18 生成扩散模型漫谈（十三）：从万有引力到扩散模型

Oct By 苏剑林 | 2022-10-18 | 53245位读者 引用

对于很多读者来说，生成扩散模型可能是他们遇到的第一个能够将如此多的数学工具用到深度学习上的模型。在这个系列文章中，我们已经展示了扩散模型与数学分析、概率统计、常微分方程、随机微分方程乃至偏微分方程等内容的深刻联系，可以说，即便是做数学物理方程的纯理论研究的同学，大概率也可以在扩散模型中找到自己的用武之地。

在这篇文章中，我们再介绍一个同样与数学物理有深刻联系的扩散模型——由“万有引力定律”启发的ODE式扩散模型，出自论文《Poisson Flow Generative Models》（简称PFGM），它给出了一个构建ODE式扩散模型的全新视角。

万有引力

中学时期我们就学过万有引力定律，大概的描述方式是：

> 两个质点彼此之间相互吸引的作用力，是与它们的质量乘积成正比，并与它们之间的距离成平方反比。

这里我们忽略质量和常数，主要关心它的方向和与距离的关系，假设引力源位于$\boldsymbol{y}$，那么位于$\boldsymbol{x}$的物体所受到的引力可以记为

$$\boldsymbol{F}(\boldsymbol{x}) = -\frac{1}{4\pi}\frac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}-\boldsymbol{y}\|^3} \tag{1}$$

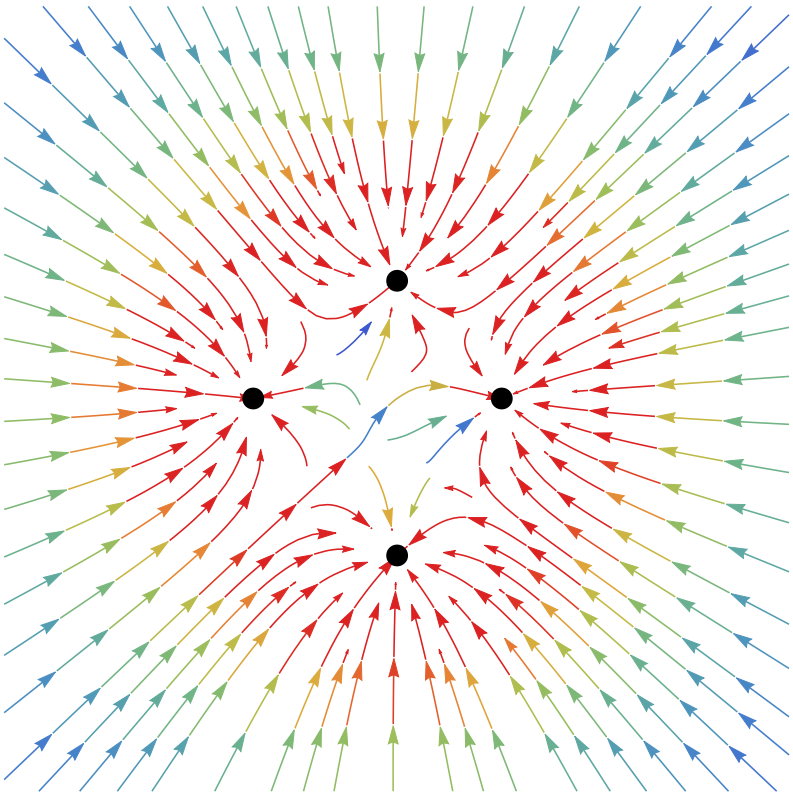
$\frac{1}{4\pi}$这个因子我们可以先不管它，它不影响后面的分析。准确来说，上式描述的是三维空间的引力场，对于d维空间来说，其引力场的形式为

$$\boldsymbol{F}(\boldsymbol{x}) = -\frac{1}{S_d(1)}\frac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}-\boldsymbol{y}\|^d} \tag{2}$$

