视觉中的扩散模型：综述

摘要：降噪扩散模型代表了计算机视觉最近的一个新兴主题，展示了在生成式建模领域的显著成果。扩散模型是基于两个阶段的深度生成式模型，一个前向扩散阶段和一个反向扩散阶段。在前向扩散阶段，通过添加高斯噪声，输入数据在几个步骤中逐渐被扰动。在反向阶段，模型的任务是通过学习逐步反演扩散过程来恢复原始输入数据。尽管扩散模型具有已知的计算负担，即由于采样过程中涉及的大量步骤而导致的速度较低，但由于生成的样本的质量和多样性而被广泛认可。在本次调查中，我们对视觉中应用的降噪扩散模型的文章进行了全面的审查，包括该领域的理论和实践贡献。首先，我们确定并提出了三个通用的扩散建模框架，它们基于去噪扩散概率模型，噪声条件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式模型之间的关系，包括变分自动编码器、生成型预测网络、基于能量的模型、自回归模型和正则化流。然后，我们介绍了应用于计算机视觉的扩散模型的多视角分类。最后，我们阐述了扩散模型目前的局限性，并为未来的研究设想了一些有趣的方向。

1. 简介

扩散模型构成了一类深度生成式模型，最近在计算机视觉成为了最热门话题之一（图1），展示了令人印象深刻的生成能力，从高水平的细节到生成实例的多样性，我们甚至可以说这些生成式模型将生成式建模领域的标准提高到了一个新的水平，特别是Imagen和潜在扩散模型(LDM)等模型。图2所示的图像样本证实了这一说法，该图像样本由Stable Diffusion生成，这是基于文本提示生成图像的LDM版本。生成的图像伪影很少，且与文本提示相当一致。值得注意的是，我们特意选择了表示非现实场景（在训练时从未见过）的文本提示，从而展示了扩散模型的高泛化能力。到目前为止，扩散模型已经应用于各种生成式建模。

到目前为止，扩散模型已经应用于各种各样的生成式建模任务，例如图像生成、图像超分辨、图像修复、图像编辑和图像-图像翻译等。此外，研究发现通过扩散模型学习的潜在表征能应用在区分任务中，例如即时分割、分类和异常检测。这证实了去噪扩散模型的广泛适用性，表明其进一步的应用尚待发现。此外，学习强潜在表征的能力创建了与表征学习的联系，这是一个研究学习强大数据表示方式的综合领域，涵盖了多种方法，从设计新的神经架构到学习策略的发展。

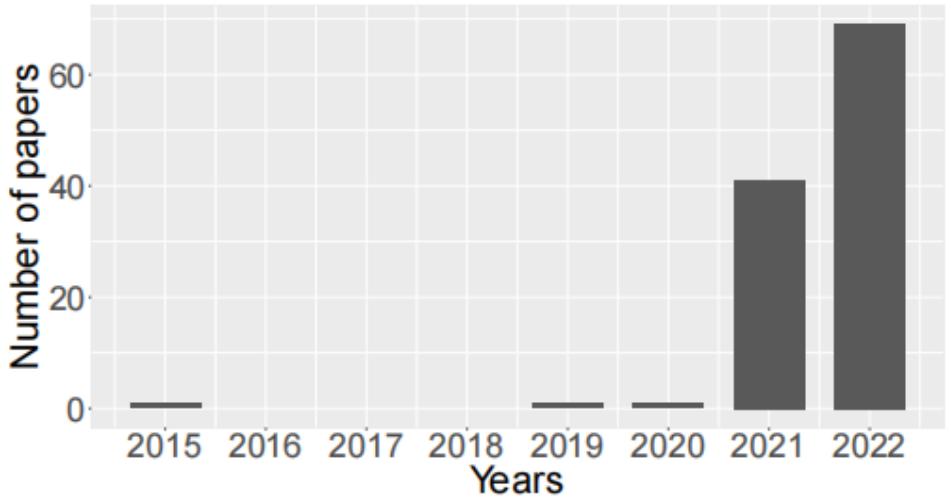

图1. 每年关于扩散模型的论文的大致数量

根据图1所示的图表，关于扩散模型的论文数量正在以非常快的速度增长。为了勾画这个快速发展的主题的过去和现在的成就的轮廓，我们提出了一个在计算机视觉去噪扩散模型的全面的文章综述。更准确地说，我们调研了属于下面定义的生成式模型类别的文章。扩散模型代表一类深度生成式模型，其基于（i）正向扩散阶段，其中输入数据通过添加高斯噪声在几个步骤上逐渐扰动，以及（ii）反向（后向）扩散阶段，其中生成式模型的任务是通过学习逐步反演扩散过程，逐步从扩散（噪声）数据中恢复原始输入数据。我们强调，至少有三个子类别的扩散模型符合上述定义。第一个子类别包括去噪扩散概率模型(DDPMs)，其灵感来自非平衡热力学理论。DDPM是使用潜在变量来估计概率分布的潜在变量模型。从这个角度来看，DDPM可以被看作是一种特殊的变分自动编码器(VAE)，其中正向扩散阶段对应于AE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，其基于通过分数匹配训练共享神经网络来估计不同噪声水平下的扰动数据分布的评分函数（定义为对数密度的梯度）。随机微分方程(SDEs)代表了模拟扩散的另一种方法，形成了扩散模型的第三子类。通过正向和反向SDE对扩散进行建模，可以得出有效的生成策略以及强大的理论结果。后一种形式(基于SDEs)可以被视为对DDPM和NCSN的推广。

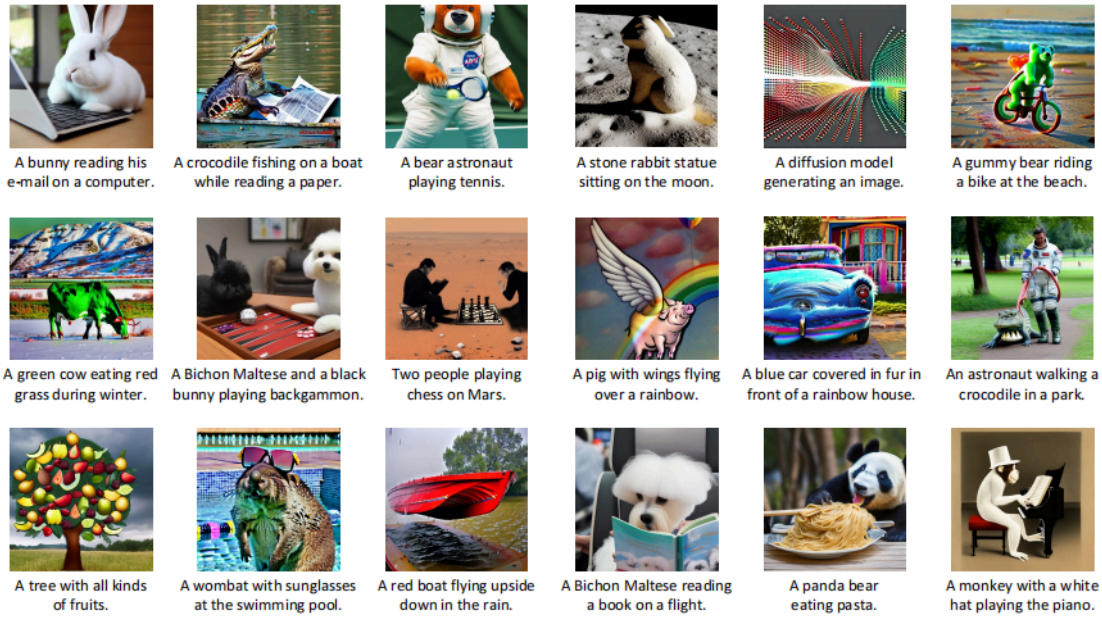

图2. Stable Diffusion基于各种文本提示生成的图像，通过

https://beta.dreamstudio.ai/dream平台

我们确定了几种定义的设计选择，并将它们合成为对应上述三个子类别的三个通用扩散建模框架。为了将通用扩散建模框架置于上下文中，我们进一步讨论了扩散模型与其他深度生成式模型之间的关系。更具体地说，我们描述了与变分自动编码器(VAE)、生成对抗网络(GAN)、基于能量的模型(EBM)、自回归的关系模型和标准化流。然后，我们介绍了一种应用于计算机视觉的扩散模型的多视角分类，基于多种标准对现有模型进行分类，例如底层框架，目标任务或降噪条件。最后，我们阐述了扩散模型目前的局限性，并为未来的研究设想了一些有趣的方向。例如，也许最可能的限制之一是推理过程中的时间效率低下，这是由于生成一个样本的评估步骤非常多，例如数千个。当然，在不影响生成样品质量的情况下克服这一限制是未来研究的重要方向。

总的来说，我们的贡献有两点：

- 由于最近在视觉领域出现了许多基于扩散模型的贡献，我们提供了一个全面而及时的文献综述，以应用于计算机视觉的去噪扩散模型，旨在为我们的读者提供对通用扩散　模型框架的快速理解。
- 我们设计了一个基于扩散模型的多视角分类旨在帮助其他研究应用于特定领域的扩散模型的研究人员快速找到相应领域的相关工作。

2. 通用框架

扩散模型是一类概率生成模型，它是一个学习逐渐削弱训练数据的过程的逆过程。因此，训练过程涉及两个阶段：前向扩散过程和后向去噪过程。前一个阶段由多个步骤组成，其中低水平噪声被添加到每个输入图像，其中噪声的比例在每个步骤上都不同。训练数

据逐渐被破坏，直到产生纯高斯噪声。后一阶段通过反演正向扩散过程来表示。使用相同的迭代过程，但相反：噪声被逐渐去除，因此，原始图像被重新创建。因此，在推理时，图像是通过从随机白噪声开始逐渐重建来生成的。在每个时间步长减去的噪声通过一个神经网络估计，通常基于U-NET架构，允许保留它的维度。

在接下来的三个小节中，我们提出了三种扩散模型的形式，即去噪扩散概率模型，噪声条件得分网络，以及基于随机微分方程的方法，该方法是前两种方法的通用。对于每个公式，我们描述了向数据添加噪声的过程，学习反演该过程的方法，以及如何在推理时生成新样本。在图3中，所有三个公式都作为一个通用框架进行了说明。我们将在最后一个小节讨论与其他深生成式模型的联系。

