扩散的图像编码成高斯潜在空间。然后将结果解码为可以扩散到反向过程的起始点的图像。

Graikos等人的目标是将扩散模型分割为两个独立的部分，即先验部分和约束部分，从而使模型能够在不进行进一步训练的情况下应用于各种任务。通过改变DDPMs的方程式，从而独立训练模型并在条件设置中使用它，前提是约束可微分。作者在条件图像合成和图像分割方面进行了实验。

### 3.10 Medical Image Generation and Translation

Wolleb等人[?]在脑肿瘤分割的背景下，引入了一种基于扩散模型的图像分割方法。训练过程包括扩散分割图，然后通过去噪得到原始图像。在反向过程中，将脑MR图像连接到中间去噪步骤中，以便通过U-Net模型传递并在其上条件化去噪过程。此外，对于每个输入，作者提出生成多个样本，由于随机性而不同。因此，集成可以生成平均分割图及其方差（与图的不确定性相关）。

Song等人[?]介绍了一种用于基于分数的模型的方法，能够解决医学图像中的逆问题，即从测量中重建图像。首先，训练一个无条件分数模型。然后，得到一个衡量过程的随机过程，可以通过近端优化步骤将条件信息融入模型中。最后，将信号映射到测量的矩阵进行分解，以允许封闭形式的采样。作者进行了多个实验，涵盖了不同的医学图像类型，包括计算机断层扫描(CT)、低剂量CT和MRI。

在医学影像领域，重构加速MRI扫描图像的过程中，Chung等人[?]提出使用基于分数的扩散模型来解决逆问题。一个分数模型在无条件环境下仅在幅值图像上进行预训练。然后，在采样过程中使用方差爆炸SDE求解器[?]。通过采用预测校正算法[?]与数据一致性映射相结合，将分割图像（实部和虚部）传入，使模型在测量上条件化。此外，作者还提出了一种扩展方法，可以在多个线圈变化的测量上进行条件化。

Ozbey等人[?]提出了一种具有对抗式推断的扩散模型。为了提高每个扩散步骤的效果，并减少步骤次数，受到[?]的启发，作者在逆过程中采用GAN模型来估计每个步骤的去噪图像。使用类似于[?]的方法，他们引入了一个循环一致结构，以允许在未配对数据集上进行训练。

Hu等人[?]旨在去除光学相干断层扫描(OCT)B扫描中的散斑噪声。第一阶段通过一种称为自我融合的方法来表示，如[?]所述，该方法选择了接近给定输入OCT体积的2D切片的附加B扫描。第二阶段包括一个扩散模型，其起始点是原始B扫描和其邻居的加权平均值。因此，通过对干净扫描进行采样可以去除噪声。

### 3.11 Anomaly Detection in Medical Images

自编码器广泛用于异常检测[?]。由于扩散模型可以被视为一种特殊类型的变分自编码器(VAE)，因此在与VAE相同的任务中使用扩散模型似乎是自然而然的选择。迄今为止，扩散模型在医学图像异常检测方面已显示出有希望的结果，如下所讨论。

Wyatt等人[?]在健康医学图像上训练了一个DDPM。在推理时，通过将生成的图像与原始图像相减来检测异常。该工作还证明，对于这种类型的任务，使用简单噪声而不是高斯噪声能够获得更好的结果。

Wolleb等人[?]提出了一种基于扩散模型的半监督异常检测方法，用于医学图像。给定两个无配对的图像，一个是健

康图像，一个是带有病变的图像，扩散模型对前者进行扩散。然后，通过二元分类器的梯度来引导去噪过程，以生成健康图像。最后，将采样的健康图像和包含病变的图像相减，得到异常图。

Pinaya等人 [?] 提出了一种基于扩散的方法来检测脑部扫描中的异常，并对这些区域进行分割。图像由VQ-VAE [?] 编码，并从码本中获取量化的潜变量表示。扩散模型在这个潜空间中运行。通过对反向过程的中间样本进行中位数步骤的平均，并应用预计算的阈值图，创建一个暗示异常位置的二值化掩码。从中间开始反向过程，使用二值化掩码对异常区域进行去噪，同时保持其余部分。最后，解码最终步骤的样本，得到一张健康图像。通过减去输入图像和合成图像，得到异常的分割图。

Sanchez等人 [?] 在医学图像中遵循相同的原则，用于检测和分割异常：扩散模型生成健康样本，然后从原始图像中减去。输入图像使用模型进行扩散，反转去噪方程并使条件失效，然后应用反向条件过程。利用无分类器的模型，通过集成在U-Net中的注意机制实现指导。训练中使用了健康和不健康的示例。

### 3.12 Video Generation

最近对扩散模型效率的改进使得这种模型能够应用于视频领域。接下来我们将介绍应用扩散模型进行视频生成的相关研究。

Ho等人 [?] 将扩散模型引入到视频生成任务中。与2D情况相比，改动只应用于架构上。作者采用了 [?] 中的3D U-Net，并展示了无条件和文本条件下的视频生成结果。较长的视频以自回归方式生成，后面的视频块依赖于前面的视频块。

Yang等人 [?] 使用扩散模型逐帧生成视频。反向过程完全依赖于由卷积递归神经网络提供的上下文向量。作者进行了一项消融研究，以确定预测下一帧的残差是否比预测实际帧更好。研究得出的结论是前一种选项效果更好。