其中$S_d(1)$是d维单位超球面的表面积。该式实际上就是d维Poisson方程的格林函数的梯度，这也就是论文标题中的“Poisson”一词的来源。

沿场线走

如果引力源有多个，那么直接将各个引力源的引力相加即可，这是引力场的线性可加性。下面我们画出了四个引力源的向量场，其中引力源用黑色点标记出，彩色线表示场线：



引力场示意图

从上述引力场图我们可以看出它的一个重要特点：

> 除了极少数外，大部分场线都是从远处出发，终止于某个引力源点。

此时，一个直观而又“异想天开”的主意是：

> 如果每个引力源都代表着一个要生成的真实样本点，那么远处的任意点只要沿着场线运动，不就都可以演变成一个真实样本点了吗？

这就是《Poisson Flow Generative Models》一文最核心的天才想法！

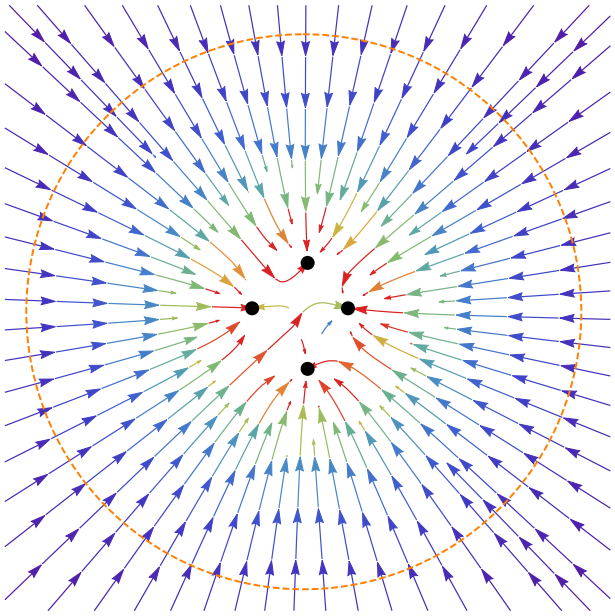
等效质心

当然，天才归天才，要将它真正变成一个可用的模型，还有很多细节要补充。比如我们刚才说“远处的任意点”，这就是扩散模型的初始分布了，那么问题就来了：“远处”是多远？“任意点”该如何采样？如果采样方式过于复杂，那也没有价值了。

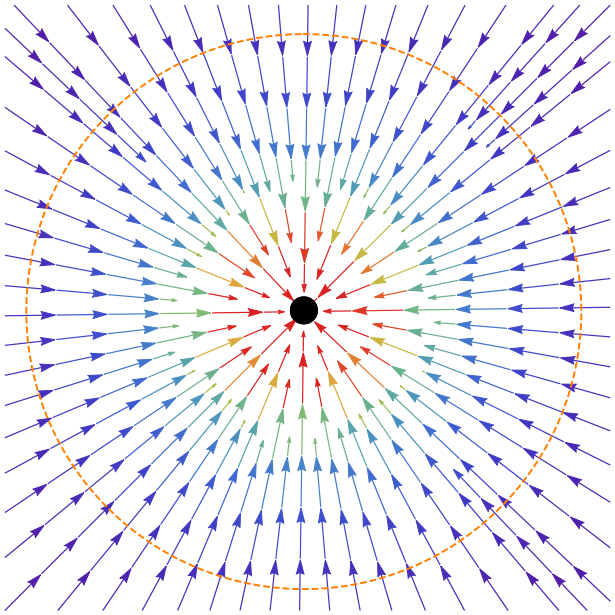
幸运的是，引力场有一个非常重要的等效性质：

> 无穷远处的多源引力场，等价于位于质心、质量叠加的质点引力场。

也就是说，当距离足够远时，我们只需要当它是位于质心的单源质点引力场。下图也画出了多源引力场及其对应的质心引力场，可以看到，当距离变大时（橙色圆圈位置），两者的引力场几乎一致了。