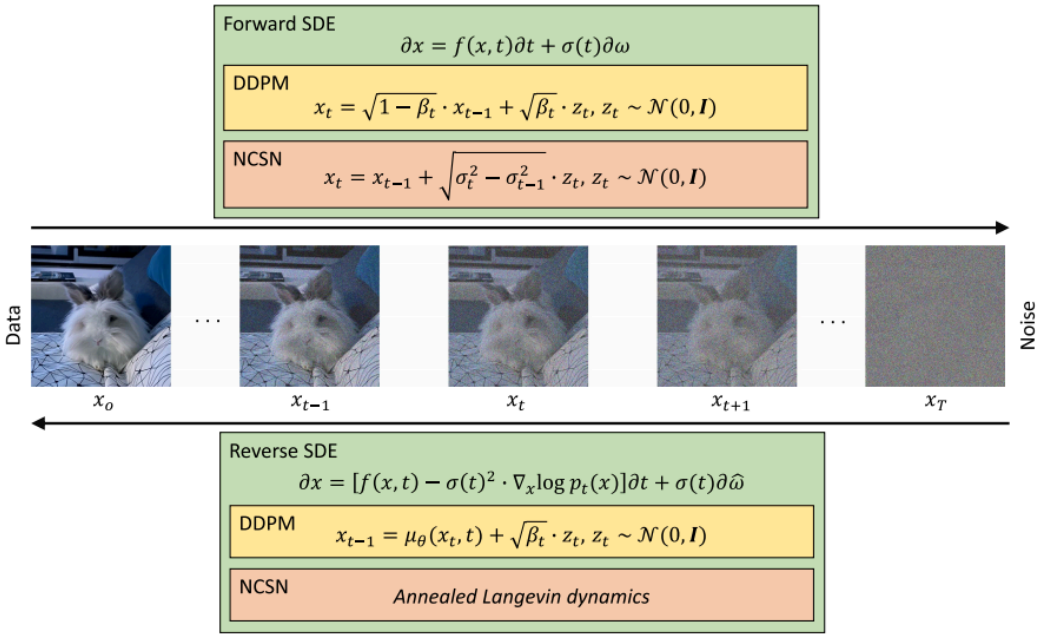

图3. 3种算法的图像表示

去噪扩散概率模型（DDPMs）

DDPMs的前向过程，使用高斯噪声缓慢破坏训练数据。P(x0)表示数据密度，其中下标0表示数据未被破坏。给定一个未损坏的训练样本$x_0 \sim p(x_0)$, 噪声版本$x_1, x_2 \ldots x_T$在如下的马尔可夫过程中获得。马尔可夫过程：

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

其中 T 是扩散步骤数，$\beta_1, \ldots, \beta_t \in [0;\ 1)$ 是表示跨扩散步骤的变量规划的超参数，I是与输入图像x_0具有相同维度的单位矩阵，N(x;μ, σ) 表示产生x的正态分布的均值和协方差。这种递归公式的一个重要特性是，当t从均匀分布中抽取时，它还允许对x_t进行直接采样，即，对于任意 t~U({1, . . . , T})，有

$$P\left(x_t|x_0\right)=N\left(x_t;\sqrt{\hat{B}_t}\cdot x_0,\left(1-\hat{\beta}_t\right)\cdot I\right) \tag{2}$$

本质上，式(2) 表明，如果我们有原始图像x_0并固定方差规划β_t，我们可以通过单个步骤对任何噪声版本x_t进行采样。

$P\left(x_t|x_0\right)$的采样是通过重新参数化技巧执行的。一般来说，为了标准化正态分布的样本x，我们减去均值并除以标准差获得正态分布 z~N(0 ， I)。重新参数化技巧执行此操作的逆操作，从z开始并通过将z与标准差相乘并加上均值来生成样本 x。如果我们将这个过程转化为我们的案例，那么x_t是从$P\left(x_t|x_0\right)$中采样的，如下所示：

$$x_t=\sqrt{\hat{\beta}_t}\cdot x_0+\sqrt{\left(1-\hat{\beta}_t\right)}\cdot z_t \tag{3}$$

z_t服从标准高斯分布

使得 $\hat{\beta}_t\rightarrow 0$，那么，根据等式 (2)，$x_T$的分布应该很好通过标准高斯分布 $\pi(x_T)$ = N(0,I) 来近似。此外，则反向步骤 $P\left(x_{t-1}|x_t\right)$与前向过程 $P\left(x_t|x_{t-1}\right)$ 具有相同的函数形式。直觉上，当x_t以非常小的步长创建时，最后的陈述是正确的，因为x_{t-1}更有可能来自靠近观察到 x_t的区域，这使我们能够使用高斯分布对该区域建模。为了符合上述特性，Ho等人选择$\left(\beta_t\right)_{t=1}^{T}<<1$是在$\beta_1=10^{-4}$和 $\beta_2=2\cdot10^{-2}$之间线性增加常数，其中 T= 1000。

对于逆过程。通过利用上述属性，我们可以从$P\left(x_0\right)$生成新样本，如果我们从样本 x_T~N(0,I) 开始并执行逆步骤，$P\left(x_{t-1}|x_t\right)=N\left(x_{t-1};u\left(x_t,t\right),\Sigma\left(x_t,t\right)\right)$，

在理想情况下，我们将使用最大似然目标训练神经网络，以便模型p分配给每个样本的x的概率尽可能的大。但是，$p(x_0)$难以处理，这是因为我们必须边缘化所有可能的反向轨迹。这个问题的解决方案是最小化负对数似然的变分下界。其中KL表示两个概率分布之间的Kullback-Leibler散度。表示两个概率分布之间的Kullback-Leibler散度。该目标的完整推导在附录A中给出。分析每个组件后，我们可以看到第二项可以删除，因为它不依赖于θ。

$$L_{vib}=-\ logp_\theta(x_0|x_1)\ +\ KL(p(x_T|x_0)||\pi(x_T)\ +\ \sum_{t>1}KL(p(x_{t-1}|x_t,\ x_0)||p_\theta(x_{t-1}|x_t))) \tag{4}$$

最后一项表明，神经网络经过训练，使得在每个时间步t，$p_\theta(x_{t-1}|x_t)$在以原始图像为条件时尽可能接近前向过程的真实后验。此外，可以证明后验$p(x_{t-1}|x_t,\ x_0)$是高斯分布，暗示KL散度的封闭形式表达式。Ho等人建议将协方差$\Sigma_\theta\left(x_t,t\right)$固定为常数值，重写$u_\theta\left(x_t,t\right)$：

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

这些简化（附录B中有更多详细信息）解锁了目标L_{vib}的新公式，它针对前向过程的随机时间步t，测量真实噪声z_t与噪声估计$z_\theta(x_t, t)$之间的距离：

其中E是期望值，$z_\theta(x_t, t)$是预测x_t中噪声的网络。我们强调x_t是通过等式(3)采样的，其中，我们使用训练集中的随机图象。

生成过程仍然由$p_\theta(x_{t-1}|x_t)$定义，但神经网络并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，并根据等式(5)确定平均值，而协方差固定为常数。算法1形式化了整个生成过程。

Algorithm 1: DDPM Sampling Method.

Input:
T – the number of diffusion steps.
$\sigma_1, \ldots, \sigma_T$ – 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** $t = T, \ldots, 1$ **do**
3: **if** $t > 1$ **then**
4: $z \sim \mathcal{N}(0, \mathbf{I})$
5: **else**
6: $z = 0$
7: $\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot (x_t - \frac{1-\alpha_t}{\sqrt{1-\hat{\beta}_t}} \cdot z_\theta(x_t, t))$
8: $x_{t-1} = \mu_\theta + \sigma_t \cdot z$

图2　算法1的基本流程

2.2　噪声条件评分网络（NCSNs）

某些数据密度p(x)的得分函数定义为相对于输入x的对数密度梯度$\nabla_x \log log\ P(x)$。Langevin　dynamics算法使用这些梯度给出的方向从随机样本(x_0)移动到高密度区域中的样本。Langevin　dynamics算法是一种受物理学启发的迭代方法，可用于数据采样。在物理学中，此方法用于确定分子系统中粒子的轨迹，该分子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拖曳力和分子间快速相互作用激发的随机力的影响。在我们的例子中，我们可以将对数密度的梯度视为将随机样本通过数据空间拖到具有高数据密度p(x)的区域的力。另一个术语ω_t在物理学中代表随机力，但对我们来说，它有助于逃避局部最小值。最后，γ的值加权了两种力的影响，因为它代表了粒子所在环境的摩擦系数。从采样的角度来看，控制更新的幅度。综上所述，Langevin dynamics算法的迭代更新如下：

$$x_i = x_{i-1} + \frac{r}{2}\nabla_x \log log\, P(x) \; + \sqrt{\gamma} w_i \tag{7}$$

其中，i∈{1, ..., N}，γ控制分数方向上更新的幅度，x_0是从先验分布中采样的，噪声w_i~N(0, I)解决了陷入局部最小值的问题，并且该方法递归地应用N趋于∞的步骤。因此，生成模型可以在用神经网络$s_\theta(x) \approx \nabla_x(x)$ 估计分数后，采用上述方法从p(x)中采样。该网络可以通过分数匹配进行训练，该方法需要优化以下目标：

$$L_s = E_x{\sim}P(x)\|S_\theta(x) - \nabla\times(x)\|_2^2 \tag{8}$$

尽管所描述的方法可用于数据生成，但Song等人强调了将该方法应用于实际数据时的几个问题。大多数问题都与流形假设有关。例如，当数据驻留在低维流形上时，分数估计s_θ(x)不一致，并且除其他影响外，这可能导致Langevin dynamics永远不会收敛到高密度区域。在同一工作中[3]，作者证明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，他们建议通过单个噪声条件得分网络（NCSN）学习所得噪声分布的得分估计。关于抽样，他们采用了等式(7)中的策略，并使用与每个噪声等级相关的分数估计。形式上，给定一系列高斯噪声尺度σ_1<σ_2<..<σ_T使得$P_{\sigma^T}(x)\approx N(0,I), p_{\sigma_1}(x)\approx p\left(x_0\right)$。我们可以通过去噪分数匹配来训练NCSN s_θ(x, σ_t)，使得对于任意t∈{1, ..., T}，有$s_\theta\left(x,\sigma_T\right) \approx \nabla_x \log log\left(p_{\sigma_T}(x)\right), P_\sigma\left(x_t|x\right) = N\left(x_t; x, \sigma_t^2 \cdot I\right)$其中$x_t$是x的噪声版本。因此，把等式(8)泛化到所有的高斯噪声尺度。并用等式（9）的形式替换所有梯度，则对于所有的t∈{1, ..., T}，训练$s_\theta(x_t, \sigma_t)$可通过最小化如下目标来实现：

$$L_s = \frac{1}{T}\sum_{t=1}^{T} \lambda\left(\sigma_T\right) E_D E_x{\sim}P\left(x_t|x\right)\|S_\theta\left(x_t, b_T\right) + \frac{x_t - x}{\sigma_t^2}\|_2^2 \tag{9}$$

其中，λ(σ_t)是加权函数。训练后，神经网络$s_\theta(x_t, \sigma_t)$将返回以噪声图像x_t和相应的时间步t作为输入的分数估计。

在推理时，Song等人介绍了算法2中正式描述的退火Langevin dynamics。他们的方法从白噪声开始，并应用等式（7）固定迭代次数。所需的梯度（分数）由以时间步T为条件的经过训练的神经网络给出。该过程继续执行以下时间步，将一个步骤的输出作为输入传播到下一步骤。 最终样本是t=0时返回的输出。

Algorithm 2: Annealed Langevin Dynamics.

Input:

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

N – the number of Langevin dynamics iterations.