Höppe等人 [?] 提出了随机掩码视频扩散 (RaMViD) 方法，该方法可用于视频生成和填充。他们的工作的主要贡献是一种新的训练策略，将帧随机分为掩码帧和非掩码帧。非掩码帧用于条件扩散，而掩码帧则通过正向过程进行扩散。

Harvey等人 [?] 介绍了灵活的扩散模型，这是一种适用于长视频生成的扩散模型类型，可使用多种采样方案。与 [?] 类似，作者通过随机选择用于扩散和条件过程的帧来训练扩散模型。在训练完模型之后，他们研究了多种采样方案的有效性，并得出结论：采样选择取决于数据集。

### 3.13 Other Tasks

有一些开创性的工作将扩散模型应用于新任务，这些任务目前很少通过扩散建模进行探索。我们在下面收集并讨论这些贡献。

Luo等人 [?] 将扩散模型应用于3D点云生成、自编码和无监督表示学习。他们从点云条件形状潜变量的似然下界变分推导出目标函数。实验使用PointNet [?] 作为底层架构。

Zhou等人 [?] 引入了基于点体素扩散 (PVD) 的新型形状生成方法，该方法在点体素表示上应用了扩散模型。该方法解决了ShapeNet和PartNet数据集上的形状生成和完成任务。

Zimmermann等人 [?] 展示了一种用于分类的基于分数的模型策略。他们将图像标签作为条件变量添加到分数函数中，并且由于ODE的公式，可以在推理时计算条件似然。因此，预测结果是具有最大似然的标签。此外，他们还研究了这种类型分类器在考虑常见图像破坏和对抗扰动的离分布场景中的影响。

Kim等人 [?] 提出使用扩散模型解决图像配准任务。这通过两个网络实现，一个是扩散网络，参考文献 [?]，另一个是基于U-Net的变形网络，如文献 [?] 所述。给定两个图像（一个静态，一个移动），前一个网络的作用是评估两个图像之间的变形，并将结果馈送到后一个网络，后者预测变形场，从而实现样本生成。该方法还具有通过整个过渡阶段合成变形的能力。作者对不同任务进行了实验，一个是2D面部表情，一个是3D脑图像。结果证实该模型能够生成定性和准确的配准场。

Jeanneret等人 [?] 将扩散模型应用于因果解释。该方法从一个噪声查询图像开始，使用无条件的DDPM生成样本。使用生成的样本计算所需的梯度来进行引导。然后，应用一个反向引导过程的步骤。输出进一步用于下一步的反向过程。

Sanchez等人 [?] 针对反事实图像生成改进了Dhariwal等人 [?] 的工作。与 [?] 一样，去噪过程由分类器梯度引导以生成所需的反事实类别样本。其关键贡献是用于检索原始图像的潜在表示的算法，该算法反转了 [?] 的确定性采样过程，并将每个原始图像映射到唯一的潜在表示。

Nie等人 [?] 展示了如何利用扩散模型作为对抗攻击的防御机制。给定一个对抗图像，它会扩散到最佳计算时间步长。然后模型对结果进行反转，产生一个纯净的样本。为了优化求解逆向时间SDE的计算，使用了Li等人 [?] 的伴随灵敏度方法来进行梯度分数计算。在Few-shot learning的背景下，Giannone等人提出了基于扩散模型的图像生成器 [?]。给定一小组作为合成条件的图像，视觉转换器对其进行编码，得到的上下文表示通过两种不同的技术集成到在去噪过程中使用的U-Net模型中。

Wang等人提出了一个基于扩散模型的语义图像合成框架 [?]。利用扩散模型的U-Net架构，输入的噪声经过编码器处理，而语义标签图通过多层空间自适应归一化操作符传递给解码器 [?]。为了进一步提高采样质量和对语义标签图的条件，还向采样方法提供了一个空白图以产生无条件噪声。最后，最终的噪声使用了这两个估计值。

关于通过各种天气条件（如雪、雨）对图像进行修复的

任务, Özdenizci等人展示了如何利用扩散模型 [?]. 他们通过按通道连接退化图像和去噪样本来将去噪过程与退化图像关联起来, 每个时间步骤都进行这样的操作。为了处理不同的图像尺寸, 在每个步骤中, 样本被划分为重叠的块, 通过模型并行传递, 并通过对重叠像素求平均值进行合并。所使用的扩散模型基于U-Net架构, 如 [?], [?]中所述, 但经过修改以接受两个连接的图像作为输入。

将图像修复任务表示为线性逆问题, Kawar等人提出了扩散模型的使用 [?]. 受到Kawar等人 [?]的启发, 线性退化矩阵通过奇异值分解进行分解, 使得输入和输出都可以映射到扩散过程所进行的矩阵的频谱空间中。利用 [?]和 [?]中预训练的扩散模型, 对超分辨率、去模糊、上色和修补等各种任务进行了评估。

### 3.14 Theoretical Contributions

Huang等人(2021年)展示了宋等人(2021年)提出的方法与最大化逆随机微分方程(reverse SDE)边际似然的下界之间的联系。此外, 他们通过对CIFAR-10和MNIST数据集进行图像生成实验来验证他们的理论贡献。