多源引力场



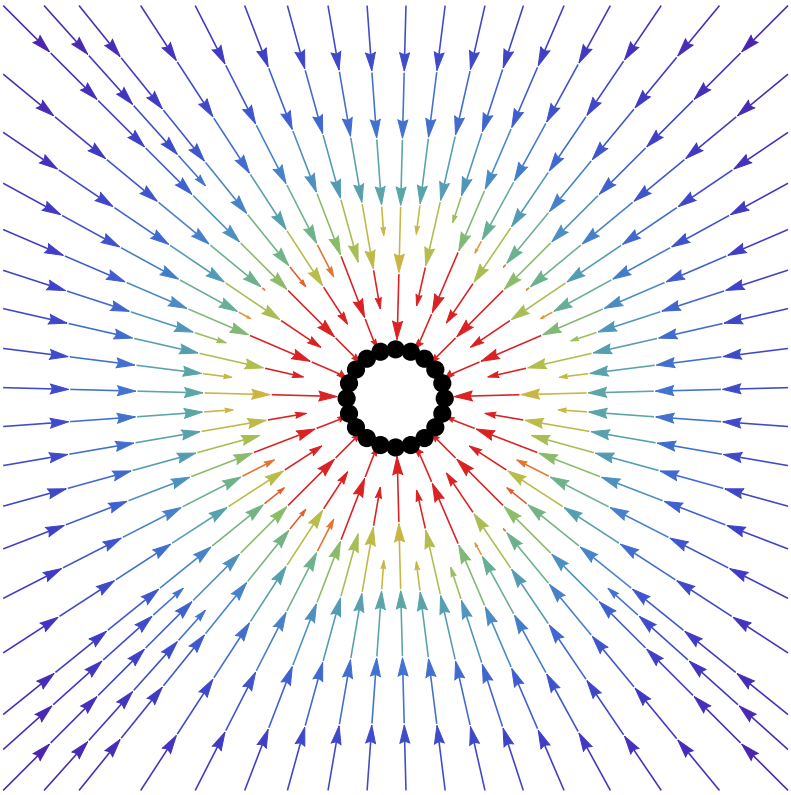
质心引力场

单个质点的引力场有什么特点？各向同性！这意味着在足够大半径时，可以认为场线是均匀地穿过以质点为球心的球面的，所以我们在一个半径足够大的球面上进行均匀采样就可以了，这就解决了初始分布的采样问题。至于“足够大”是多大，我们后面再说。

模式坍缩

所以生成模型就这样构建好了？还没有。引力场的各向同性使得对应的初始分布易于采样，然而也会造成引力源的相互抵消现象，从而出现“模式坍缩（Mode Collapse）”。

具体来说，我们先来画出均匀分布在球壳上的引力场，它的分布是这样的：



引力源各向同性的引力场

发现特点了没？在球壳外部是正常的各向同性分布，但是在球壳内部是“空”的！也就是说球壳内部的引力场相互抵消了，相当于一个真空地带，这个现象笔者在十多年前的科普博客《球壳内部的均匀力场》也介绍过。

抵消现象意味着任意选一个球面，球面上均匀分布的引力源由于引力相互抵消，那么就相当于不存在该引力源了。而我们说了，本文构建生成模型的方式，是通过远处的任意点沿着场线运动，直达某个引力源。如果引力源相互抵消，那么就意味着永远也

达不到某些引力源，即意味着某些真实样本无法生成，生成结果就会缺失多样性，这就是“模式坍缩”现象。

增加一维

看上去模式坍缩无论如何都是不能避免的。因为在构建生成模型的时候，我们通常假设真实样本服从一个连续型的分布，这样一来，任选一个球面，哪怕真实样本在该球面的分布不是均匀的，我们也能从中挑出一个均匀分布的“子集”，该“子集”的引力就相互抵消了，相当于这些数据点不存在了，继而发生模式坍缩。

那么这条路真的走到尽头了吗？并没有！这时候PFGM的第二个“天才想法”来了：**增加一维！**

刚才我们分析了，模式坍缩无法避免，是因为连续性分布的假设导致各向同性无法避免。要想避免模式坍缩，就要想办法杜绝分布的各向同性。可是真实样本的分布是目标分布，这是不能改变的，然而我们可以给它增加一维，如果我们在$d+1$维空间去讨论，那么原来的d维分布可以视为$d+1$维空间的一个平面，平面就不可能各向同性了。举个低维空间的例子。我们知道对于二维空间来说，“圆”是各向同性的，但对于三维空间来说，“球”才是各向同性的，二维空间中各向同性的“圆”，在三维空间看来就不是各向同性了。

所以，假设要生成的真实样本原来是$\boldsymbol{x}\in\mathbb{R}^d$的，我们引入一个新维度$t$，使得数据点变为$(\boldsymbol{x},t)\in\mathbb{R}^{d+1}$，而原来真实样本服从的分布是$\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x})$的，现在改为$(\boldsymbol{x},t)\sim\delta(t)\tilde{p}(\boldsymbol{x})$，其中$\delta(t)$是狄拉克分布，其实就是将真实样本放到$d+1$维空间的$t=0$平面上。这样处理后，在$d+1$维空间中，真实样本点的$t$取值总是0，因此就不能出现各向同性现象了（类比刚才“三维空间中的圆”的例子）。