$\gamma_1, \ldots, \gamma_T$ – 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** $t = T, \ldots, 1$ **do**

3: **For** $i = 1, \ldots, N$ **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: $x_{t-1}^0 = x_t^N$

图3　算法2的基本流程

2.3　随机微分方程（SDE）

与前两种方法类似，中提出的方法逐渐将数据分布p(x_0)转换为噪声。然而，它概括了前两种方法，因为在这种情况下，扩散过程被认为是连续的，从而成为随机微分方程（SDE）的解。这种扩散的逆过程可以用逆时间SDE来建模，它需要每个时间步的密度得分函数。因此，Song等人的生成模型使用神经网络来估计得分函数，并使用数值SDE求解器从p(x_0)生成样本。与NCSN的情况一样，神经网络接收扰动数据和时间步长作为输入，并生成分数函数的估计。前向扩散过程的SDE

$$\left(x_t\right)_t^T = 0 \tag{10}$$

其中，t∈[0, T]，具有以下形式：

$$\frac{\partial x}{\partial t} = f(x,t) + \sigma(t) \cdot w_t \tag{11}$$

其中w_t是高斯噪声，f是x和t的函数，用于计算漂移系数，σ是计算扩散系数的时间相关函数。为了将扩散过程作为该SDE的解决方案，应设计漂移系数，使其逐渐使数据x_0无效，而扩散系数控制添加多少高斯噪声。相关的逆时SDE定义如下：

$$\partial x = \left[f(x,t) - \sigma^2(t) \cdot \nabla_x \log log\, p_t(x)\right] \cdot \partial t + \sigma|t| \cdot \partial \hat{w} \tag{12}$$

其中，ˆw表示时间反转时的布朗运动，从T到0。逆时间SDE表明，如果我们从纯噪声开始，我们可以通过消除导致数据破坏的漂移来恢复数据。消除是通过减去

$$\sigma^2(t) \cdot \nabla_x(x) \tag{13}$$

我们可以通过优化与等式11中相同的目标来训练神经网络，其中是权重函数，t∈U[0, T]。我们强调，当漂移系数f是仿射时，$p_t(x_t|x_0)$是高斯分布。当f不符合这个属性时，我们不能使用去噪分数匹配，但我们可以回退到切片分数匹配。

Algorithm 3: Euler-Maruyama Sampling Method.

Input:
$\Delta t < 0$ – a negative step close to 0.
f – a function of x and t that computes the drift coefficient.
σ – 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: $\quad \Delta x = [f(x,t) - \sigma(t)^2 \cdot \nabla_x \log p_t(x)] \cdot \Delta t + \sigma(t) \cdot \Delta\hat{\omega}$
4: $\quad x = x + \Delta x$
5: $\quad t = t + \Delta t$

图4　算法3的基本流程

这种方法的采样可以使用应用于等13中定义的SDE的任何数值方法来执行。实际上，求解器不适用于连续公式。例如，Euler–Maruyama方法固定一个微小的负步长△t并执行算法3，直到初始时间步长t=T变为t=0。在第3步，布朗运动由下式给出，其中z~N(0, I)。这种方法的采样可以使用应用于等式(13)中定义的SDE的任何数值方法来执行。实际上，求解器不适用于连续公式。例如，Euler–Maruyama方法固定一个微小的负步长△t并执行算法3，直到初始时间步长t=T变为t=0。布朗运动由下式给出，其中z~N(0, I)。

Song等人提出了采样技术方面的一些贡献。他们引入了预测器-校正器　采样器，可以生成更好的示例。该算法首先采用数值方法从反时SDE中进行采样，然后使用基于分数的方法作为校正器，　例如上一小节中描述的退火Langevin　dynamics。此外，他们表明常微分方程（ODE）也可以用来模拟逆过程。因此，SDE解释解锁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点是其效率。

2.4与其他生成模型之间的关系

我们下面讨论扩散模型和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，以生成对抗网络结束。

扩散模型与VAE有更多共同点。

1.例如，在这两种情况下，数据都被映射到潜在空间，并且生成过程学习将潜在表示转换为数据。

2. 此外，在这两种情况下，目标函数都可以作为数据似然的下限导出。

3. 然而，这两种方法之间存在本质区别。

4. VAE的潜在表示包含有关原始图像的压缩信息，而扩散模型在前向过程的最后一步后完全破坏了数据。

5. 扩散模型的潜在表示与原始数据具有相同的维度，而VAE在维度减小时效果更好。

6. 最终，到VAE潜在空间的映射是可训练的，这对于扩散模型的前向过程来说是不正确的，因为如前所述，潜在空间是通过向原始图像逐渐添加高斯噪声来获得的。

7. 上述相似点和不同点可能是这两种方法未来发展的关键。例如，已经存在一些通过将扩散模型应用于VAE的潜在空间来构建更有效的扩散模型的工作。

自回归模型

1. 他们的生成过程通过以先前生成的像素为条件逐像素生成图像来生成新样本。

2. 这种方法意味着单向偏差，清楚地代表了此类生成模型的局限性。

3. 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。他们的方法学习通过马尔可夫链反转多项式扩散过程，其中每个转移都作为自回归模型实现。提供给自回归模型的全局信息由马尔可夫链的前一步给出。

归一化流

1. 该变换是通过一组可逆函数完成的，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。

2. 此类模型的一个重要特征是易于处理的似然。因此，训练的目标是负对数似然。

3. 与扩散模型相比，两种模型的共同点是数据分布到高斯噪声的映射。然而，这两种方法之间的相似之处到此为止，因为归一化流通过学习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，这些属性意味着对网络架构的额外约束以及可学习的前向过程。

4. 连接这两种生成算法的方法是DiffFlow，DiffFlow扩展了扩散模型和归一化流，使得反向和正向过程都是可训练的和随机的。

基于能量的模型

1. 由于这一特性，并且与之前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来表示。

2. 然而，由于这种灵活性，EBM的训练很困难。

3. 实践中使用的一种流行的训练策略是分数匹配。

4. 关于采样，除其他策略外，还有基于得分函数的马尔可夫链蒙特卡罗（MCMC）方法。

5. 因此，扩散模型第2. 2小节的公式可以被认为是基于能量的框架的特殊情况，正是训练和采样仅需要得分函数时的情况。

在扩散模型最近兴起之前，GAN就生成样本的质量而言被许多人认为是最先进的生成模

型。

1. GAN因其对抗性目标而难以训练，并且经常遭受模式崩溃。相比之下，扩散模型具有稳定的训练过程，并提供更多的多样性，因为它们是基于似然的。尽管有这些优点，但与GAN相比，扩散模型仍然效率低下，在推理过程中需要进行多次网络评估。

GAN和扩散模型之间比较的一个关键方面是它们的潜在空间。虽然GAN具有低维潜在空间，但扩散模型保留了图像的原始大小。此外，扩散模型的潜在空间通常被建模为随机高斯分布，类似于VAE。

3. 在语义属性方面，人们发现GAN的潜在空间包含与视觉属性相关的子空间。由于这个属性，可以通过潜在空间的变化来操纵属性。相反，当扩散模型需要这种变换时，首选过程是引导技术，它不利用潜在空间的任何语义属性。然而，Song等人证明扩散模型的潜在空间具有明确定义的结构，说明该空间中的插值导致图像空间中的插值。

4. 综上所述，从语义的角度来看，扩散模型的潜在空间的探索比GAN的情况要少得多，但这可能是社区未来遵循的研究方向之一。

3. 扩散模型的分类

TABLE 1
Our multi-perspective categorization of diffusion models applied in computer vision. To classify existing models, we consider three criteria: the task, the denoising condition, and the underlying approach (architecture). Additionally, we list the data sets on which the surveyed models are applied. We use the following abbreviations in the architecture column: D3PM (Discrete Denoising Diffusion Probabilistic Models), DSB (Diffusion Schrödinger Bridge), BDDM (Bilateral Denoising Diffusion Models), PNDM (Pseudo Numerical Methods for Diffusion Models), ADM (Ablated Diffusion Model), D2C (Diffusion-Decoding Models with Contrastive Representations), CCDF (Come-Closer-Diffuse-Faster), VQ-DDM (Vector Quantised Discrete Diffusion Model), BF-CNN (Bias-Free CNN), FDM (Flexible Diffusion Model), RVD (Residual Video Diffusion), RaMViD (Random Mask Video Diffusion).

Paper	Task	Denoising Condition	Architecture	Data Sets
Austin *et al.* [78]	image generation	unconditional	D3PM	CIFAR-10
Bao *et al.* [20]	image generation	unconditional	DDIM, Improved DDPM	CelebA, ImageNet, LSUN Bedroom, CIFAR-10
Benny *et al.* [79]	image generation	unconditional	DDPM, DDIM	CIFAR-10, ImageNet, CelebA
Bond-Taylor *et al.* [80]	image generation	unconditional	DDPM	LSUN Bedroom, LSUN Church, FFHQ
Choi *et al.* [81]	image generation	unconditional	DDPM	FFHQ, AFHQ-Dog, CUB, MetFaces
De *et al.* [82]	image generation	unconditional	DSB	MNIST, CelebA
Deasy *et al.* [83]	image generation	unconditional	NCSN	MNIST, Fashion-MNIST, CIFAR-10, CelebA
Deja *et al.* [84]	image generation	unconditional	Improved DDPM	Fashion-MNIST, CIFAR-10, CelebA
Dockhorn *et al.* [21]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10
Ho *et al.* [2]	image generation	unconditional	DDPM	CIFAR-10, CelebA-HQ, LSUN
Huang *et al.* [59]	image generation	unconditional	DDPM	CIFAR-10, MNIST
Jing *et al.* [30]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10, CelebA-256-HQ, LSUN Church
Jolicoeur *et al.* [85]	image generation	unconditional	NCSN	CIFAR-10, LSUN Church, Stacked-MNIST
Jolicoeur *et al.* [86]	image generation	unconditional	DDPM++, NCSN++	CIFAR-10, LSUN Church, FFHQ
Kim *et al.* [87]	image generation	unconditional	NCSN++, DDPM++	CIFAR-10, CelebA, MNIST
Kingma *et al.* [88]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Kong *et al.* [89]	image generation	unconditional	DDIM, DDPM	LSUN Bedroom, CelebA, CIFAR-10
Lam *et al.* [90]	image generation	unconditional	BDDM	CIFAR-10, CelebA