## 4 CLOSING REMARKS AND FUTURE DIRECTIONS

在本文中, 我们回顾了研究界在开发和应用扩散模型到各种计算机视觉任务方面所取得的进展。我们确定了基于DDPMs、NCSNs和SDEs的三种主要扩散建模方法。每种方法在图像生成方面都取得了显著的结果, 超越了生成对抗网络(GANs), 同时增加了生成样本的多样性。尽管研究仍处于早期阶段, 扩散模型的优异结果已经得到了实现。虽然我们观察到主要关注点是有条件和无条件的图像生成, 但仍有许多任务有待探索和进一步改进。

**限制。** 扩散模型最显著的缺点仍然是在推理时需要执行多个步骤才能生成一个样本。尽管已经在这个方向上进行了大量的研究, 生成对抗网络(GANs)在生成图像方面仍然更快。其他扩散模型的问题可能与常用的使用CLIP嵌入进行文本到图像生成的策略有关。例如, Ramesh等人[?]指出, 他们的模型在图像中生成可读文本方面存在困难, 并通过指出CLIP嵌入不包含拼写信息来解释这种行为。因此, 当使用这些嵌入来对去噪过程进行条件约束时, 模型可能会继承这种问题。

**未来方向。** 为了降低不确定性水平, 扩散模型通常避免在采样过程中采用大步长。确实, 采取小步长确保了学习到的高斯分布解释了每个步骤生成的数据样本。应用梯度下降来优化神经网络时也观察到类似的行为。实际上, 在负梯度的负方向采取大步长, 即使用非常大的学习率, 可能会导致更新模型到具有高不确定性的区域, 无法对损失值进行控制。在未来的工作中, 将从高效优化器借鉴的更新规则转移到扩散模型中, 可能会导致更高效的采样(生成)过程。

除了当前更加高效的扩散模型研究趋势外, 未来的工作可以研究扩散模型在其他计算机视觉任务中的应用, 例如图像去雾、视频异常检测或视觉问答。尽管我们找到了一些研究医学图像异常检测的工作[?], [?], [?], 但此任务也可以在其他领域中进行探索, 比如视频监控或工业检测。

一个有趣的研究方向是评估扩散模型学到的表示空间在判别任务中的质量和效用。这可以通过至少两种不同的方式进行。直接方式是在去噪模型提供的潜在表示上学习一些判别模型, 来解决某个分类或回归任务。间接方式是通过扩散模型生成的逼真样本来扩充训练集。后一种方向可能更适用于目标检测等任务, 在图像中修复扩散模型可以很好地融合新对象。

另一个未来的方向是利用有条件的扩散模型模拟视频中的可能未来。生成的视频可以进一步作为输入传递给强化学习模型。

最近的扩散模型[?]在文本到视频合成方面显示出令人印象深刻的潜力, 与先前的技术相比, 显著减少了伪影, 达到了前所未有的生成性能。然而, 我们认为这个方向在未来的工作中需要更多关注, 因为生成的视频时间较短。因此, 对于建模长期时序关系和物体之间的交互仍然是一个未解决的挑战。

未来, 关于扩散模型的研究还可以扩大到学习多用途模型, 即解决多个任务。创建一个扩散模型来生成多种类型的输出, 同时以各种类型的数据进行条件限制, 例如文本、类标签或图像, 可能会让我们更接近理解发展人工通用智能(AGI)所需的必要步骤。

## ACKNOWLEDGMENTS

### ACKNOWLEDGMENT

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project no. PN-III-P2-2.1-PED-2021-0195, contract no. 690/2022, within PNCDI III.

## 参考文献



**Florinel-Alin Croitoru** 是罗马尼亚布加勒斯特大学的博士研究生。他于2019年在布加勒斯特大学数学与计算机科学学院获得学士学位。2021年, 他以一篇关于足球视频中动作监控的论文获得人工智能硕士学位。他的研究领域包括机器学习, 计算机视觉和深度学习。



**Vlad Hondu** 是罗马尼亚布加勒斯特大学的博士生。他在曼彻斯特大学获得了机械电子工程学士学位，然后在伦敦帝国理工学院获得了计算机科学硕士学位，专攻可视化计算和机器人技术，并以人工智能为重点。他曾在劳斯莱斯公司担任软件工程师一年，并在曼彻斯特大学的机器人小组进行了一个暑期实习。他目前在一家机器学习工程师，开发自然语言处理产品。

**Radu Ionescu** 是罗马尼亚布加勒斯特大学的教授。他于2013年在罗马尼亚布加勒斯特大学完成了博士学位，获得了罗马尼亚Ad Astra协会颁发的2014年度优秀博士研究奖。他的研究兴趣包括机器学习、计算机视觉、图像处理、计算语言学和医学成像。他在国际会议（包括CVPR、NeurIPS、ICCV、ACL、EMNLP、NAACL、TPAMI、IJCV、CVIU）上发表了100多篇文章，并出版了一本由Springer出版社的研究专著。Radu还在2013年的ICIAP会议上获得了“Caianiello最佳青年论文奖”。Radu还获得了2017年的“青年科学与工程研究员奖”，以及2018年罗马尼亚的“Danubius青年科学家奖”。

