A Survey on Probabilistic Modeling of Data: From Traditional to Modern

LU Hongtao*, HU Yuting

(School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Probabilistic modeling of data is the core in machine learning and modern generative AI. This survey reviews the methodological evolution from traditional statistical formulations to recent deep generative frameworks under a unified view of probability distribution learning. Representative methods are organized into three connected routes: Maximum-likelihood-based modeling, score-matching-based modeling, and flow-based modeling. On the traditional side, the survey revisits Gaussian assumptions, Gaussian mixture models, expectation-maximization (EM) algorithms, and variational inference, emphasizing how tractability-flexibility trade-offs shape model design. On the modern side, it discusses variational autoencoders (VAEs), generative adversarial net (GAN)-related generative mechanisms, diffusion probabilistic models, score-based stochastic differential equation (SDE) formulations, normalizing flows, and flow matching, with focus on objective functions, parameterization choices, and sampling dynamics. A structured comparison is provided from the perspectives of explicit likelihood, trajectory modeling, computational efficiency, controllability, and deployment stability. To bridge methodology and practice, the paper summarizes benchmark-oriented observations and application trends in image generation, video and audio synthesis, inverse problems, and science-and-control scenarios. It also identifies practical bottlenecks, including dependence on high-quality large-scale data, limited semantic operability of latent representations, and inference latency caused by multi-step sampling. Finally, future directions are discussed around coordinated advances in path design, training objectives, numerical solvers, and guidance strategies, together with unified evaluation over quality, efficiency, safety, and compliance for trustworthy large-scale deployment.

Highlights:

1. A unified probabilistic perspective links classical methods and modern generative models through likelihood, score, and flow formulations.
2. The survey compares diffusion and flow matching under shared criteria of quality, efficiency, stability, and controllability.
3. Application analysis emphasizes inverse problems, multimodal generation, and deployment bottlenecks in data, representation, and inference.

Key words: probabilistic modeling; maximum likelihood estimation; score matching; diffusion model; flow matching; generative model

---

Foundation item: National Natural Science Foundation of China (No. 62176155).

Received: 2026-01-30; **Revised**: 2026-03-08

***Corresponding author**, **E-mail**: htlu@sjtu.edu.cn.

ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 461－488
DOI:10. 16337/j. 1004-9037. 2026. 02. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

数据的概率建模综述:从传统到现代

卢宏涛，胡宇庭

(上海交通大学计算机学院,上海 200240)

摘　要: 人工智能技术发展日新月异,各类模型、算法及其应用领域受到较大关注。数据的概率建模是人工智能和机器学习的核心问题,但是其关注度普遍较低。这一方面是由于概率建模理论抽象,另一方面是相关综述较少。然而人工智能领域的原创性突破大多都与数据概率建模有关,因此本文以数据的概率建模为主线,对机器学习中从传统到现代的主流方法进行综述,从高斯混合模型、期望最大化(Expectation-maximization,EM)算法和变分推理等传统方法到变分自编码器、生成对抗网、分数匹配、扩散模型、归一化流和流匹配等现代方法都统一到数据的概率建模框架下。这些方法虽然提出的时间跨度很大,解决的问题有所不同,但它们都可以解释为最大似然估计或分数匹配框架,区别在于对数据及模型的假设不同。因此,本文构建了一种对从传统机器学习到最新生成模型的统一理解方式,将概率建模方法分为基于最大似然估计的方法、基于分数匹配的方法和基于流的方法,揭示了它们之间的内在联系,为人工智能生成方法的进一步发展提供了理论基础方面的解读。

关键词: 概率建模;最大似然估计;分数匹配;扩散模型;流匹配;生成模型

中图分类号: TP18　　**文献标志码:** A

引用格式: 卢宏涛,胡宇庭.数据的概率建模综述:从传统到现代[J].数据采集与处理,2026,41(2):461-488. LU Hongtao,HU Yuting. A survey on probabilistic modeling of data: From traditional to modern[J].Journal of Data Acquisition and Processing,2026,41(2):461-488.