Liu *et al.* [91]	image generation	unconditional	PNDM	CIFAR-10, CelebA
Ma *et al.* [92]	image generation	unconditional	NCSN, NCSN++	CIFAR-10, CelebA, LSUN Bedroom, LSUN Church, FFHQ
Nachmani *et al.* [93]	image generation	unconditional	DDIM, DDPM	CelebA, LSUN Church
Nichol *et al.* [6]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Pandey *et al.* [19]	image generation	unconditional	DDPM	CelebA-HQ, CIFAR-10
San *et al.* [94]	image generation	unconditional	DDPM	CelebA, LSUN Bedroom, LSUN Church
Sehwag *et al.* [95]	image generation	unconditional	ADM	CIFAR-10, ImageNet
Sohl-Dickstein *et al.* [1]	image generation	unconditional	DDPM	MNIST, CIFAR-10, Dead Leaf Images
Song *et al.* [13]	image generation	unconditional	NCSN	FFHQ, CelebA, LSUN Bedroom, LSUN Tower, LSUN Church Outdoor
Song *et al.* [15]	image generation	unconditional	DDPM++	CIFAR-10, ImageNet 32×32
Song *et al.* [7]	image generation	unconditional	DDIM	CIFAR-10, CelebA, LSUN
Vahdat *et al.* [17]	image generation	unconditional	NCSN++	CIFAR-10, CelebA-HQ, MNIST
Wang *et al.* [96]	image generation	unconditional	DDIM	CIFAR-10, CelebA
Wang *et al.* [97]	image generation	unconditional	StyleGAN2, ProjectedGAN	CIFAR-10, STL-10, LSUN Bedroom, LSUN Church, AFHQ, FFHQ
Watson *et al.* [98]	image generation	unconditional	DDPM	CIFAR-10, ImageNet
Watson *et al.* [8]	image generation	unconditional	Improved DDPM	CIFAR-10, ImageNet 64×64
Xiao *et al.* [99]	image generation	unconditional	NCSN++	CIFAR-10
Zhang *et al.* [71]	image generation	unconditional	DDPM	CIFAR-10, MNIST
Zheng *et al.* [100]	image generation	unconditional	DDPM	CIFAR-10, CelebA, CelebA-HQ, LSUN Bedroom, LSUN Church
Bordes *et al.* [101]	conditional image generation	conditioned on latent representations	Improved DDPM	ImageNet
Campbell *et al.* [102]	conditional image generation	unconditional, conditioned on sound	DDPM	CIFAR-10, Lakh Pianoroll
Chao *et al.* [103]	conditional image generation	conditioned on class	Score SDE, Improved DDPM	CIFAR-10, CIFAR-100
Dhariwal *et al.* [5]	conditional image generation	unconditional, classifier guidance	ADM	LSUN Bedroom, LSUN Horse, LSUN Cat
Ho *et al.* [104]	conditional image generation	conditioned on label	DDPM	LSUN, ImageNet
Ho *et al.* [77]	conditional image generation	unconditional, classifier-free guidance	ADM	ImageNet 64×64, ImageNet 128×128
Karras *et al.* [105]	conditional image generation	unconditional, conditioned on class	DDPM++, NCSN++, DDPM, DDIM	CIFAR-10, ImageNet 64×64
Liu *et al.* [106]	conditional image generation	conditioned on text, image, style guidance	DDPM	FFHQ, LSUN Cat, LSUN Horse, LSUN Bedroom
Liu *et al.* [22]	conditional image generation	conditioned on text, 2D positions, relational descriptions between items, human facial attributes	Improved DDPM	CLEVR, Relational CLEVR, FFHQ
Lu *et al.* [107]	conditional image generation	unconditional, conditioned on class	DDIM	CIFAR-10, CelebA, ImageNet, LSUN Bedroom
Salimans *et al.* [108]	conditional image generation	unconditional, conditioned on class	DDIM	CIFAR-10, ImageNet, LSUN
Singh *et al.* [109]	conditional image generation	conditioned on noise	DDIM	ImageNet
Sinha *et al.* [16]	conditional image generation	unconditional, conditioned on label	D2C	CIFAR-10, CIFAR-100, fMoW, CelebA-64, CelebA-HQ-256, FFHQ-256
Ho *et al.* [34]	image-to-image translation	conditioned on image	Improved DDPM	ctest10k, places10k
Li *et al.* [37]	image-to-image translation	conditioned on image	DDPM	Face2Comic, Edges2Shoes, Edges2Handbags
Sasaki *et al.* [110]	image-to-image translation	conditioned on image	DDPM	CMP Facades, KAIST Multispectral Pedestrian
Wang *et al.* [36]	image-to-image translation	conditioned on image	DDIM	ADE20K, COCO-Stuff, DIODE
Wolleb *et al.* [38]	image-to-image translation	conditioned on image	Improved DDPM	BRATS
Zhao *et al.* [35]	image-to-image translation	conditioned on image	DDPM	CelebaA-HQ, AFHQ
Gu *et al.* [111]	text-to-image generation	conditioned on text	VQ-Diffusion	CUB-200, Oxford 102 Flowers, MS-COCO
Jiang *et al.* [23]	text-to-image generation	conditioned on text	Transformer-based encoder-decoder	DeepFashion-MultiModal
Ramesh *et al.* [112]	text-to-image generation	conditioned on text	ADM	MS-COCO, AVA
Rombach *et al.* [11]	text-to-image generation	conditioned on text	LDM	OpenImages, WikiArt, LAION-2B-en, ArtBench
Saharia *et al.* [12]	text-to-image generation	conditioned on text	Imagen	MS-COCO, DrawBench
Shi *et al.* [9]	text-to-image generation	unconditional, conditioned on text	Improved DDPM	Conceptual Captions, MS-COCO
Zhang *et al.* [113]	text-to-image generation	unconditional, conditioned on text	DDIM	CIFAR-10, CelebA, ImageNet
Daniels *et al.* [25]	super-resolution	conditioned on image	NCSN	CIFAR-10, CelebA

Hu *et al.* [118]	multi-task (image generation, inpainting)	unconditional, conditioned on image	VQ-DDM	CelebA-HQ, LSUN Church
Khrulkov *et al.* [119]	multi-task (image generation, image-to-image translation)	conditioned on class	Improved DDPM	AFHQ, FFHQ, MetFaces, ImageNet
Kim *et al.* [120]	multi-task (image translation, multi-attribute transfer)	conditioned on image, portrait, stroke	DDIM	ImageNet, CelebA-HQ, AFHQ-Dog, LSUN Bedroom, Church
Luo *et al.* [121]	multi-task (point cloud generation, auto-encoding, unsupervised representation learning)	conditioned on shape latent	DDPM	ShapeNet
Lyu *et al.* [122]	multi-task (image generation, image editing)	unconditional, conditioned on class	DDPM	CIFAR-10, CelebA, ImageNet, LSUN Bedroom, LSUN Cat
Preechakul *et al.* [123]	multi-task (latent interpolation, attribute manipulation)	conditioned on latent representation	DDIM	CelebA-HQ
Rombach *et al.* [10]	multi-task (super-resolution, image generation, inpainting)	unconditional, conditioned on image	VQ-DDM	ImageNet, CelebA-HQ, FFHQ, LSUN
Shi *et al.* [124]	multi-task (super-resolution, inpainting)	conditioned on image	Improved DDPM	MNIST, CelebA
Song *et al.* [3]	multi-task (image generation, inpainting)	unconditional, conditioned on image	NCSN	MNIST, CIFAR-10, CelebA
Kadkhodaie *et al.* [125]	multi-task (Spatial super-resolution, Deblurring, Compressive sensing, Inpainting, Random missing pixels)	conditioned on linear measurements	BF-CNN	MNIST, Set5, Set68, Set14
Song *et al.* [4]	multi-task (image generation, inpainting, colorization)	unconditional, conditioned on image, class	NCSN++, DDPM++	CelebA-HQ, CIFAR-10, LSUN
Hu *et al.* [126]	medical image-to-image translation	conditioned on image	DDPM	ONH
Chung *et al.* [127]	medical image generation	conditioned on measurements	NCSN++	fastMRI knee
Özbey *et al.* [128]	medical image generation	conditioned on image	Improved DDPM	IXI, Gold Atlas - Male Pelvis
Song *et al.* [129]	medical image generation	conditioned on measurements	NCSN++	LIDC, LDCT Image and Projection, BRATS
Wolleb *et al.* [41]	medical image segmentation	conditioned on image	Improved DDPM	BRATS
Sanchez *et al.* [130]	medical image segmentation and anomaly detection	conditioned on image and binary variable	ADM	BRATS
Pinaya *et al.* [44]	medical image segmentation and anomaly detection	conditioned on image	DDPM	MedNIST, UK Biobank Images, WMH, BRATS, MSLUB
Wolleb *et al.* [45]	medical image anomaly detection	conditioned on image	DDIM	CheXpert, BRATS
Wyatt *et al.* [46]	medical image anomaly detection	conditioned on image	ADM	NFBS, 22 MRI scans
Harvey *et al.* [131]	video generation	conditioned on frames	FDM	GQN-Mazes, MineRL Navigate, CARLA Town01
Ho *et al.* [132]	video generation	unconditional, conditioned on text	DDPM	101 Human Actions
Yang *et al.* [133]	video generation	conditioned on video representation	RVD	BAIR, KTH Actions, Simulation, Cityscapes
Höppe *et al.* [134]	video generation and infilling	conditioned on frames	RaMViD	BAIR, Kinetics-600, UCF-101
Giannone *et al.* [135]	few-shot image generation	conditioned on image	Improved DDPM	CIFAR-FS, mini-ImageNet, CelebA
Jeanneret *et al.* [136]	counterfactual explanations	unconditional	DDPM	CelebA
Sanchez *et al.* [137]	counterfactual estimates	conditional	ADM	MNIST, ImageNet
Kawar *et al.* [27]	image restoration	conditioned on image	DDIM	FFHQ, ImageNet
Özdenizci *et al.* [138]	image restoration	conditioned on image	DDPM	Snow100K, Outdoor-Rain, RainDrop
Kim *et al.* [139]	image registration	conditioned on image	DDPM	Radboud Faces, OASIS-3
Nie *et al.* [140]	adversarial purification	conditioned on image	Score SDE, Improved DDPM, DDIM	CIFAR-10, ImageNet, CelebA-HQ
Wang *et al.* [141]	semantic image generation	conditioned on semantic map	DDPM	Cityscapes, ADE20K, CelebAMask-HQ
Zhou *et al.* [142]	shape generation and completion	unconditional, conditional shape completion	DDPM	ShapeNet, PartNet
Zimmermann *et al.* [43]	classification	conditioned on label	DDPM++	CIFAR-10

考虑到不同的分离标准，我们使用多视角分类法对扩散模型分类。也许分离模型的最重要标准是由（i）它们所应用的任务和（ii）它们所需的输入信号定义的。此外，由于制定扩散模型有多种方法，（iii）底层框架是扩散模型分类的另一个关键因素。最后，训练和评估过程中使用的(iv)数据集也非常重要，因为它们提供了在同一任务上比较不同模型的方法。我们根据上面列举的标准对扩散模型进行分类，如表1所示。

在本节的其余部分中，我们提出了对扩散模型的一些贡献，选择目标任务作为分离方法的主要标准。我们选择这种分类标准是因为它对于扩散模型的研究来说相当平衡且具有代表性，有助于从事特定任务的读者快速掌握相关作品。虽然主要任务通常与图像生成相关，但已经进行了大量工作来匹配甚至超越GAN在其他主题上的性能，例如超分辨率、修复、图像编辑、图像到图像翻译或分割。

3.1无条件图像生成

下面介绍的扩散模型用于在无条件设置下生成样本。这种模型不需要监督信号，完全不受监督。我们认为这是图像生成的最基本和通用的设置。

3.1.1　去噪扩散概率模型

Sohl-Dickstein等人的工作将扩散模型形式化，如第2.1节所述。所提出的神经网络基于包含多尺度卷积的卷积架构。

Austin等人扩展了Sohl-Dickstein等人的方法，针对离散扩散模型，研究前向过程中使用的转移矩阵的不同选择。他们的结果与之前用于图像生成任务的连续扩散模型具有竞争力。

Ho等人扩展了中提出的工作，提出通过估计每一步图像中的噪声来学习逆过程。这一变化导致了一个应用的去噪分数匹配的标。为了预测图像中的噪声，作者使用了中介绍的Pixel-CNN++架构。

在Ho等人提出的工作之上。Nichol等人介绍了一些改进，观察到线性噪声表对于低分辨率来说不是最佳的。他们提出了一种新的选择，可以避免在转发过程结束时快速破坏信息。此外，他们表明需要学习方差才能提高扩散模型在对数似然方面的性能最后一项更改允许更快的采样，大约需要50个步骤。