—  
—  
— —

Mubarak Shah 佛罗里达中央大学（UCF）计算机视觉研究中心创始主任和UCF受托主席教授。他是现代技术与发明院、IEEE、AAAS、IAPR和SPIE的会士。他是国际视频计算图书系列的编辑，曾任《机器视觉和应用》的主编，还是ACM Computing Surveys和IEEE TPAMI的副编辑。

他的研究兴趣包括视频监控、视觉跟踪、人类活动识别、拥挤场景的视觉分析、视频注册、无人机视频分析等。他曾担任ACM杰出演讲者和IEEE杰出访问学者。他获得了ACM SIGMM技术成就奖、IEEE杰出工程教育家奖、Harris Corporation工程成就奖、ICCV 2005年“我在哪里？”挑战问题荣誉奖、2013年NGA最佳研究海报展示奖、ACM Multimedia 2013年大赛亚军，并且在2005年和2010年的ACM Multimedia Conference中获得了最佳论文奖的亚军。在UCF，他获得了Pegasus教授奖、杰出研究奖、博士生导师卓越奖、教学与学习奖、教学激励计划

奖和研究激励奖。

## 附录 A

### VARIATIONAL BOUND.

我们强调，下面所展示的推导也在 [?], [?] 中有所呈现。数据的对数密度的变分下界可以按照VAE的情况推导 [?], 其中潜在变量是噪声图像  $x_{1:T}$ ，观察变量是原始图像  $x_0$ 。我们从将数据的对数似然  $\log p_\theta(x_0)$  写成联合概率  $p_\theta(x_{0:T})$  的边缘对数形式开始推导：

$$\begin{aligned} \log p_\theta(x_0) &= \log \int p_\theta(x_{0:T}) dx_{1:T} \\ &= \log \int p_\theta(x_{0:T}) \cdot \frac{p(x_{1:T}|x_0)}{p(x_{1:T}|x_0)} dx_{1:T} \\ &= \log \int p(x_{1:T}) \cdot \frac{p_\theta(x_{0:T})}{p(x_{1:T}|x_0)} dx_{1:T} \\ &= \log \mathbb{E}_{x_{1:T} \sim p(x_{1:T}|x_0)} \left[ \frac{p_\theta(x_{0:T})}{p(x_{1:T}|x_0)} \right]. \end{aligned} \quad (15)$$

Jensen不等式指出，对于一个随机变量  $Y$  和一个凸函数  $f$ ，以下结论成立：

$$f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)]. \quad (16)$$

如果我们将公式(16) 应用于公式(15)，并且由于  $\log$  函数是凹函数而改变不等号的方向，那么我们得到如下结果：

$$\begin{aligned} \log p_\theta(x_0) &\geq \mathbb{E}_{x_{1:T} \sim p(x_{1:T}|x_0)} \left[ \log \frac{p_\theta(x_{0:T})}{p(x_{1:T}|x_0)} \right] \cdot (-1) \\ -\log p_\theta(x_0) &\leq \mathbb{E}_{x_{1:T} \sim p(x_{1:T}|x_0)} \left[ \log \frac{p(x_{1:T}|x_0)}{p_\theta(x_{0:T})} \right]. \end{aligned} \quad (17)$$

方程式 (17) 表明，我们可以最小化不等式的右侧，而不是最小化生成模型下数据的期望负对数似然。我们进一步关注这一项，最终我们将从方程 (4) 推导出目标函数。

根据定义，正向和逆向过程是马尔可夫过程。基于这一点，我们可以将方程 (17) 中的概率重写如下：

$$\begin{aligned} p(x_{1:T}|x_0) &= p(x_T|x_{1:T-1}, x_0) \cdot p(x_{1:T-1}|x_0) \\ &= p(x_T|x_{T-1}) \cdot p(x_{T-1}|x_{1:T-2}, x_0) \cdot p(x_{1:T-2}|x_0) \\ &= \dots = \prod_{t=1}^T p(x_t|x_{t-1}), \\ p_\theta(x_{0:T}) &= p_\theta(x_0|x_{1:T}) \cdot p_\theta(x_{1:T}) \\ &= p_\theta(x_0|x_1) \cdot p_\theta(x_1|x_{2:T}) \cdot p_\theta(x_{2:T}) \\ &= \dots = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t). \end{aligned} \quad (18)$$

我们将公式 (17) 右侧的概率替换为公式 (18) 中的乘积，并应用对数的性质，将乘积转换为求和：

$$\begin{aligned} \mathbb{E}_p \left[ \log \frac{p(x_{1:T}|x_0)}{p_\theta(x_{0:T})} \right] &= \\ &= \mathbb{E}_{x_{1:T} \sim p(x_{1:T}|x_0)} \left[ -\log p_\theta(x_T) + \sum_{t=1}^T \log \frac{p(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)} \right]. \end{aligned} \quad (19)$$

术语  $p(x_t|x_{t-1})$  可以通过贝叶斯规则转化为  $\frac{p(x_{t-1}|x_t) \cdot p(x_t)}{p(x_{t-1})}$ ，但正向过程的真实后验  $p(x_{t-1}|x_t)$  是难以计算的。然而，如果我们额外给出初始图像  $x_0$  的条件，后验将变得可计算。此外，由于正向过程是马尔科夫过程，我们知道  $p(x_t|x_{t-1}, x_0) = p(x_t|x_{t-1})$  是成立的。因此，如果我们对  $p(x_t|x_{t-1}, x_0)$  应用贝叶斯规则，我们将获得对真实后验的附加条件：

$$p(x_t|x_{t-1}, x_0) = \frac{p(x_{t-1}|x_t, x_0) \cdot p(x_t|x_0)}{p(x_{t-1}|x_0)}. \quad (20)$$