引　　言

机器学习的本质是对数据进行建模,概率建模是主要的方法,其主线贯穿从传统机器学习到深度学习,再到当前的生成模型和大模型。本文以数据概率建模的思想,统一描述这些不同时期、不同背景的方法,由此可以深入理解它们之间的深刻联系。

假设随机变量$\boldsymbol{x}\in\mathbf{R}^d$表示观察变量,其概率分布为$p_x(\cdot)$,实际问题中的$p_x(\cdot)$通常是未知且非常复杂的。观察到$N$个服从$p_x(\cdot)$分布的数据$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,本文的目标是估计出$p_x(\cdot)$的一个近似分布。最大似然估计是一个古老的、一般性的概率建模框架[1],但直到今日仍然起着非常重要的作用。假设概率密度函数$p_x(\cdot)$具有特定的函数形式,比如高斯函数或其他的函数形式,函数中包含了参数$\boldsymbol{\theta}$,概率分布可以写成显式包含参数的形式$p(\boldsymbol{x},\boldsymbol{\theta})$。最大似然估计方法是关于参数$\boldsymbol{\theta}$最大化似然函数,表达式为

$$\max_{\boldsymbol{\theta}}\prod_{n=1}^{N}p(\boldsymbol{x}_n,\boldsymbol{\theta}) \tag{1}$$

基金项目:国家自然科学基金(62176155)。
收稿日期:2026-01-30;**修订日期:**2026-03-08

式中似然函数$\prod_{n=1}^{N} p(\boldsymbol{x}_n,\boldsymbol{\theta})$是参数$\boldsymbol{\theta}$的函数,它本身不是概率密度函数。最大似然估计的基本思想是估计出的参数使得概率分布可以最好地(最大可能性地)解释观察数据。由于似然函数中N项连乘计算复杂,很多情况下对似然函数取对数,最大似然通过如下最大化对数似然实现,乘积变为求和,即

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n,\boldsymbol{\theta}) \tag{2}$$

由于对数函数是单调递增函数,式(2)的解与式(1)的解一致。对数似然理论上可定义为模型概率密度的对数$\ln p(\boldsymbol{x},\boldsymbol{\theta})$关于数据分布$p_x$的期望$E_{p_x}[\ln p(\boldsymbol{x},\boldsymbol{\theta})]$,对于连续变量,可以写成积分$E_{p_x}[\ln p(\boldsymbol{x},\boldsymbol{\theta})]=\int p_x \ln p(\boldsymbol{x},\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}$。当仅有$N$个观察数据时,期望即由式(2)样本均值近似。

另一类参数估计的方法是分数匹配。分数匹配方法2005年在文献[2]中被提出,近年来随着扩散模型的兴起(扩散模型可以解释成一种分数匹配方法),分数匹配再次受到关注。分数匹配主要应用于估计非归一化的概率模型。在实际应用中,很多情况下难以得到准确的概率密度函数,仅能得到相差一个乘积倍数的非归一化的概率密度函数,此时概率模型$p(\boldsymbol{x},\boldsymbol{\theta})$可以写为

$$p(\boldsymbol{x},\boldsymbol{\theta})=\frac{1}{Z(\boldsymbol{\theta})} q(\boldsymbol{x},\boldsymbol{\theta}) \tag{3}$$

式中:$Z(\boldsymbol{\theta})$为一个未知的归一化常数,在实际问题中是很难计算得到的;$q(\boldsymbol{x},\boldsymbol{\theta})$为数据变量$\boldsymbol{x}$和参数$\boldsymbol{\theta}$的函数,本身不一定满足概率密度函数的归一化要求。概率模型的分数函数$s(\boldsymbol{x},\boldsymbol{\theta})$定义为其对数概率密度关于数据变量$\boldsymbol{x}$的梯度[2],即

$$s(\boldsymbol{x},\boldsymbol{\theta})=\nabla_x \ln p(\boldsymbol{x},\boldsymbol{\theta}) \tag{4}$$