豁然开朗

乍一看，增加一维只是一个数学上的小技巧，但细细品味之下，我们会越发感觉它妙不可言，很多在原来d维空间中不好处理的细节问题，在$d+1$维空间中就豁然开朗了。

根据式(2)和引力的线性叠加性，我们可以写出此时$d+1$维空间的引力场为

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{x},t) =& -\frac{1}{S_{d+1}(1)}\iint \frac{(\boldsymbol{x}-\boldsymbol{x}_0,t-t_0)}{(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2+(t-t_0)^2)^{(d+1)/2}}\delta(t_0)\tilde{p}(\boldsymbol{x}_0)d\boldsymbol{x}_0 dt_0\\ =& -\frac{1}{S_{d+1}(1)}\int \frac{(\boldsymbol{x}-\boldsymbol{x}_0,t)}{(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2+t^2)^{(d+1)/2}}\tilde{p}(\boldsymbol{x}_0)d\boldsymbol{x}_0\\ \triangleq& (\boldsymbol{F}_{\boldsymbol{x}},\boldsymbol{F}_t)\end{aligned} \tag{3}$$

其中$\boldsymbol{F}_{\boldsymbol{x}}$是$\boldsymbol{F}(\boldsymbol{x},t)$的前$d$个分量，$\boldsymbol{F}_t$是它的第$d+1$个分量。下一节我们再来讨论$\boldsymbol{F}(\boldsymbol{x},t)$怎么学，现在假设$\boldsymbol{F}(\boldsymbol{x},t)$已经知道了，那么接下来就是要沿着场线运动，也就是运动轨迹要时刻跟$\boldsymbol{F}(\boldsymbol{x},t)$同方向，即

$$(d\boldsymbol{x},dt) = (\boldsymbol{F}_{\boldsymbol{x}},\boldsymbol{F}_t)d\tau \quad\Rightarrow\quad \frac{d\boldsymbol{x}}{dt} = \frac{\boldsymbol{F}_{\boldsymbol{x}}}{\boldsymbol{F}_t} \tag{4}$$

这就是生成过程所需要的微分方程（ODE）。在之前的d维方案中，除了有模式坍缩问题外，什么时候终止也是一个不好处理的细节问题。直观来想，就是沿着场线运动，直到撞上一个真实样本后就停止，但什么时候才是“撞上”，这个并不好判断。而在$d+1$维方案中，我们知道真实样本都是在$t=0$这个面上，所以可以很自然地以$t=0$为终止信号了。

至于初始分布，按照前面的讨论，应该是“半径足够大的、$d+1$维的球面均匀分布”，但既然我们是以$t=0$为终止信号，那么我们不妨固定一个足够大的$t=T$（大致是$40\sim 100$这个量级），然后在$t=T$这个平面上做采样，这样一来生成过程就变成了微分方程(4)从$t=T$到$t=0$的运动过程了，生成过程的始和终都变得相当明朗。

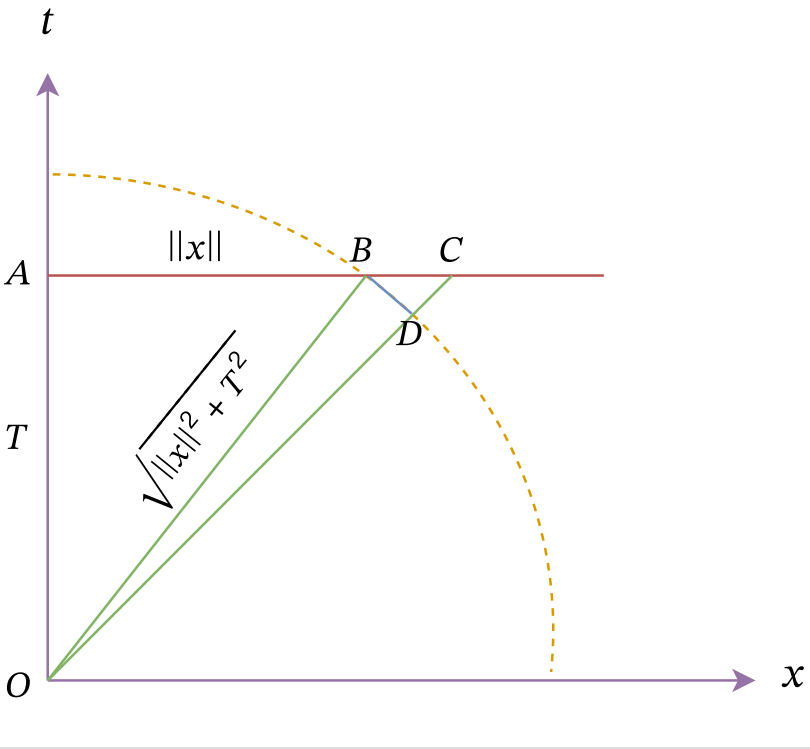
当然，如果在固定的$t=T$平面上采样，那肯定就不是均匀的了。事实上有：

$$p_{prior}(\boldsymbol{x}) \propto \frac{1}{(\|\boldsymbol{x}\|^2+T^2)^{(d+1)/2}} \tag{5}$$