Song等人用非马尔可夫过程替换。生成过程发生变化，模型首先预测正常样本，然后用于估计链中的下一步。这一变化导致采样过程更快，对生成样本的质量影响很小。由此产生的框架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），这是一种在编码器产生的潜在表示上训练扩散模型的生成方法。该框架提出的DDPM架构，通过将潜在表示映射到图像来生成图像。

作者提出了一种在推理时给定当前输入的情况下估计噪声参数的方法。他们的改变提高了FID同时需要更少的步骤。作者用VGG-11来估计噪声参数，并使用DDPM来生成图像。

Nachmani等的工作建议用另外两个分布（两个高斯分和Gamma分布的混合）替换扩散过程的高斯噪声分布。由于Gamma分布具有更高的建模能力，结果显示出更好的FID值和更快的收敛速度。

Lam等人学习采样的噪声计划。训练的噪声计划与以前一样保性。训练得分网络后，他们假设它接近最优值，以便将其用于噪声计划训练。推理由两个步骤组成。首先，通过固定两个初始超参数来确定计划。第二步是按照确定的计划进行通常的逆过程。

Bond-Taylor等人提出了一个两阶段的过程，其中他们将矢量量化（vector　quantization）应用于图像以获得离散表示。transformer来反转离散扩散过程，其中

元素在每一步被随机掩蔽。 采样过程更快，因为扩散应用于高度压缩的表示，从而允许更少的去噪步骤(50-256)。

Watson等人提出了一种动态编程算法，可以找到最佳推理计划，时间复杂度为O(T)，其中T是步骤数。他们使用DDPM架构在CIFAR-10和ImageNet上进行图像生成实验。

在另一项工作中，Watson等人[8]首先介绍如何将重参数化技巧集成到扩散模型的后向过程中，以优化一系列快速采样器。 他使用内核初始距离（Kernel Inception Distance）作为损失函数，展示了如何使用随机梯度下降来完成优化。接下来，他们提出了一个特殊的参数化采样器系列，该采样器使用与以前相同的过程，可以通过更少的采样步骤获得有竞争力的结果。使用FID和初始分数(IS)作为指标，该方法似乎优于某些扩散模型基线。

类似于Bond-Taylor等人和Watson等人，肖等人尝试提高采样速度，同时也保持样本的质量、覆盖度和多样性。他们的方法是将GAN集成到去噪过程中，以区分真实样本（前向过程）和假样本（来自生成器的去噪样本），目标是最小化软化反向KL散度。然而，可通过直接生成干净的（完全去噪的）样本并在其上调节假样本来修改模型。使用具有适用于GAN生成器的自适应组归一化的NCSN++架构，它们在图像合成和基于笔画的图像生成中实现了类似的FID值，采样率比其他扩散模型快约20至2000倍。

Kingma等人引入了一类扩散模型，可以获得图像密度估计的最先进的可能性。他们将傅里叶特征添加到网络的输入中以预测噪声，并调查观察到的改进是否特定于此类模型。他们的结果证实了这一假设，即以前最先进的模型并没有从这一变化中受益。作为理论贡献，他们表明扩散损耗仅在其极端情况下受到信噪比函数的影响。

Bao等人提出了一种不需要使用非马尔可夫扩散过程进行训练的推理框架。通过首先导出关于评分函数的最佳均值和方差的分析估计，并使用预训练的基于评分的模型来获取评分值，它们显示出更好的结果，同时时间效率提高了20到40倍。该分数通过蒙特卡罗采样近似得出。然而，为了减少预训练DDPM模型的任何偏差，分数被限制在一些预先计算的范围内。

Zheng等人建议在任意步骤截断该过程，并提出一种通过放宽将高斯随机噪声作为前向扩散的最终输出的约束来反转该分布的扩散的方法。为了解决从不可处理分布开始逆向过程的问题，使用隐式生成分布来匹配扩散数据的分布。代理分布通过GAN或条件转移进行拟合。我们注意到生成器使用与扩散模型的采样器相同的U-Net模型，因此没有添加额外的参数进行训练。

Deja等人首先分析扩散模型的后向过程，并假设它由两个模型组成：生成器和降噪器因此，他们建议明确地将过程分为两个部分：通过自动编码器的降噪器和通过扩散模型的生成器。两种模型都使用相同的U-Net架构。

Wang 等人从 Arjovsky 等人和 Sønderby 等人提出的想法开始，通过添加噪声来增强

鉴别器的输入数据。这是通过从高斯混合分布注入噪声来实现的，该高斯混合分布由不同时间步长的干净图像的加权扩散样本组成。 噪声注入机制适用于真实图像和假图像。 这些实验是在涵盖多种分辨率和高多样性的广泛数据集上进行的。

3.1.2 基于分数的生成模型

从之前的工作开始, Song 等人提出了一些基于理论和实证分析的改进。它们涉及训练和采样阶段. 对于训练，作者展示了选择噪声尺度以及如何将噪声调节纳入NCSN 的新策略。对于采样，他们建议对参数应用指数移动平均值（EMA）, 并为 Langevin dynamics 选择超参数，以便步长验证某个方程. 拟议的更改解锁了 NCSN 在高分辨率图像上的应用。

Jolicoeur-Martineau 等人引入对抗性目标和去噪分数匹配来训练基于分数的模型。 此外，他们提出了一种称为一致退火采样（Consistent Annealed Sampling）的新采样程序，并证明它比annealed Langevin方法更稳定。 他们的图像生成实验表明，新的目标会返回更高质量的示例，而不会影响多样性。

Song 等人提高基于分数的扩散模型的可能性。 他们通过结合分数匹配损失的新加权函数来实现这一目标。 对于图像生成实验，他们使用DDPM++ 架构。

作者提出了一种基于分数的生成模型作为迭代比例拟合（IPF）的实现，这是一种用于解决薛定谔桥（Schr ¨ odinger bridge）问题的技术。这种新颖的方法在图像生成以及数据集插值上进行了测试，这是可能的，因为先验可以是任何分布。

Vahdat 特等人训练潜在表征的扩散模型。 他们使用 VAE 对潜在空间进行编码和解码。 这项工作实现了高达56倍的采样速度。 对于图像生成实验，作者采用NCSN++ 架构。

3.1.3 随机微分方程

DiffFlow 作为一种新的生成建模方法被引入，它结合了归一化流和扩散概率模型。从扩散模型的角度来看，由于可学习的前向过程跳过了不需要的噪声区域，该方法的采样过程效率提高了20倍。

Jolicoeur-Martineau 等人引入了一种新的 SDE 求解器，它比 Euler-Maruyama 快2到5倍, 并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使用预训练模型进行评估。

Wang 等人提出了一种基于薛定谔桥的新深度生成模型. 这是一个两阶段方法，其中第一阶段学习目标分布的平滑版本，第二阶段导出实际目标。

Dockhorn 等人专注于基于评分的模型，通过向数据添加另一个变量（速度）来利用临界阻尼朗之万（Langevin）扩散过程，这是该过程中唯一的噪声源. 考虑到新的扩散空间，所得到的得分函数被证明更容易学习。作者扩展了他们的工作，开发了一种更合适的评分

目标（称为混合评分匹配），以及一种采样方法，通过积分求解 SDE。 作者采用 NCSN++ 和 DDPM++ 架构来接受数据和速度，在无条件图像生成上进行评估，并优于类似的基于分数的扩散模型。

受高斯噪声分布导致的基于高维分数的扩散模型的限制的启发，Deasy 等人扩展去噪分数匹配以推广到正态噪声分布。通过添加更重的尾部分布，他们在多个数据集上的实验显示了有希望的结果，因为在某些情况下（取决于分布的形状）生成性能得到了提高。该方法擅长的一个重要场景是具有不平衡类的数据集。

Jing 等人尝试通过减少实现扩散的空间来缩短扩散模型采样过程的持续时间，即扩散过程中的时间步长越大，子空间越小。 数据在特定时间被投影到一组有限的子空间上，每个子空间都与一个分数模型相关联。这会降低计算成本，同时提高性能。这项工作仅限于自然图像合成。在评估无条件图像生成中的方法时，作者与最先进的模型相比获得了相似或更好的性能，同时推理时间更短。该方法被证明也适用于修复任务。

Kim 等人建议将扩散过程改为非线性过程。这是通过使用可训练的归一化流模型来实现的，该模型对潜在空间中的图像进行编码，现在可以将图像线性扩散到噪声分布。然后将类似的逻辑应用于去噪过程。 该方法应用于NCSN++和DDPM++框架，而归一化流模型基于ResNet。

Ma 等人旨在使后向扩散过程更加省时，同时保持合成性能。 在基于分数的扩散模型系列中，他们开始分析频域中的反向扩散，随后将空频滤波器应用于采样过程，旨在将有关目标分布的信息集成到初始噪声采样中。 作者使用 NCSN和 NCSN++进行了实验，其中所提出的方法清楚地显示了图像合成的速度改进（采样步骤减少了20倍），同时对于低图像和高图像保持了相同令人满意的生成质量。 分辨率图像。

3.2 条件图像生成

接下来我们展示应用于条件图像合成的扩散模型。该条件通常基于各种源信号，在大多数情况下使用一些类标签。 有些方法同时执行无条件和条件生成，这也在此处讨论。

3.2.1 去噪扩散概率模型

Dhariwal 等人引入了一些架构更改来改进扩散模型的 FID。 他们还提出了分类器指导，这是一种使用分类器的梯度来指导采样过程中的扩散的策略。他们进行了无条件和条件图像生成实验。

Bordes 等人通过可视化并将其与原始图像进行比较来检查自我监督任务产生的表示。他们还比较了不同来源生成的表示。 因此，扩散模型用于生成以这些表示为条件的样本。作者对 Dhariwal 等人提出的 U-Net 架构进行了一些修改，例如添加条件批量归一化层，并通过全连接层映射向量表示。

中提出的方法允许扩散模型从数据流形的低密度区域生成图像。他们使用两个新的损

失来指导相反的过程。 第一个损失将扩散引导至低密度区域，而第二个损失则强制扩散停留在流形上。 此外，他们证明他们的扩散模型不会记住低密度邻域的示例，从而生成新颖的图像。 作者采用了类似于 Dhariwal 等人的架构。

Kong 等人定义了连续扩散步骤和噪声水平之间的双射。通过定义的双射，他们能够构建一个需要更少步骤的近似扩散过程。该方法使用之前的DDIM和DDPM架构在图像生成上进行了测试。

Pandey 等构建了一个生成器-精炼器框架，其中生成器是VAE，精炼器是由VAE输出调节的DDPM。VAE的潜在空间可用于控制生成图像的内容，因为DDPM仅添加细节。训练框架后，生成的 DDPM 能够泛化到不同的噪声类型。 更具体地说，如果逆过程不以 VAE 的输出为条件，而是以不同的噪声类型为条件，则 DDPM 能够重建初始图像。

Ho 等人介绍了级联扩散模型（CDM），一种以ImageNet类为条件生成高分辨率图像的方法。 他们的框架包含多个扩散模型，其中流程中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。这些模型以类别和低分辨率图像为条件。