同样地,对观察数据也可以定义其分数函数$s_x(\cdot)=\nabla_x \ln p_x(\cdot)$。分数匹配的思想是优化参数$\boldsymbol{\theta}$,使模型的分数函数尽量地匹配数据的分数函数。具体地,分数匹配可表示为平方误差最小化问题,即

$$\hat{\boldsymbol{\theta}}=\operatorname*{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{5}$$

其中,目标函数为

$$J(\boldsymbol{\theta})=\frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbf{R}^d} p_x(\boldsymbol{\xi})\|s(\boldsymbol{\xi},\boldsymbol{\theta})-s_x(\boldsymbol{\xi})\|^2 \mathrm{d}\boldsymbol{\xi} \tag{6}$$

根据分数函数的定义(式(4)),分数函数也等于$\nabla_x \ln q(\boldsymbol{x},\boldsymbol{\theta})$,因此不依赖于归一化常数$Z(\boldsymbol{\theta})$。分数匹配的结果也与$Z(\boldsymbol{\theta})$无关。因此,不同于最大似然估计,分数匹配避免了计算归一化常数的困难,在某些情况下能取得更好、更鲁棒的结果。目标函数(6)中包含了数据的分数函数,但这是未知的。因此优化问题(5)不可求解。幸运的是,文献[2]的定理1基于微积分的分部积分方法证明了如果分数函数可微,且满足一定的条件,则目标函数(6)仅依赖于模型的分数函数,而不依赖于观测数据的分数函数。具体地,$J(\boldsymbol{\theta})$可表示为

$$J(\boldsymbol{\theta})=\frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbf{R}^d} p_x(\boldsymbol{\xi}) \sum_{i=1}^{d}\left[\partial_i s_i(\boldsymbol{\xi},\boldsymbol{\theta})+\frac{1}{2}s_i(\boldsymbol{\xi},\boldsymbol{\theta})^2\right]\mathrm{d}\boldsymbol{\xi}+\mathrm{const} \tag{7}$$

式中:const为不依赖于参数$\boldsymbol{\theta}$的常数,$s_i(\boldsymbol{\xi},\boldsymbol{\theta})$为$d$维分数函数$s(\boldsymbol{\xi},\boldsymbol{\theta})$的第$i$个分量函数,$\partial_i$表示对第$i$个分量求偏导数。

近年来,基于流的生成模型取得了极大的成功,但基于流的方法最早也是基于最大似然估计的框架。由于当前最先进的方法是基于流匹配的方法,本文将按照最大似然估计、分数匹配和归一化流3个方向对数据的概率建模进行分类介绍。

1 数据概率建模的最大似然估计方法

1.1 传统方法

由于实际数据的分布往往非常复杂,是未知且不可求解的,概率建模的思想是对其进行各种简化假设,使其变为可求解的问题。但过强的假设使得模型过分的简化,又会使得假设分布远离真实分布,导致模型的数据建模性能变差。因此,需要尽量弱的假设,使模型空间具有比较大的灵活性和表达能力。机器学习方法总是在可求解性和灵活性之间寻求折中。

1.1.1 高斯分布

不失一般性,本文主要考虑连续的随机变量。最简单的连续随机变量是其概率密度函数具有解析表达式,如均匀分布、高斯分布和柯西分布等。如果假设数据服从高斯分布,即

$$p(\boldsymbol{x},\boldsymbol{\theta})=N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

式中均值向量$\boldsymbol{\mu}\in\mathrm{R}^d$和协方差矩阵$\boldsymbol{\Sigma}\in\mathrm{R}^{d\times d}$构成需要估计的参数$\boldsymbol{\theta}$。高斯分布是很强的假设,此时最大似然估计的结果有闭式解,均值和协方差矩阵的估计结果分别是样本均值向量和样本协方差矩阵[1],与直觉高度符合。

对于高斯分布,一个有趣的结果是其分数匹配估计结果与最大似然估计结果一致[2]。这一结论对其他分布一般不成立。

1.1.2 高斯混合模型和期望最大化算法

高斯混合模型比单个的高斯具有更丰富的表达能力,但如果假设数据的分布符合混合高斯分布,即

$$p(\boldsymbol{x},\boldsymbol{\theta})=\sum_{k=1}^{K}\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