推导过程见下面的框。可以看到，概率密度只依赖于模长$\|\boldsymbol{x}\|$，所以从该分布采样的方案是先按照特定的分布采样模长，然后再按均匀分布采样方向，将两者进行组合。对于模长的采样，记$r=\|\boldsymbol{x}\|$，我们将它换元到超球坐标，就可以得到

$p_{prior}(r) \propto r^{d-1}(r^2+T^2)^{-(d+1)/2}$，然后利用逆累积概率函数法进行采样就行了（参考《变分自编码器（七）：球面上的VAE（vMF-VAE）》）。

初始分布的推导：场线是均匀穿过$d+1$维超球面上的，所以$(\boldsymbol{x}, T)$处的密度是反比于$S_{d+1}(\boldsymbol{x}, T)$，即$\propto \frac{1}{(\|\boldsymbol{x}\|^2+T^2)^{d/2}}$，但现在不是在球面，而是在$t=T$的平面上，所以我们要将球面投影到平面上，示意图如下：



球面到t=T平面的投影

如上图，当B、D两点充分接近时，有$\Delta OAB \sim \Delta BDC$，所以

$$\frac{|BC|}{|BD|} = \frac{|OB|}{|OA|} = \frac{\sqrt{\|\boldsymbol{x}\|^2+T^2}}{T} \tag{6}$$

也就是说，原本球面上单位长度的弧，投影到平面上后长度变为了$\frac{\sqrt{\|\boldsymbol{x}\|^2+T^2}}{T}$倍，由于只有一个维度变化，所以原来球面上的面积元，投影后也变为$\frac{\sqrt{\|\boldsymbol{x}\|^2+T^2}}{T}$倍，因此根据

概率反比面积，我们可以得到

$$p_{prior}(\boldsymbol{x}) \propto \frac{1}{S_{d+1}(\boldsymbol{x},T)} \times \frac{T}{\sqrt{\|\boldsymbol{x}\|^2 + T^2}} \propto \frac{1}{(\|\boldsymbol{x}\|^2 + T^2)^{(d+1)/2}} \tag{7}$$

场的训练

现在，初始分布有了，微分方程也有了，所以就只差向量场函数$\boldsymbol{F}(\boldsymbol{x},t)$的训练了。从微分方程$(4)$可以看出，它只依赖于向量场的相对值，因此向量场的缩放不影响最终结果。根据式(3)，向量场可以写为

$$\boldsymbol{F}(\boldsymbol{x},t) = \mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0)}\left[-\frac{(\boldsymbol{x}-\boldsymbol{x}_0,t)}{(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + t^2)^{(d+1)/2}}\right] \tag{8}$$

根据我们在《生成扩散模型漫谈（五）：一般框架之SDE篇》、《生成扩散模型漫谈（七）：最优扩散方差估计（上）》等文章多次用到一个结论：

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathop{\text{argmin}}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}}\left[\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right] \tag{9}$$

我们可以引入函数$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$来学习$\boldsymbol{F}(\boldsymbol{x},t)$，训练目标为

$$\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0)}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) + \frac{(\boldsymbol{x}-\boldsymbol{x}_0,t)}{(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + t^2)^{(d+1)/2}}\right\|^2\right] \tag{10}$$

然而，上述目标的$\boldsymbol{x},t$还需要采样，它的采样方式没有明确定义。这就是PFGM的主要特点之一，它直接定义了反向过程（生成过程），不需要定义前向过程，而这一步的采样实际上就相当于前向过程。为此，原论文考虑地每个真实样本进行扰动的方式来构建$\boldsymbol{x},t$的样本：

$$\boldsymbol{x} = \boldsymbol{x}_0 + \|\boldsymbol{\varepsilon}_{\boldsymbol{x}}\|(1+\tau)^m \boldsymbol{u},\quad t = |\varepsilon_t|(1+\tau)^m \tag{11}$$