Benny 等人研究了逆向过程中预测图像而不是噪声的优点和缺点。他们的结论是，一些发现的问题可以通过插入两种类型的输出来解决。他们修改了以前的架构以返回噪声和图像，以及在执行插值时控制噪声重要性的值。该策略是在 DDPM 和 DDIM 架构之上进行评估的。

Choi 等人研究噪声水平对扩散模型学习的视觉概念的影响。 他们将目标函数的传统加权方案修改为一种新的加权方案，强制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类（粗略、内容和清理）(coarse, content and clean-up)，即小SNR为粗略，中SNR为内容，大SNR为清理。加权函数为最后一组分配较低的权重。

Singh 等人提出了一种用于条件图像生成的新方法。 他们没有在整个采样过程中调节信号，而是提出了一种调节噪声信号（从采样开始的地方）的方法。 使用反转梯度，噪声被注入有关条件类的定位和方向的信息，同时保持相同的随机高斯分布。

Liu 等人描述了扩散模型和基于能量的模型的相似功能，并利用了后者模型的组成结构。 建议结合多个扩散模型进行条件图像合成。 在逆过程中，多个扩散模型的组合，每个模型与不同的条件相关，可以通过合并或求反来实现。

3.2.2 基于分数的生成模型

Song 等人和 Dhariwal 等人的作品基于分类器指导的基于评分的条件扩散模型启发了 Chao 等人开发一个新的训练目标，减少分数模型和真实分数之间的潜在差异。 分类器的损失被修改为缩放交叉熵，并添加到修改后的分数匹配损失中。

3.2.3 随机微分方程

Ho 等人引入了一种不需要分类器的指导方法。 它只需要一个条件扩散模型和一个无条件版本，但他们使用相同的模型来学习这两种情况。 无条件模型是在类标识符等于 0 的情况下进行训练的。这个想法基于从贝叶斯规则导出的隐式分类器。

Liu 等人研究了使用传统数值方法来求解逆过程的ODE公式。 他们发现与以前的方法相比，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：梯度部分和迁移部分。迁移部分（标准方法具有线性迁移部分）被替换，使得结果尽可能接近目标流形。 最后一步，他们展示了这种变化如何解决使用传统方法时发现的问题。Tachibana 等人解决了DDPM的缓慢采样问题。他们建议通过增加随机微分方程求解器（去噪部分）的阶数（从一阶到二阶）来减少采样步骤数。 在保留网络架构和分数匹配功能的同时，他们对采样器采用 Itˆo-Taylor 展开方案，并替换一些导数项以简化计算。 它们减少了逆步骤的数量，同时保留了性能。 除此之外，另一个贡献是新的噪声计划。

Karras等人尝试将基于扩散评分的模型分成彼此独立的各个组件。 这种分离允许修改单个部分而不影响其他单元，从而促进扩散模型的改进。利用这个框架，作者首先提出了一个使用 Heun的方法作为ODE求解器的采样过程，在保持FID分数的同时减少了神经功能评估。 他们进一步表明，随机采样过程带来了巨大的性能优势。 第二个贡献与通过对神经网络的输入和相应目标进行预处理以及使用图像增强来训练基于分数的模型有关。

在无条件和类条件图像生成的背景下，Salimans 等人提出了一种减少采样步骤数量的技术。他们将由确定性 DDIM 表示的训练有素的教师模型的知识提炼成具有相同架构的学生模型，但采样步骤数减半。换句话说，学生的目标是连续走老师的两步。此外，可以重复该过程，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，探索了模型的三个版本和两个损失函数，以促进蒸馏过程并减少采样步骤数（从8192到4）。

Campbell 等人展示了能够处理离散数据的去噪扩散模型的连续时间公式。 该工作通过转移率矩阵对前向连续时间马尔可夫链扩散过程进行建模，并通过逆转移率矩阵的参数近似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有关。

Song 等人提出将扩散模型解释为ODE，Lu等人将其重新表述为可以使用指数积分器求解的形式。Lu等人的其他贡献是一个ODE求解器，它使用泰勒展开（一阶到三阶）来近似新公式的积分项，以及一个适应时间步安排的算法，速度快了4到16倍。

3.3 图像到图像的转换

Saharia 等人提出了一个使用扩散模型进行图像到图像转换的框架，重点关注四个任务：着色、修复、取消裁剪和JPEG恢复。所提出的框架在所有四个任务中都是相同的，这意味着它不会对每个任务进行自定义更改。 作者首先比较了 L1 和 L2 损失，表明 L2 是首选，因为它会带来更高的样本多样性。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。

为了转换一组未配对的图像，Sasaki 等人提出了一种涉及两个联合训练的扩散模型的方法。在反向去噪过程中，每个模型的每一步都以另一个模型的中间样本为条件。 此外，扩散模型的损失函数使用循环一致性损失进行正则化。

Zhao等人的目标是通过利用来自具有同等重要性的源域的数据来改进当前基于图像到图像转换分数的扩散模型。采用在源域和目标域上训练的基于能量的函数来指导 SDE 求解器。 这导致保留了与域无关的特征的图像的生成，同时将特定于源域的特征转换为目标域。 能量函数基于两个特征提取器，每个特征提取器特定于一个域。

Wang 等人利用预训练的力量，采用 GLIDE 模型并对其进行训练以获得丰富的语义潜在空间。从预训练版本开始，替换头部以适应任何条件输入，该模型在一些特定的图像生成下游任务上进行了微调。这是分两步完成的，第一步是冻结解码器并仅训练新的编码器，第二步是同时训练它们。最后，作者采用对抗性训练并对无分类器指导进行标准化，以提高生成质量。

Li 等人引入了一种基于布朗桥（Brownian bridges）和GAN的图像到图像转换的扩散模型。 所提出的过程首先使用 VQ-GAN 对图像进行编码。在生成的量化潜在空间中，扩散过程（公式化为布朗桥）在源域和目标域的潜在表示之间进行映射。 最后，另一个 VQGAN 对量化向量进行解码，以便在新域中合成图像。 这两个 GAN 模型在各自的特定领域独立训练。

Wolleb 等人继续他们的先前工作， 通过用特定于任务的另一个模型替换分类器来扩展扩散模型。 因此，在采样过程的每一步中，都会注入特定于任务的网络的梯度。 该方法通过回归器（基于编码器）或分割模型（使用 U-Net 架构）进行演示，而扩散模型则基于现有框架。此设置的优点是无需重新训练整个扩散模型（特定于任务的模型除外）。

3.4 文本到图像的合成

也许扩散模型最令人印象深刻的结果是在文本到图像的合成上获得的，其中组合不相关的概念（例如对象、形状和纹理）以生成不寻常的示例的能力得到了体现。 为了证实这一说法，我们使用稳定扩散根据各种文本提示生成图像，结果如图5所示。

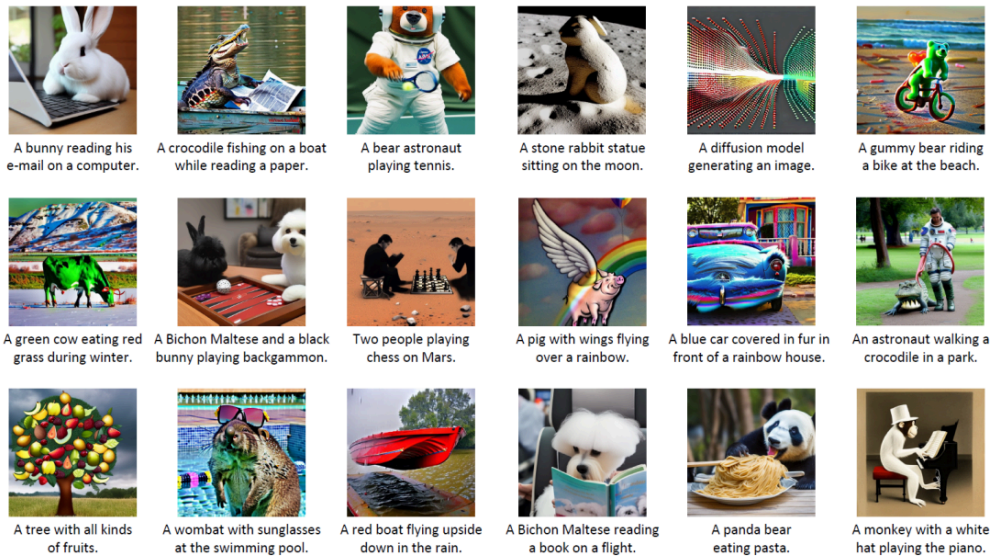

Fig. 2. Images generated by Stable Diffusion [10] based on various text prompts, via the https://beta.dreamstudio.ai/dream platform.

图5 文本生成图像的结果

Imagen 被介绍为一种文本到图像合成的方法。 它由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扩散模型组成。 这些模型还以编码器返回的文本嵌入为条件。 此外，作者还引入了一组新的标题（captions） (DrawBench)，用于文本到图像的评估。 在架构方面，作者开发了 Efficient U-Net 来提高效率，并将该架构应用到他们的文本到图像生成实验中。

Gu 等人引入了 VQ-Diffusion 模型，这是一种文本到图像合成的方法，它不存在先前方法的单向偏差。通过其掩蔽机制，该方法避免了推理过程中错误的累积。 该模型有两个阶段，第一阶段基于 VQ-VAE，它学习通过离散 tokens 表示图像，第二阶段是在 VQ-VAE 的离散潜在空间上运行的离散扩散模型。 扩散模型的训练以标题（captions）嵌入为条件。 受掩码语言模型的启发，一些标记被替换为[mask]token。

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扩散模型。 这是一种两阶段方法，其中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生成最终图像。 为了生成图像嵌入，作者在潜在空间中使用了扩散模型。 他们进行主观的人类评估来评估他们的生成结果。

为了解决扩散模型缓慢采样的不便，Zhang 等人他们的工作重点是一种新的离散化方案，该方案可以减少误差并允许更大的步长，即更少的采样步骤数。 通过在得分函数中使用高阶多项式外推以及用于求解反向 SDE 的指数积分器，网络评估的数量大大减少，同时保持生成能力。

Shi 等人结合VQ-VAE和扩散模型来生成图像。从VQ-VAE开始，编码功能被保留，而解码器被扩散模型取代。 作者使用中的U-Net 架构，将图像标记注入到中间块中。

Rombach 等人引入了使用相同过程创建艺术图像的修改：从数据库中提取图像的CLIP潜在空间中的 k 近邻，然后通过使用这些嵌 入指导反向去噪过程来生成新图像 。 由于

CLIP潜在空间由文本和图像共享，因此也可以通过文本提示来引导扩散。 然而，在推理时，数据库被替换为另一个包含艺术图像的数据库。 因此，该模型会生成新数据库风格的图像。

Jiang 等人提出了一个框架，可以在给定三个输入的情况下生成具有丰富服装表示的全身人体图像：人体姿势、衣服形状的文本描述以及服装纹理的另一个文本。 该方法的第一阶段将前文本提示编码为嵌入向量，并将其注入到生成表单地图的模块（基于编码器-解码器）中。 在第二阶段，基于扩散的转换器从多个多级码本（每个特定于纹理）对后一个文本提示的嵌入表示进行采样，这是 VQ-VAE中建议的机制。 最初，对较粗级别的码本索引进行采样，然后使用前馈网络预测较精细级别的索引。 文本使用 Sentence-BERT进行编码。