式中:π_k为第k个高斯分布发生的概率,$\sum_{k=1}^{K}\pi_k=1$,$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$分别为第k个高斯分布的均值向量和协方差矩阵,模型参数$\boldsymbol{\theta}=\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$。此时,最大似然估计无闭式解[1],这是因为对数似然函数对多个高斯的和计算对数,不再像单个高斯分布情况下那么容易计算对数。高斯混合模型最大似然估计的经典求解方法是期望最大化(Expectation-maximization,EM)算法[1],包含E步和M步的交替迭代。给定观测数据集$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,E步对每个数据点$\boldsymbol{x}_n$($n=1,2,\cdots,N$)计算其属于第$k$个高斯分量的概率$\gamma_{nk}$(该数值对应于后文的关于隐变量的后验概率),在这一步每个高斯分量的参数是前一步计算出来的;在M步,基于前一步的参数和γ_{nk}对每个高斯分量的均值向量、协方差矩阵和π_k进行迭代更新。求解高斯混合模型的EM算法可以直接从最大对数似然推导出来[1]。

高斯混合模型的EM算法还可以从带有隐变量的图模型角度进行解释[1],并推广到更一般的EM算法、变分推理和变分自编码器(Variational autoencoder, VAE)中。图模型[1]是用图的方式可视化地表示联合概率分布的方法,其中节点表示随机变量,节点之间的有向边(边也可以是无向的)表示随机变量之间的依赖性。对于高斯混合模型的例子,每个数据点$\boldsymbol{x}$有一个相应的隐变量$\boldsymbol{z}$,表示$\boldsymbol{x}$来自于哪个高斯分量,通过图模型,高斯混合模型的联合概率分布表示为$p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{z})\,p(\boldsymbol{x}|\boldsymbol{z})$。

1.1.3 一般EM算法

文献[1]给出了一个一般EM算法,用以求解一般的具有隐变量的图模型最大似然估计问题。把N个观测数据作为行向量构成一个矩阵$\boldsymbol{X}$,对应的隐变量向量构成矩阵$\boldsymbol{Z}$,根据全概率公式,对数似然函数可以写为

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta})=\ln\left\{\sum_{\boldsymbol{Z}}p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right\} \tag{8}$$

式中:$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$表示对数似然函数依赖于参数$\boldsymbol{\theta}$,而不是条件概率;条件概率也使用相同的符号"|",从上下文中可以区分。这里假设隐变量是离散的,对于连续隐变量,求和变成积分,下面的结论也成立。式(8)的右边由于是对求和号进行对数运算,因此即使$\boldsymbol{X}$、$\boldsymbol{Z}$的联合分布$p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$是指数族分布,由于求和号的阻隔,对数不能直接作用于$p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$,式(8)的最大似然问题一般也是不可解的。文献[1]把$\boldsymbol{X}$称为不完整数据,把$(\boldsymbol{X},\boldsymbol{Z})$一起称为完整数据。完整数据的对数似然函数为$\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$,一般假设这个对数似然函数容易计算(比如完整数据的分布属于指数族)。但在实际问题中并不知道隐变量$\boldsymbol{Z}$,仅有不完整数据$\boldsymbol{X}$。假如能计算出给定观察数据$\boldsymbol{X}$下$\boldsymbol{Z}$的后验概率$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$,由于直接最大化式(8)比较困难,可以转而最大化完整数据对数似然函数$\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$关于后验概率$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$的期望,即最大化如下目标函数

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})=\sum_{\boldsymbol{Z}}p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \tag{9}$$

式中$\boldsymbol{\theta}^{\text{old}}$是EM算法前一步迭代估计出的参数。目标函数(9)的含义可解释为完整数据的对数似然,但由于不知道隐变量,因此通过求期望,把隐变量的影响平均掉(求期望)。

一般EM算法的E步即是先计算给定观察数据$\boldsymbol{X}$下隐变量$\boldsymbol{Z}$的后验概率$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$,然后计算联合似然函数的期望(式(9));M步则是优化目标函数(9),得到新的参数估计值。如此反复迭代,直到收敛。可以证明,高斯混合模型的EM算法可以纳入这个一般EM算法的框架。