如果我们将公式 (20) 的推导应用到公式 (19) 中的所有  $t \geq 2$ ，结果如下

所示:

$$\begin{aligned}\mathcal{L}_{\text{Vlb}} &= \mathbb{E}_p \left[ -\log p_\theta(x_T) + \sum_{t=1}^T \log \frac{p(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)} \right] \\ &= \mathbb{E}_p[-\log p_\theta(x_T)] + \mathbb{E}_p \left[ \sum_{t=2}^T \log \frac{p(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} \right] \quad (21) \\ &\quad + \mathbb{E}_p \left[ \sum_{t=2}^T \log \frac{p(x_t|x_0)}{p(x_{t-1}|x_0)} + \log \frac{p(x_1|x_0)}{p_\theta(x_0|x_1)} \right].\end{aligned}$$

观察到第二个求和的项相互抵消,  $\mathcal{L}_{\text{Vlb}}$  变成:

$$\begin{aligned}\mathcal{L}_{\text{Vlb}} &= \mathbb{E}_p[-\log p_\theta(x_T)] + \mathbb{E}_p \left[ \sum_{t=2}^T \log \frac{p(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} \right] \quad (22) \\ &\quad + \mathbb{E}_p \left[ \log \frac{p(x_T|x_0)}{p(x_1|x_0)} + \log \frac{p(x_1|x_0)}{p_\theta(x_0|x_1)} \right].\end{aligned}$$

最后, 如果我们重新安排这些术语并将对数率转化为Kullback-Leibler分歧, 那么结果就是方程(4)中给出的表述:

$$\begin{aligned}\mathcal{L}_{\text{Vlb}} &= \mathbb{E}_p[-\log p_\theta(x_0|x_1)] + \mathbb{E}_p \left[ \sum_{t=2}^T \log \frac{p(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} \right] \\ &\quad + \mathbb{E}_p \left[ \log \frac{p(x_T|x_0)}{p_\theta(x_T)} \right] \quad (23) \\ &= -\log p_\theta(x_0|x_1) + KL(p(x_T|x_0)\|p_\theta(x_T)) \\ &\quad + \sum_{t=2}^T KL(p(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)).\end{aligned}$$

## 附录 B

### NOISE ESTIMATION.

在本节中, 我们重点讨论Ho等人提出的简化方法和达到Eq. (6)简单目标所需的调整。

第一个简化方法是避免训练 $p_\theta(x_{t-1}|x_t)$ 的协方差, 而是事先将其固定为 $\sigma_t^2 \cdot \mathbf{I}$ 。在实践中, Ho等人建议使用 $\sigma_t^2 = \beta_t$ 。这个改变影响到 $\mathcal{L}_{\text{Vlb}}$ 的Kullback-Leibler项, 因为如果协方差不可训练, 那么散度可以重写为两个分布均值之间的距离加上一个与 $\theta$ 无关的常数:

$$\begin{aligned}\mathcal{L}_{KL} &= KL(p(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)) \\ &= \frac{1}{2 \cdot \sigma_t^2} \cdot \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 + C, \quad (24)\end{aligned}$$

其中 $\tilde{\mu}(x_t, x_0)$ 是 $p(x_{t-1}|x_t, x_0)$ 的均值,  $\mu_\theta(x_t, t)$ 是 $p_\theta(x_{t-1}|x_t)$ 的均值,  $C$ 是一个常数。我们强调一下, 此时神经网络的输出是 $\mu_\theta(x_t, t)$ 。

下一步的改变基于以下观察: 均值 $\tilde{\mu}(x_t, x_0)$ 可以表示为 $x_t$ 和 $z_t$ 的函数, 如下所示:

$$\tilde{\mu}(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \hat{\beta}_t}} \cdot z_t \right). \quad (25)$$

这意味着根据Eq. (24),  $\mu_\theta(x_t, t)$ 必须近似于这个表达式。然而,  $x_t$ 是模型的输入。因此, Ho等人[?]建议以相同的方式重新参数化 $\mu_\theta(x_t, t)$ :

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \hat{\beta}_t}} \cdot z_\theta(x_t, t) \right), \quad (26)$$

其中 $z_\theta(x_t, t)$ 现在是神经网络的输出, 即给定有噪声的图像 $x_t$ 的噪声 $z_t$ 的估计。

如果我们将 $\mathcal{L}_{KL}$ 中的均值替换为方程(25)和方程(26)中的参数化形式, 则结果如下:

$$\mathcal{L}_{KL} = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1 - \hat{\beta}_t)} \|z_t - z_\theta(x_t, t)\|^2. \quad (27)$$

这个术语实质上是图像 $x_t$ 的真实噪声与网络估计之间的时间加权距离。Ho等人[?]进一步简化了这个术语, 丢弃了权重 $\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1 - \hat{\beta}_t)}$ , 得到了一