其中$(\boldsymbol{\varepsilon}_{\boldsymbol{x}},\varepsilon_t)\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}_{(d+1)\times(d+1)})$，$m\sim U[0,M]$，$\boldsymbol{u}$是$d$维单位球面上均匀分布的

单位向量，而τ, σ, M则都是常数。~~这个设计有颇多的主观性，大家自行欣赏和领会即可，这里不做过多展开。~~后续讨论请看这里。

最后，原论文的训练目标跟本文的式(10)略有不同，大致相当于

$$\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) + \text{Normalize}\left(\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0)}\left[\frac{(\boldsymbol{x}-\boldsymbol{x}_0, t)}{(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + t^2)^{(d+1)/2}}\right]\right)\right\|^2 \tag{12}$$

实际训练时由于只能采样有限个$\boldsymbol{x}_0$对括号里边的期望进行估算，因此该目标实际上是一个有偏估计。当然有偏不一定就比无偏更差，具体为什么原论文使用有偏估计，这里暂时不得而知，猜测是因为有偏估计由于对向量进行了归一化操作，可能会使得训练进程更加稳定，但是也因为有偏的估计进行了归一化，所以需要较大的batch_size才能比较准，这对实验成本提了要求。

实验结果

可以说，PFGM是一个彻底的新框架，它不再像之前一样依赖于高斯假设，并且得到了一个确实有着全新内涵的模型。然而，我们不能“为新而新”，如果新的框架没有做出更有说服力的结果，那么新就是没有意义的。

当然，原论文的实验结果，肯定了PFGM的价值，比如得到了更好的评估指标、更快的生成速度，以及对超参数（包括模型架构）有更好的鲁棒性等，这里就不一一展示了，大家自行读原论文就好。我看了看原论文中了NeurIPS 2022，不得不说这确实是实至名归的顶会论文啊！

官方Github：https://github.com/Newbeeer/Poisson_flow

Table 1: CIFAR-10 sample quality (FID, Inception) and number of function evaluation (NFE).

	Invertible?	Inception ↑	FID ↓	NFE ↓
PixelCNN [36]	✗	4.60	65.9	1024
IGEBM [8]	✗	6.02	40.6	60
ViTGAN [24]	✗	9.30	6.66	1
StyleGAN2-ADA [17]	✗	9.83	2.92	1
StyleGAN2-ADA (cond.) [17]	✗	10.14	2.42	1
NCSN [31]	✗	8.87	25.32	1001
NCSNv2 [32]	✗	8.40	10.87	1161
DDPM [16]	✗	9.46	3.17	1000
NCSN++ VE-SDE [33]	✗	9.83	2.38	2000
NCSN++ deep VE-SDE [33]	✗	9.89	2.20	2000
Glow [19]	✓	3.92	48.9	1
DDIM, T=50 [30]	✓	-	4.67	50
DDIM, T=100 [30]	✓	-	4.16	100
NCSN++ VE-ODE [33]	✓	9.34	5.29	194
NCSN++ deep VE-ODE [33]	✓	9.17	7.66	194
DDPM++ backbone				
VP-SDE [33]	✗	9.58	2.55	1000
sub-VP-SDE [33]	✗	9.56	2.61	1000
VP-ODE [33]	✓	9.46	2.97	134
sub-VP-ODE [33]	✓	9.30	3.16	146
PFGM (ours)	✓	**9.65**	**2.48**	**104**
DDPM++ deep backbone				
VP-SDE [33]	✗	9.68	2.41	1000
sub-VP-SDE [33]	✗	9.57	2.41	1000
VP-ODE [33]	✓	9.47	2.86	134
sub-VP-ODE [33]	✓	9.40	3.05	146
PFGM (ours)	✓	**9.68**	**2.35**	**110**

PFGM的实验结果（部分）

文章小结

本文介绍了一个由“万有引力定律”启发的ODE式扩散模型，它突破了以往众多扩散模型对高斯假设的依赖，是一个基于场论来构建ODE式扩散模型的全新框架，整个模型颇多启发性，值得仔细研读。

如果您需要引用本文，请参考：

苏剑林. (Oct. 18, 2022). 《生成扩散模型漫谈（十三）：从万有引力到扩散模型》[Blog post]. Retrieved from https://spaces.ac.cn/archives/9305

```
@online{kexuefm-9305,
    title={生成扩散模型漫谈（十三）：从万有引力到扩散模型},
    author={苏剑林},
    year={2022},
    month={Oct},
    url={\url{https://spaces.ac.cn/archives/9305}},
}
```