1.1.4 更一般的EM算法

文献[1]中还介绍了一个更一般的EM算法框架,它也是变分推理和其他概率模型的基础。1.1.3节中的一般EM算法假设能计算出隐变量$\boldsymbol{Z}$的后验概率$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$,但在很多实际问题中,隐变量是未知的,其后验概率也难以计算。在此情况下,引入关于$\boldsymbol{Z}$的另外一个概率密度函数$q(\boldsymbol{Z})$(称为建议分布,可能更简单或者具有某些特定性质)来近似并替换$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$。由于$q(\boldsymbol{Z})$的引入,观察数据的对数似然可以写成如下形式

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta})=L(q,\boldsymbol{\theta})+\text{KL}(q\|p) \tag{10}$$

$$L(q,\boldsymbol{\theta})=\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{11}$$

$$\text{KL}(q\|p)=-\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{12}$$

式(12)是$q(\boldsymbol{Z})$和$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$之间的KL(Kullback-Leibler)散度,式(11)只依赖于近似分布$q(\boldsymbol{Z})$和参数$\boldsymbol{\theta}$。要注意的是式(12)中包含了隐变量的后验概率,式(11)中包含了完整数据的联合分布,展开后发现$L(q,\boldsymbol{\theta})$等于完整数据的对数似然关于$q(\boldsymbol{Z})$的期望与$q(\boldsymbol{Z})$熵的和。由于KL散度总是非负的,当且仅当$q=p$时等于0,因此$L(q,\boldsymbol{\theta})$是对数似然$\log p(\boldsymbol{X}|\boldsymbol{\theta})$的下界。如果选择$q(\boldsymbol{Z})=p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$,则式(10)中的KL散度等于0,式(10)中的对数似然等于其下界(11),式(11)等于完整数据的对数似然关于后验分布的期望(这部分正是式(9)中的目标函数,这也说明了定义该目标函数的原因)加上后验概率的熵。

式(10)在后续方法中起着根本作用,其证明有多种方式,下面给出其中一种证明,推导公式为

$$\text{KL}(q\|p)=\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}=\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\{\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\}=$$

$$\sum_Z q(\boldsymbol{Z})\{\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})+\ln p(\boldsymbol{X}|\boldsymbol{\theta})\}=$$
$$\sum_Z q(\boldsymbol{Z})\{\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\}+\ln p(\boldsymbol{X}|\boldsymbol{\theta})=-L(q,\boldsymbol{\theta})+\ln p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{13}$$

式中:第3个等号基于贝叶斯公式;第4个等号基于$q(\boldsymbol{Z})$是一个概率密度函数,其和为1,且$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$与$q(\boldsymbol{Z})$无关。

更一般的EM算法中,E步为在前一步参数$\boldsymbol{\theta}^{\text{old}}$基础上,关于近似分布$q(\boldsymbol{Z})$最大化似然函数的下界$L(q,\boldsymbol{\theta})$。从式(10)看出,这在KL散度等于0时取得,即$q(\boldsymbol{Z})=p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$。由于这个后验概率不能计算,只能取各种近似,导致不能取得最优解,这与一般EM算法一致。在M步,固定$q(\boldsymbol{Z})=p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$,关于参数$\boldsymbol{\theta}$最大化似然函数(10),由于KL散度等于0,似然函数等于其下界,即式(11)。容易证明,此时式(11)的下界函数$L(q,\boldsymbol{\theta})$与1.1.3节的$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$函数只相差一个常数,其解与一般EM算法的M步一致。

1.1.5 变分推理

变分推理[1]要解决的就是上述最大似然估计中后验概率分布$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$不可计算的场景。此时不可能令$q(\boldsymbol{Z})=p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$,只能以容易计算的$q(\boldsymbol{Z})$对$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$进行近似。优化过程类似于1.1.4节,最大化如下变分下界$L(q)$(此时将参数$\boldsymbol{\theta}$吸收到隐变量$\boldsymbol{Z}$中),或等价地,最小化$q(\boldsymbol{Z})$与$p(\boldsymbol{Z}|\boldsymbol{X})$之间的KL散度,即

$$L(q)=\int q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\mathrm{d}\boldsymbol{Z} \tag{14}$$