3.5 图像超分辨率

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以低分辨率版本为条件的高质量图像。 这项工作采用架构以及以下数据集：CelebA-HQ、FFHQ 和 ImageNet。Daniels 等人使用基于分数的模型从两个分布的 Sinkhorn 耦合中进行采样。他们的方法用神经网络对双变量进行建模，然后解决最优转移问题。 训练神经网络后，可以通过 Langevin dynamics 和基于分数的模型进行采样。 他们使用 U-Net 架构进行图像超分辨率实验。

3.6图像编辑

Meng 等人在各种引导图像生成任务中利用扩散模型，即笔划绘画或基于笔划的编辑和图像合成。 从包含某种形式的引导的图像开始，其属性（例如形状和颜色）被保留，同时通过逐渐添加噪声（扩散模型的前向过程）来平滑变形。 然后，对结果进行去噪（逆过程），以根据指导创建逼真的图像。 通过求解反向 SDE，使用通用扩散模型合成图像，无需任何自定义数据集或修改训练。

文章中介绍了基于自然语言描述编辑图像特定区域的第一种方法。 要修改的区域由用户通过掩码指定。 该方法依靠 CLIP 指导根据文本输入生成图像，但作者观察到，最后将输出与原始图像相结合并不能产生全局连贯的图像。 因此，他们修改了去噪过程来解决这个问题。 更准确地说，在每一步之后，作者都会在潜在图像上应用掩模，同时还添加原始图像的噪声版本。

扩展了中提出的工作，Avrahami 等人应用潜在扩散模型使用文本在本地编辑图像。VAE 将图像和自适应时间掩模（要编辑的区域）编码到发生扩散过程的潜在空间中。 每个样本都经过迭代去噪，同时受到感兴趣区域内文本的指导。 然而，受混合扩散的启发，图像与当前时间步长的潜在空间中的屏蔽区域相结合。 最后，使用 VAE 对样本进行解码以生成新图像。该方法表现出卓越的性能，同时速度相对更快。

3.7 图像修复

Nichol 等人训练一个以文本描述为条件的扩散模型，并研究无分类器和基于 CLIP 指导的有效性。 他们通过第一种选择获得了更好的结果。 此外，他们还微调图像修复模型，解锁基于文本输入的图像修改。

Lugmay 等人提出了一种与掩模形式无关的修复方法。 他们为此使用了无条件扩散模型，但修改了其逆过程。 他们通过从掩模图像中采样已知区域来生成步骤 t - 1 的图像，并通过对步骤 t 获得的图像应用去噪来生成未知区域。 通过这个过程，作者观察到未知区域具有正确的结构，但在语义上也是不正确的。 此外，他们通过多次重复所提出的步骤来解决该问题，并且在每次迭代时，他们用从步骤 t - 1 生成的去噪版本获得的新样本替换步骤 t 中的先前图像。

3.8 图像分割

Baranchuk 等人演示了如何在语义分割中使用扩散模型。 从 U-Net 的解码器（用于去噪过程）获取不同尺度的特征图（中间块）并将它们连接起来（对特征图进行上采样以获得相同的维度），它们可用于对每个特征图进行分类 通过进一步附加多层感知器的集合来实现像素。 作者表明，这些在去噪过程的后续步骤中提取的特征图包含丰富的表示。实验表明，基于扩散模型的分割优于大多数基线。

Amit 等人提出通过扩展 U-Net 编码器的架构，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通过两个不同的编码器并通过求和组合在一起。 然后将结果提供给 U-Net 的编码器-解码器。 由于每个时间步长注入随机噪声，因此会生成单个输入图像的多个样本并用于计算平均分割图。 U-Net 架构基于之前的工作，而输入图像生成器是使用 Residual Dense Blocks构建的。 去噪样本生成器是一个简单的 2D 卷积层。

3.9 多任务方法

一系列扩散模型已应用于多个任务，表现出良好的跨任务泛化能力。 我们将在下面讨论此类贡献。

Song 等人提出了噪声条件评分网络（NCSN），这是一种估计不同噪声尺度下的评分函数的方法。 为了进行采样，他们引入了 Langevin dynamics 的退火版本，并用它来报告图像生成和修复的结果。 NCSN 架构主要基于提出的工作，进行了一些小的更改，例如用实例规范化替换批量规范化。

Kadkhodaie 等人训练神经网络来恢复被高斯噪声损坏的图像，这些图像是使用限制在特定范围内的随机标准差生成的。 训练后，神经网络的输出与作为输入接收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。 该属性基于中先前完成的工作。 对于图像生成，作者使用上述差异作为梯度（分数）估计，并通过采用类似的退火Langevin dynamics 的迭代方法，从网络的隐式数据先验中进行采样。 然而，这两种采样方法有一

些不同之处，例如迭代更新中注入的噪声遵循不同的策略。 注入的噪声根据网络的估计进行调整，它是固定的。 此外，梯度估计是通过分数匹配学习的，而 Kadkhodaie 等人依靠前面提到的属性来计算梯度。 Kadkhodaie 等人的贡献通过使算法适应线性逆问题（例如去模糊和超分辨率）而进一步发展。

扩散模型的 SDE 公式概括了之前的几种方法。Song 等人提出了正向和反向扩散过程作为SDE的解决方案。 该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。 作者进行了图像生成、修复和着色方面的实验。

Batzolis 等人在扩散模型中引入了一种新的前向过程，称为非均匀扩散。 这是由使用不同的 SDE 扩散的每个像素决定的。 这个过程中使用了多个网络，每个网络对应不同的扩散尺度。 该论文进一步演示了一种新颖的条件采样器，可以在两种基于去噪分数的采样方法之间进行插值。该模型在无条件合成、超分辨率、修复和边缘到图像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 用于对反向步骤进行自回归建模，其中编码器的表示是使用上一步的输出获得的。 ImageBART 在无条件、类条件和文本条件图像生成以及本地编辑方面进行评估。

Gao 等人引入扩散恢复可能性，这是一种基于能量的模型的新训练程序。 他们学习一系列基于能量的扩散过程边缘分布模型。 因此，他们不是用正态分布来近似逆过程，而是从基于边际能量的模型中得出条件分布。 作者对图像生成和修复进行了实验。

Batzolis 等人分析了先前基于分数的条件图像生成扩散模型。 此外，他们提出了一种称为条件多速扩散估计器（conditional multi-speed diffusive estimator，CMDE）的条件图像生成新方法。 该方法基于以下观察：以相同速率扩散目标图像和条件图像可能不是最佳的。 因此，他们建议使用 SDE 来扩散具有相同漂移和不同扩散速率的两个图像。 该方法在修复、超分辨率和边缘到图像合成方面进行了评估。

Liu 等人引入了一个框架，允许根据参考图像进行文本、内容和风格指导。 核心思想是使用最大化图像和文本所学习的表示之间的相似性的方向。 图像和文本嵌入由 CLIP 模型生成。 为了满足在噪声图像上训练 CLIP 的需求，作者提出了一种不需要文本标题的自我监督程序。 该过程使用成对的正常图像和噪声图像来最大化正对之间的相似性并最小化负对之间的相似性（对比目标）。

Choi 等人提出了一种不需要进一步训练的新方法，使用无条件扩散模型进行条件图像合成。 给定参考图像（即条件），通过消除低频内容并将其替换为参考图像中的内容，每个样本都更接近它。 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。作者展示了如何将该方法应用于各种图像到图像的翻译任务，例如，绘画到图像，并使用涂鸦进行编辑。

Hu 等人建议将扩散模型应用于离散 VAE 给出的离散表示。 他们考虑了 CelebA-HQ 和 LSUN Church 数据集，评估了图像生成和修复实验中的想法。

Rombach 等人引入潜在扩散模型，其中正向和反向过程发生在自动编码器学习的潜在空间上。 它们还在架构中包含了交叉注意力，这进一步改进了条件图像合成。 该方法在超分辨率、图像生成和修复方面进行了测试。

Preechakul 等人介绍的方法包含一个学习描述性潜在空间的语义编码器。 该编码器的输出用于调节 DDIM 实例。 所提出的方法允许 DDPM 在插值或属性操作等任务上表现良好。

Chung 等人引入了一种采样算法，该算法减少了条件情况所需的步骤数。 与标准情况相比，逆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中间噪声图像，并从此时开始恢复采样。 该方法在修复、超分辨率和磁共振成像（MRI）重建方面进行了测试。

作者微调了预训练的 DDIM 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，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。评估中考虑的任务是未见过的域之间的图像转换和多属性迁移。

从 Meng 等人将扩散模型表述为 SDE 开始，Khrulkov 等人研究潜在空间和生成的编码器图。 根据 Monge 公式，结果表明这些编码器图是最佳传输图，但这仅针对多元正态分布进行了证明。 作者使用 Dhariwal 等人的模型实现，通过数值实验和实际实验进一步支持了这一点。

Shi 等人首先观察如何将基于无条件分数的扩散模型表示为薛定谔桥，该模型可以使用迭代比例拟合（Iterative Proportional Fitting）的修改版本来求解。 先前的方法被重新表述为接受条件，从而使条件合成成为可能。 对迭代算法进行进一步调整，以优化收敛所需的时间。 该方法首先使用 Kovachki 等人的合成数据进行了验证，显示出估计真实情况的能力有所提高。 作者还进行了超分辨率、修复和生化需氧量的实验，后一项任务受到 Marzouk 等人的启发。

受 Retrieval Transformer的启发，Blattmann 等人提出了一种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。 图像由具有固定参数的编码器进一步编码，并投影到 CLIP特征空间中。 最后，扩散模型的逆过程以该潜在空间为条件。该方法可以进一步扩展以使用其他条件信号，例如，文本，通过简单地使用信号的编码表示来增强潜在空间。

Lyu 等人引入了一种新技术来减少扩散模型的采样步骤数，同时提高了性能。 这个想法是在更早的步骤中停止扩散过程。 由于采样不能从随机高斯噪声开始，因此使用 GAN 或 VAE 模型将最后的扩散图像编码到高斯潜在空间中。 然后将结果解码为图像，

该图像可以扩散到后向过程的起点。

Graikos 等人的目标是将扩散模型分成两个独立的部分，先验（基础部分）和约束（条件）。 这使得模型无需进一步训练即可应用于各种任务。 考虑到约束变得可鉴别（differentiable），从更改 DDPM 方程可以独立训练模型并在条件设置中使用它。 作者进行了条件图像合成和图像分割的实验。

3.10 医学图像生成和转换

Wolleb 等人介绍了一种基于扩散模型的脑肿瘤分割背景下的图像分割方法。 训练包括扩散分割图，然后对其进行去噪以获得原始图像。 在后向过程中，大脑 MR 图像被连接到中间去噪步骤中，以便通过 U-Net 模型，从而调节其去噪过程。 此外，对于每个输入，作者建议生成多个样本，由于随机性，这些样本会有所不同。 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 等人引入了一种基于评分的模型方法，该方法能够解决医学成像中的逆问题，即根据测量重建图像。 首先，训练无条件评分模型。 然后，导出测量的随机过程，该过程可用于通过近端优化步骤将条件信息注入模型中。 最后，将信号映射到测量的矩阵被分解，以允许以封闭形式进行采样。 作者对不同的医学图像类型进行了多项实验，包括计算机断层扫描（CT）、低剂量 CT 和 MRI。