个形式, 也涵盖了 $\mathcal{L}_{\text{Vlb}}$ 的第一项。因此, 通过进行这些最后的更改, 最终的目标变为方程(6)的简化版本:

$$\mathcal{L}_{\text{Simple}} = \mathbb{E}_{t \sim [1, T]} \mathbb{E}_{x_0 \sim p(x_0)} \mathbb{E}_{z_t \sim \mathcal{N}(0, \mathbf{I})} \|z_t - z_\theta(x_t, t)\|^2. \quad (28)$$

表 1

我们对计算机视觉中应用的扩散模型进行了多视角分类。为了对现有的模型进行分类，我们考虑了三个标准：任务、去噪条件和底层方法（架构）。此外，我们列出了调查模型所应用的数据集。我们在架构一列中使用了以下缩写：D3PM（离散去噪扩散概率模型）、DSB（扩散Schrödinger桥）、BDDM（双边去噪扩散模型）、PNDM（扩散模型的伪数值方法）、ADM（消融扩散模型）、D2C（具有对比表示的扩散解码模型）、CCDF（趋近-扩散-加速）、VQ-DDM（向量量化离散扩散模型）、BF-CNN（无偏CNN）、FDM（灵活扩散模型）、RVD（残差视频扩散）、RaMViD（随机掩码视频扩散）。

Paper	Task	Denoising Condition	Architecture	Data Sets
Austin <i>et al.</i> [?]	image generation	unconditional	D3PM	CIFAR-10
Bao <i>et al.</i> [?]	image generation	unconditional	DDIM, Improved DDPM	CelebA, ImageNet, LSUN Bedroom, CIFAR-10
Benny <i>et al.</i> [?]	image generation	unconditional	DDPM, DDIM	CIFAR-10, ImageNet, CelebA
Bond-Taylor <i>et al.</i> [?]	image generation	unconditional	DDPM	LSUN Bedroom, LSUN Church, FFHQ
Choi <i>et al.</i> [?]	image generation	unconditional	DDPM	FFHQ, AFHQ-Dog, CUB, Met-Faces
De <i>et al.</i> [?]	image generation	unconditional	DSB	MNIST, CelebA
Deasy <i>et al.</i> [?]	image generation	unconditional	NCSN	MNIST, Fashion-MNIST, CIFAR-10, CelebA
Deja <i>et al.</i> [?]	image generation	unconditional	Improved DDPM	Fashion-MNIST, CIFAR-10, CelebA
Dockhorn <i>et al.</i> [?]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10
Ho <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, CelebA-HQ, LSUN
Huang <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, MNIST
Jing <i>et al.</i> [?]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10, CelebA-256-HQ, LSUN Church
Jolicoeur <i>et al.</i> [?]	image generation	unconditional	NCSN	CIFAR-10, LSUN Church, Stacked-MNIST
Jolicoeur <i>et al.</i> [?]	image generation	unconditional	DDPM++, NCSN++	CIFAR-10, LSUN Church, FFHQ
Kim <i>et al.</i> [?]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10, CelebA, MNIST
Kingma <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Kong <i>et al.</i> [?]	image generation	unconditional	DDIM, DDPM	LSUN Bedroom, CelebA, CIFAR-10
Lam <i>et al.</i> [?]	image generation	unconditional	BDDM	CIFAR-10, CelebA
Liu <i>et al.</i> [?]	image generation	unconditional	PNDM	CIFAR-10, CelebA
Ma <i>et al.</i> [?]	image generation	unconditional	NCSN, NCSN++	CIFAR-10, CelebA, LSUN Bedroom, LSUN Church, FFHQ
Nachmani <i>et al.</i> [?]	image generation	unconditional	DDIM, DDPM	CelebA, LSUN Church
Nichol <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Pandey <i>et al.</i> [?]	image generation	unconditional	DDPM	CelebA-HQ, CIFAR-10
San <i>et al.</i> [?]	image generation	unconditional	DDPM	CelebA, LSUN Bedroom, LSUN Church
Sehwag <i>et al.</i> [?]	image generation	unconditional	ADM	CIFAR-10, ImageNet
Sohl-Dickstein <i>et al.</i> [?]	image generation	unconditional	DDPM	MNIST, CIFAR-10, Dead Leaf Images
Song <i>et al.</i> [?]	image generation	unconditional	NCSN	FFHQ, CelebA, LSUN Bedroom, LSUN Tower, LSUN Church Outdoor
Song <i>et al.</i> [?]	image generation	unconditional	DDPM++	CIFAR-10, ImageNet 32×32
Song <i>et al.</i> [?]	image generation	unconditional	DDIM	CIFAR-10, CelebA, LSUN
Vahdat <i>et al.</i> [?]	image generation	unconditional	NCSN++	CIFAR-10, CelebA-HQ, MNIST
Wang <i>et al.</i> [?]	image generation	unconditional	DDIM	CIFAR-10, CelebA
Wang <i>et al.</i> [?]	image generation	unconditional	StyleGAN2, ProjectedGAN	CIFAR-10, STL-10, LSUN Bedroom, LSUN Church, AFHQ, FFHQ
Watson <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Watson <i>et al.</i> [?]	image generation	unconditional	Improved DDPM	CIFAR-10, ImageNet 64×64
Xiao <i>et al.</i> [?]	image generation	unconditional	NCSN++	CIFAR-10
Zhang <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, MNIST
Zheng <i>et al.</i> [?]	image generation	unconditional	DDPM	CIFAR-10, CelebA, CelebA-HQ