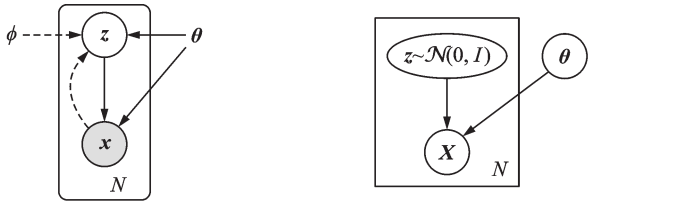
一种常用的变分推理是使用可分解分布,假设$q(\boldsymbol{Z})$可分解为$\boldsymbol{Z}$的一些不相交子集变量分布的乘积,即$q(\boldsymbol{Z})=\prod_{i=1}^{M}q_i(\boldsymbol{Z}_i)$,其中隐变量$\boldsymbol{Z}$分成了$M$个不相交的子集$\boldsymbol{Z}_i(i=1,2,\cdots,M)$,$q_i(\boldsymbol{Z}_i)$为关于$\boldsymbol{Z}_i$的分布,具体形式不做任何假设。这种变分推理方式在物理中称为平均场理论。分解分布本质上是假设隐变量可以分解为独立的多个变量子集,可以用图模型来表示联合分布$q(\boldsymbol{Z})$,其中各组子集$\boldsymbol{Z}_i$之间无边连接,可以通过子集之间的独立性来限制分布的形式。最大似然的变分推理现在转化为对每个分量分布$q_i(\boldsymbol{Z}_i)$来最大化下界$L(q)$,$q_i(\boldsymbol{Z}_i)$的详细推导见文献[1]。

以上传统概率建模方法都是基于最大似然估计框架。如果假设参数$\boldsymbol{\theta}$本身是随机变量,满足一定的先验分布,则最大似然估计变为最大后验估计(Maximum a posterior, MAP)。优化过程与最大似然的区别是在似然目标函数中多了与$\boldsymbol{\theta}$的先验分布相关的正则化项。

1.2 变分自编码器

VAE是文献[3]首先提出的一种近似推理和学习方法,旨在解决具有连续隐变量的有向概率图模型中不可求解的后验分布问题,其仍然是一个最大似然估计框架。文献[4]以更易于理解的方式重述了VAE。VAE的背景与前述传统模型一致,假设观察到N个iid数据集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,每个数据点对应着一个隐变量,相应的隐变量集合$\boldsymbol{Z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_N\}$。与高斯混合模型不同的是,这里的隐变量是连续的,而且可以假设它们是一维标准高斯分布噪声。VAE的概率图模型表示如图1所示。图1(a)中的图模型可以从生成模型的角度解释,先随机采样一个高斯噪声样本$\boldsymbol{z}$,然后以条件概率$p(\boldsymbol{x}|\boldsymbol{z})$采样生成一个样本$\boldsymbol{x}$。图1(b)与高斯混合模型的图模型很相似,不同的是高斯混合模型的隐变量是服从伯努利分布的离散二值变量,这里的隐变量是服从标准高斯分布的连续随机变量。1.1.2节的高斯混合模型也可以从类似的生成模型的角度解释。扩散模型也是类似的从高斯噪声生成数据。图1(a)中从$\boldsymbol{x}$到$\boldsymbol{z}$的虚线边即是不可处理的后验概率$p(\boldsymbol{z}|\boldsymbol{x})$的近似概率$q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\phi})$,具有参数$\boldsymbol{\phi}$,在VAE中用神经

网络进行近似，对应VAE的编码器，ϕ是神经网络的可学习参数。而从z到x的条件概率$p(x|z,\theta)$对应VAE的解码器，具有参数θ，对应着解码器神经网络的可学习参数。