Chung 等人在医学成像领域内，但专注于通过加速 MRI 扫描重建图像，建议使用基于分数的扩散模型来解决逆问题。 评分模型仅在无条件设置下的幅度图像上进行预训练。然后，在采样过程中采用方差爆炸 SDE 求解器。通过采用与数据一致性映射交织的预测校正算法，分割图像（实部和虚部）被馈送，从而能够根据测量来调节模型。 此外，作者还提出了该方法的扩展，可以对多个线圈变化的测量进行调节。

O ¨ zbey 等人提出了一种具有对抗性推理的扩散模型。 为了增加每个扩散步骤，从而减少步骤，作者在反向过程中采用了 GAN 模型来估计每个步骤的去噪图像。 他们使用与类似的方法，引入了循环一致的架构，以允许对不成对的数据集进行训练。

Hu 等人的目标的目的是消除光学相干断层扫描（OCT）b 扫描中的散斑噪声。 第一阶段由一种称为自融合的方法表示，其中选择接近输入 OCT 体积的给定 2D 切片的附加 b 扫描。 第二阶段由扩散模型组成，其起点是原始 b 扫描及其邻居的加权平均值。 因此，可以通过对干净的扫描进行采样来消除噪声。

3.11 医学图像中的异常检测

自动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定类型的 VAE，因此采用扩散模型来完成与 VAE 相同的任务似乎很自然。 到目前为止，扩散模型在检测医学图像异常方面显示出了有希望的结果，如下文进一步讨论的。

Wyatt 等人在健康医学图像上训练 DDPM。 通过从原始图像中减去生成的图像，在

推理时检测异常。 这项工作还证明，使用单纯形噪声代替高斯噪声可以为此类任务带来更好的结果。

Wolleb 等人提出了一种基于扩散模型的弱监督方法，用于医学图像中的异常检测。给定两张未配对的图像，一张是健康的，一张是有病变的，前者会被模型扩散。 然后，由二元分类器的梯度引导去噪过程，以生成健康图像。 最后，将采样的健康图像与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大脑扫描中的异常，并对这些区域进行分割。 图像由 VQ-VAE进行编码，并从码本中获得量化的潜在表示。 扩散模型在这个潜在空间中运行。 对来自后向过程的中值步骤的中间样本进行平均，然后应用预先计算的阈值图，创建暗示异常位置的二进制掩码。 从中间开始后向过程，使用二值掩模对异常区域进行去噪，同时保持其余区域。 最后，对最后一步的样本进行解码，产生健康的图像。通过输入图像和合成图像相减得到异常的分割图。

Sanchez 等人遵循相同的原理来检测和分割医学图像中的异常：扩散模型生成健康样本，然后从原始图像中减去这些样本。 使用模型对输入图像进行扩散，反转去噪方程并使条件无效，然后应用后向条件过程。 利用无分类器模型，通过 U-Net 中集成的注意力机制实现引导。训练使用了健康和不健康的例子。

3.12 视频生成

最近在提高扩散模型效率方面取得的进展使其能够在视频领域得到应用。 接下来我们将展示将扩散模型应用于视频生成的作品。

Ho 等人将扩散模型引入视频生成任务。 与 2D 情况相比，更改仅应用于架构。 作者采用中的3D U-Net，展示了无条件和文本条件视频生成的结果。 较长的视频以自回归方式生成，其中后面的视频块以前面的视频块为条件。

Yang 等人使用扩散模型逐帧生成视频。 相反的过程完全以卷积循环神经网络提供的上下文向量为条件。 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的情况返回更好的结果。 结论是前一种选择效果更好。

H ¨ oppe 等人提出了随机掩模视频扩散（RaMViD），一种可用于视频生成和填充的方法。 他们工作的主要贡献是一种新颖的训练策略，该策略将帧随机分为屏蔽帧和未屏蔽帧。 未屏蔽的帧用于调节扩散，而屏蔽的帧则通过前向过程进行扩散。

Harvey 等人的工作引入了灵活的扩散模型，这是一种可以与多种采样方案一起使用来生成长视频的扩散模型。作者通过随机选择扩散中使用的帧和用于调节过程的帧来训练扩散模型。 训练模型后，他们研究了多种采样方案的有效性，得出的结论是采样选择取决于数据集。

3.13 其他任务

有一些开创性的工作将扩散模型应用于新任务，而这些任务很少通过扩散模型进行探索。 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、自动编码和无监督表示学习。 他们从以潜在形状为条件的点云可能性的变分下界导出目标函数。 实验是使用 PointNet作为底层架构进行的。

Zhou 等人介绍了点体素扩散（PVD），这是一种在点体素表示上应用扩散的形状生成的新方法。 该方法解决了 ShapeNet 和 PartNet 数据集上的形状生成和完成（completion）。

Zimmermann 等人展示了一种应用基于分数的模型进行分类的策略。 他们将图像标签作为条件变量添加到评分函数中，并且借助 ODE 公式，可以在推理时计算条件似然。因此，预测是具有最大似然性的标签。 此外，考虑到常见的图像损坏和对抗性扰动，他们研究了这种类型的分类器对分布外场景的影响。

Kim 等人建议使用扩散模型来解决图像配准任务。 这是通过两个网络实现的，一个是扩散网络，另一个是基于 U-Net 的变形网络。 给定两张图像（一张静态，一张移动），前一个网络的作用是评估两个图像之间的变形，并将结果输入到后一个网络，后者预测变形场（deformation fields），从而生成样本。 该方法还能够合成整个过渡过程中的变形。 作者针对不同的任务进行了实验，一项针对 2D 面部表情，一项针对 3D 大脑图像。 结果证实该模型能够产生定性且准确的配准字段。

Jeanneret 等人应用扩散模型进行反事实解释。 该方法从带噪声的查询图像开始，并生成具有无条件 DDPM 的样本。 利用生成的样本，计算引导所需的梯度。 然后，应用反向引导过程的一个步骤。 输出进一步用于接下来的反向步骤。

Sanchez 等人改编 Dhariwal 等人的工作用于反事实图像生成。 去噪过程由分类器梯度引导，以从所需的反事实类中生成样本。 关键贡献是用于检索原始图像的潜在表示的算法，从降噪过程开始。 他们的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 等人演示了如何使用扩散模型作为对抗性攻击的防御机制。 给定一个对抗性图像，它会被扩散，直到达到最佳计算的时间步长。 然后模型将结果反转，最终产生纯化的样品。 为了优化求解反时 SDE 的计算，Li 等人的伴随（adjoint）灵敏度方法被用于梯度分数计算。

在少样本学习的背景下，Giannone 等人提出了一种基于扩散模型的图像生成器。 给定一小组调节合成的图像，视觉 transformer 对这些图像进行编码，并将生成的上下文表示（通过两种不同的技术）集成到去噪过程中使用的 U-Net 模型中。

Wang 等人提出了一种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，将输入噪声提供给编码器，同时使用多层空间自适应归一化算子将语义标签

图传递给解码器。为了进一步提高采样质量和语义标签图上的条件，还向采样方法提供空图以生成无条件噪声。 然后，最终噪声使用两个估计值。

关于恢复受各种天气条件（例如雪、雨）负面影响的图像的任务，Ozdenizci 等人演示了如何使用扩散模型。 他们通过在每个时间步将降级图像按通道连接到去噪样本来调节降级图像的去噪过程。 为了处理不同的图像尺寸，在每一步中，样本都被分成重叠的块，并行地通过模型，并通过对重叠像素进行平均来组合回来。 所采用的扩散模型基于 U-Net 架构，但经过修改以接受两个串联图像作为输入。

Kawar 等人将图像恢复任务表述为线性逆问题，建议使用扩散模型。 受到卡瓦尔等人的启发 ，使用奇异值分解对线性退化矩阵进行分解，使得输入和输出都可以映射到进行扩散过程的矩阵的谱空间上。 对各种任务进行评估：超分辨率、去模糊、着色和修复。

3.14 理论贡献

Huang等人演示了Song等人提出的方法如何与最大化反向 SDE 边际似然下限相关。此外，他们通过CIFAR-10和MNIST上的图像生成实验验证了他们的理论贡献。

4. 结束语和未来方向

在本文中，我们回顾了研究界在开发扩散模型并将其应用于各种计算机视觉任务方面取得的进展。 我们基于 DDPM、NCSN 和 SDE 确定了扩散模型的三种主要公式。 每个公式在图像生成方面都取得了显着的效果，超越了GAN，同时增加了生成样本的多样性。扩散模型在研究仍处于早期阶段时就取得了突出的成果。 尽管我们观察到主要焦点是条件和无条件图像生成，但仍有许多任务需要探索并需要实现进一步的改进。

局限性

扩散模型最显着的缺点仍然是需要在推理时执行多个步骤才能生成一个样本。 尽管在这个方向上进行了大量研究，但 GAN 生成图像的速度仍然更快。

扩散模型的其他问题可以与采用 CLIP 嵌入进行文本到图像生成的常用策略相关联。例如，Ramesh 等人强调他们的模型很难在图像中生成可读文本，并通过声明 CLIP 嵌入不包含有关拼写的信息来激发该行为。 因此，当这种嵌入用于调节去噪过程时，模型会继承此类问题。

未来发展方向

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取大步长。 事实上，采取小步骤可以确保每一步生成的数据样本都能由学习到的高斯分布来解释。 当应用梯度下降来优化神经网络时，会观察到类似的行为。 事实上，在梯度的负方向上迈出一大步，即使用非常大的学习率，可能会导致模型更新到具有高不确定性的区域，无法控制损失值。 在未来的工作中，将从高效优化器借用的更新规则转移到扩散模型可能会导致更有效的采样

（生成）过程。

除了当前研究更有效的扩散模型的趋势之外，未来的工作还可以研究应用于其他计算机视觉任务的扩散模型，例如图像去雾、视频异常检测或视觉问答。 即使我们发现了一些研究医学图像中异常检测的作品，这项任务也可以在其他领域进行探索，例如视频监控或工业检查。

一个有趣的研究方向是评估判别任务中扩散模型学习到的表示空间的质量和效用。 这可以至少以两种不同的方式进行。

以直接的方式，通过在去噪模型提供的潜在表示之上学习一些判别模型，来解决某些分类或回归任务。

以间接的方式，通过使用扩散模型生成的真实样本来增强训练集。

后一个方向可能更适合对象检测等任务，其中修复扩散模型可以很好地混合图像中的新对象。

未来的另一个工作方向是采用条件扩散模型来模拟视频中可能的未来。 生成的视频可以进一步作为强化学习模型的输入。

与之前的最先进技术相比，最近的扩散模型显示出令人印象深刻的文本到视频合成能力，显着减少了伪影的数量并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的关注，因为生成的视频相当短。 因此，对对象之间的长期时间关系和交互进行建模仍然是一个开放的挑战。

未来，扩散模型的研究还可以扩展到学习同时解决多个任务的多用途模型。 创建扩散模型以生成多种类型的输出，同时以各种类型的数据为条件，例如，文本、类别标签或图像，可能会让我们更进一步了解开发通用人工智能（AGI）的必要步骤。