Ho <i>et al.</i> [?]	conditional image generation	conditioned on label	DDPM	LSUN, ImageNet
Ho <i>et al.</i> [?]	conditional image generation	unconditional, classifier-free guidance	ADM	ImageNet 64×64, ImageNet 128×128
Karras <i>et al.</i> [?]	conditional image generation	unconditional, conditioned on class	DDPM++, NCSN++, DDPM, DDIM	CIFAR-10, ImageNet 64×64
Liu <i>et al.</i> [?]	conditional image generation	conditioned on text, image, style guidance	DDPM	FFHQ, LSUN Cat, LSUN Horse, LSUN Bedroom
Liu <i>et al.</i> [?]	conditional image generation	conditioned on text, 2D positions, relational descriptions between items, human facial attributes	Improved DDPM	CLEVR, Relational CLEVR, FFHQ
Lu <i>et al.</i> [?]	conditional image generation	unconditional, conditioned on class	DDIM	CIFAR-10, CelebA, ImageNet, LSUN Bedroom
Salimans <i>et al.</i> [?]	conditional image generation	unconditional, conditioned on class	DDIM	CIFAR-10, ImageNet, LSUN
Singh <i>et al.</i> [?]	conditional image generation	conditioned on noise	DDIM	ImageNet
Sinha <i>et al.</i> [?]	conditional image generation	unconditional, conditioned on label	D2C	CIFAR-10, CIFAR-100, fMoW, CelebA-64, CelebA-HQ-256, FFHQ-256
Ho <i>et al.</i> [?]	image-to-image translation	conditioned on image	Improved DDPM	ctest10k, places10k
Li <i>et al.</i> [?]	image-to-image translation	conditioned on image	DDPM	Face2Comic, Edges2Shoes, Edges2Handbags
Sasaki <i>et al.</i> [?]	image-to-image translation	conditioned on image	DDPM	CMP Facades, KAIST Multi-spectral Pedestrian
Wang <i>et al.</i> [?]	image-to-image translation	conditioned on image	DDIM	ADE20K, COCO-Stuff, DIODE
Wolleb <i>et al.</i> [?]	image-to-image translation	conditioned on image	Improved DDPM	BRATS
Zhao <i>et al.</i> [?]	image-to-image translation	conditioned on image	DDPM	CelebaA-HQ, AFHQ
Gu <i>et al.</i> [?]	text-to-image generation	conditioned on text	VQ-Diffusion	CUB-200, Oxford 102 Flowers, MS-COCO
Jiang <i>et al.</i> [?]	text-to-image generation	conditioned on text	Transformer-based encoder-decoder	DeepFashion-MultiModal
Ramesh <i>et al.</i> [?]	text-to-image generation	conditioned on text	ADM	MS-COCO, AVA
Rombach <i>et al.</i> [?]	text-to-image generation	conditioned on text	LDM	OpenImages, WikiArt, LAION-2B-en, ArtBench
Saharia <i>et al.</i> [?]	text-to-image generation	conditioned on text	Imagen	MS-COCO, DrawBench
Shi <i>et al.</i> [?]	text-to-image generation	unconditional, conditioned on text	Improved DDPM	Conceptual Captions, MS-COCO
Zhang <i>et al.</i> [?]	text-to-image generation	unconditional, conditioned on text	DDIM	CIFAR-10, CelebA, ImageNet
Daniels <i>et al.</i> [?]	super-resolution	conditioned on image	NCSN	CIFAR-10, CelebA
Saharia <i>et al.</i> [?]	super-resolution	conditioned on image	DDPM++	FFHQ, CelebA-HQ, ImageNet-1K
Avrahami <i>et al.</i> [?]	image editing	conditioned on image and mask	DDPM, ADM	ImageNet, CUB, LSUN Bedroom, MS-COCO
Avrahami <i>et al.</i> [?]	region image editing	text guidance	DDPM	PaintByWord
Meng <i>et al.</i> [?]	image editing	conditioned on image	Score SDE, DDPM, Improved DDPM	LSUN, CelebA-HQ
Lugmayr <i>et al.</i> [?]	inpainting	unconditional	DDPM	CelebA-HQ, ImageNet
Nichol <i>et al.</i> [?]	inpainting	conditioned on image, text guidance	ADM	MS-COCO
Amit <i>et al.</i> [?]	image segmentation	conditioned on image	Improved DDPM	Cityscapes, Vaihingen, MoNuSeg
Baranchuk <i>et al.</i> [?]	image segmentation	conditioned on image	Improved DDPM	LSUN, FFHQ-256, ADE-Bedroom-30, CelebA-19
Batzolis <i>et al.</i> [?]	multi-task (inpainting, super-resolution, edge-to-image)	conditioned on image	DDPM	CelebA, Edges2Shoes
Batzolis <i>et al.</i> [?]	multi-task (image generation, super-resolution, inpainting, image-to-image translation)	unconditional	DDIM	ImageNet, CelebA-HQ, CelebA, Edges2Shoes
Blattmann <i>et al.</i> [?]	multi-task (image generation)	unconditional, conditioned	LDM	ImageNet