(a) VAE的图模型[3] (a) Graphical model of VAE[3]　(b) VAE生成过程图模型[4] (b) Graphical model of VAE generation process[4]

图1 VAE的图模型及其生成过程图模型

Fig.1 Graphical model of VAE and graphical model of VAE generation process

VAE、生成对抗网、扩散模型等生成模型都将数据归结到高斯噪声，然后再从高斯噪声生成数据。这是因为高斯分布是最简单的连续随机变量，很容易从高斯分布采样。生成模型生成数据的本质就是对真实复杂分布进行采样，对复杂分布的实际数据通过神经网络映射（VAE）、不断添加噪声（扩散模型）、微分方程变换（流匹配）等机制将其分布变换为简单分布（如高斯噪声），然后再通过神经网络等学习反向的过程，将高斯分布的噪声映射回复杂分布的数据样本。

VAE的理论推导过程类似于更一般的EM算法中的式（13），由于这里隐变量是连续的，用积分代替求和，可得

$$\mathrm{KL}(q\|p)=\int_z q(\boldsymbol{Z})\ln\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\mathrm{d}\boldsymbol{Z}=\int_z q(\boldsymbol{Z})\{\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\}\mathrm{d}\boldsymbol{Z}+\ln p(\boldsymbol{X}|\boldsymbol{\theta})=$$

$$\int_z q(\boldsymbol{Z})\{\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})-\ln p(\boldsymbol{Z})\}\mathrm{d}\boldsymbol{Z}+\ln p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{15}$$

为与前面表示一致，式（15）中用黑体大写符号对应图1中的小写或非黑体符号。式（15）推导的第2个等号对应于式（13）的第4个等号，第3个等号进一步利用贝叶斯公式把$\boldsymbol{Z}$分离出来。将式（15）重新整理，写成期望的形式，并写出概率密度函数对变量和参数的显式依赖，即

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta})-\mathrm{KL}(q((\boldsymbol{Z}|\boldsymbol{X},\phi))\|p(\boldsymbol{Z}|\boldsymbol{X}))=E_q[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})]-\mathrm{KL}(q((\boldsymbol{Z}|\boldsymbol{X},\phi)\|p(\boldsymbol{Z})) \tag{16}$$

式（16）即文献[4]中的式（5），右边是对数似然函数的变分下界。在观察数据集上变分下界的样本近似替代期望也称为证据下界（Evidence lower bound，ELBO）。变分下界可以用神经网络来近似，通过随机梯度算法进行迭代优化。值得注意的是，式（16）两边的KL散度不同，左边的p是$\boldsymbol{X}$条件下$\boldsymbol{Z}$的后验概率$p(\boldsymbol{Z}|\boldsymbol{X})$，右边KL散度中的概率$p(\boldsymbol{Z})$是隐变量$\boldsymbol{Z}$的先验概率，此处即标准高斯分布。

VAE利用两个多层神经网络分别近似从$\boldsymbol{X}$到$\boldsymbol{Z}$的后验概率$q(\boldsymbol{Z}|\boldsymbol{X},\phi)$和从$\boldsymbol{Z}$到$\boldsymbol{X}$的生成概率$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})$，两个网络中间通过隐变量$z$连接起来，并通过重参数化实现端到端的训练，其结构如图2所示①。VAE是深度隐变量模型（Deep latent variable model，DLVM）的一种特例[5]。DLVM用深度网络来近似联合概率或后验概率，其面临的困难在于ELBO关于参数的梯度如何计算及反传。重参数化[4-5]可以解决这一问题，其在扩散模型[6]中也有应用。重参数化本质是一种变量替换[5]，通过一系列的变量替换使随机变量的分布从简单分布变换为实际数据的复杂分布，这也是归一化流[7]和流