Hu <i>et al.</i> [?]	multi-task (image generation, in-painting)	unconditional, conditioned on image	VQ-DDM	CelebA-HQ, LSUN Church
Khrulkov <i>et al.</i> [?]	multi-task (image generation, image-to-image translation)	conditioned on class	Improved DDPM	AFHQ, FFHQ, MetFaces, ImageNet
Kim <i>et al.</i> [?]	multi-task (image translation, multi-attribute transfer)	conditioned on image, portrait, stroke	DDIM	ImageNet, CelebA-HQ, AFHQ-Dog, LSUN Bedroom, Church
Luo <i>et al.</i> [?]	multi-task (point cloud generation, auto-encoding, unsupervised representation learning)	conditioned on shape latent	DDPM	ShapeNet
Lyu <i>et al.</i> [?]	multi-task (image generation, image editing)	unconditional, conditioned on class	DDPM	CIFAR-10, CelebA, ImageNet, LSUN Bedroom, LSUN Cat
Preechakul <i>et al.</i> [?]	multi-task (latent interpolation, attribute manipulation)	conditioned on latent representation	DDIM	CelebA-HQ
Rombach <i>et al.</i> [?]	multi-task (super-resolution, image generation, inpainting)	unconditional, conditioned on image	VQ-DDM	ImageNet, CelebA-HQ, FFHQ, LSUN
Shi <i>et al.</i> [?]	multi-task (super-resolution, in-painting)	conditioned on image	Improved DDPM	MNIST, CelebA
Song <i>et al.</i> [?]	multi-task (image generation, in-painting)	unconditional, conditioned on image	NCSN	MNIST, CIFAR-10, CelebA
Kadkhodaie <i>et al.</i> [?]	multi-task (Spatial super-resolution, Deblurring, Compressive sensing, Inpainting, Random missing pixels)	conditioned on linear measurements	BF-CNN	MNIST, Set5, Set68, Set14
Song <i>et al.</i> [?]	multi-task (image generation, in-painting, colorization)	unconditional, conditioned on image, class	NCSN++, DDPM++	CelebA-HQ, CIFAR-10, LSUN
Hu <i>et al.</i> [?]	medical image-to-image translation	conditioned on image	DDPM	ONH
Chung <i>et al.</i> [?]	medical image generation	conditioned on measurements	NCSN++	fastMRI knee
Özbey <i>et al.</i> [?]	medical image generation	conditioned on image	Improved DDPM	IXI, Gold Atlas - Male Pelvis
Song <i>et al.</i> [?]	medical image generation	conditioned on measurements	NCSN++	LIDC, LDCT Image and Projection, BRATS
Wolleb <i>et al.</i> [?]	medical image segmentation	conditioned on image	Improved DDPM	BRATS
Sanchez <i>et al.</i> [?]	medical image segmentation and anomaly detection	conditioned on image and binary variable	ADM	BRATS
Pinaya <i>et al.</i> [?]	medical image segmentation and anomaly detection	conditioned on image	DDPM	MedNIST, UK Biobank Images, WMH, BRATS, MSLUB
Wolleb <i>et al.</i> [?]	medical image anomaly detection	conditioned on image	DDIM	CheXpert, BRATS
Wyatt <i>et al.</i> [?]	medical image anomaly detection	conditioned on image	ADM	NFBS, 22 MRI scans
Harvey <i>et al.</i> [?]	video generation	conditioned on frames	FDM	GQN-Mazes, MineRL Navigate, CARLA Town01
Ho <i>et al.</i> [?]	video generation	unconditional, conditioned on text	DDPM	101 Human Actions
Yang <i>et al.</i> [?]	video generation	conditioned on video representation	RVD	BAIR, KTH Actions, Simulation, Cityscapes
Höppe <i>et al.</i> [?]	video generation and infilling	conditioned on frames	RaMViD	BAIR, Kinetics-600, UCF-101
Giannone <i>et al.</i> [?]	few-shot image generation	conditioned on image	Improved DDPM	CIFAR-FS, mini-ImageNet, CelebA
Jeanneret <i>et al.</i> [?]	counterfactual explanations	unconditional	DDPM	CelebA
Sanchez <i>et al.</i> [?]	counterfactual estimates	conditional	ADM	MNIST, ImageNet
Kawar <i>et al.</i> [?]	image restoration	conditioned on image	DDIM	FFHQ, ImageNet
Özdenizci <i>et al.</i> [?]	image restoration	conditioned on image	DDPM	Snow100K, Outdoor-Rain, RainDrop
Kim <i>et al.</i> [?]	image registration	conditioned on image	DDPM	Radboud Faces, OASIS-3
Nie <i>et al.</i> [?]	adversarial purification	conditioned on image	Score SDE, Improved DDPM, DDIM	CIFAR-10, ImageNet, CelebA-HQ
Wang <i>et al.</i> [?]	semantic image generation	conditioned on semantic map	DDPM	Cityscapes, ADE20K, CelebAMask-HQ
Zhou <i>et al.</i> [?]	shape generation and completion	unconditional, conditional shape completion	DDPM	ShapeNet, PartNet
Zimmermann <i>et al.</i> [?]	classification	conditioned on label	DDPM++	CIFAR-10