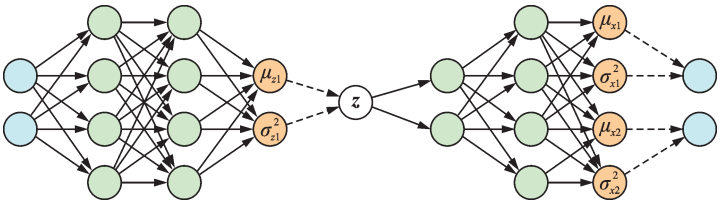


图2 VAE结构图①

Fig.2 Architecture of VAE①

①https://cseweb.ucsd.edu/~dasgupta/254-deep-ul/fangchen.pdf

匹配方法[8]的基本思路。VAE的优点是重构数据的自监督学习,不需要数据标签,训练稳定,而且隐藏层提供了数据的表示,因此VAE不仅在图像、视频、语音等生成领域都获得了成功应用,而且在提出10年后获得2024年国际表示学习大会(ICLR)首个时间检验奖。

VAE也有许多的推广,包括推广到动态系统[9],一种具有多级随机隐变量的模型[10]。近年来提出的扩散模型[6,11]也是一种多级随机隐变量模型。隐变量分布是自回归模型的深度自回归模型[12],原始的VAE假设隐变量是连续的,VQ-VAE[13-14]可应用于离散隐变量。VQ-VAE输出数据的离散表征,而且隐变量的先验分布不是像VAE那样给定为高斯分布,而是学习出来的自回归分布。

尽管VAE奠定了深度生成模型中的概率推断范式,但其在高质量视觉生成任务上仍有显著局限性。首先,在常见高斯观测假设下,ELBO中的重构项通常退化为像素级均方误差,模型倾向于学习条件均值,导致生成结果容易出现过度平滑,纹理与边缘细节不足[15-16]。其次,当解码器表达能力过强时,训练会出现后验坍塌,即隐变量对重构的贡献被弱化,潜空间表征能力下降[10,17]。此外,标准VAE的潜空间虽然可以连续且可插值,但因子解耦与可控编辑能力通常不足[16-17]。

针对上述局限性,研究者们提出了多条改进路线:超越像素级的自编码方法通过学习特征空间相似度替代逐像素重构,缓解了视觉模糊[15];层次化VAE(如NVAE)通过多层隐变量结构增强复杂分布建模能力[10];beta-VAE通过调节KL散度项权重强化潜变量解耦,提高语义可解释性与可控性[17]。进一步地,LDM将VAE作为高效潜表示模块,把扩散过程迁移到低维潜空间,在计算效率与生成质量之间取得了更好的平衡[18]。

1.3 生成对抗网络

生成对抗网络(Generative adversarial net, GAN)[19]包含一个生成器G和一个判别器D,生成器和判别器都由多层感知机(Multilayer perceptron,MLP)实现,生成器的输入是噪声,可以解释为满足高斯分布的隐变量z,生成器G通过MLP将高斯噪声映射为期望的输出数据,输出数据的分布通过判别器D来控制,判别器对生成器的输出进行分类,判断其来自于真实数据还是输出数据。GAN整体是一个两参与者的Minmax博弈,当训练算法收敛后,文献[19]证明了若达到全局最优解,则生成器输出数据的分布等于数据的分布。判别器D的目标函数可解释为通过估计某个数据点来自于数据的分布还是来自于生成器的分布概率的最大似然估计。且训练GAN相当于最小化数据分布和生成器分布之间的JS(Jensen-Shannon)散度,其定义为两个分布之和与它们平均分布之间KL散度的平均,即

$$\mathrm{JSD}(p_{\mathrm{Data}}\|p_{\mathrm{G}})=\mathrm{KL}\left(p_{\mathrm{Data}}\|\frac{p_{\mathrm{Data}}+p_{\mathrm{G}}}{2}\right)+\mathrm{KL}\left(p_{\mathrm{G}}\|\frac{p_{\mathrm{Data}}+p_{\mathrm{G}}}{2}\right) \tag{17}$$

不同于KL散度,JS散度是对称的;但JS散度等于零也当且仅当两个分布相同,最小化JS散度与最小化KL散度有类似效果,说明GAN可以解释为一种隐式最大似然估计方法。GAN在高锐度图像生成方面具有优势,但训练动力学较为敏感。由于生成器与判别器通过对抗博弈共同优化,梯度场往往呈现非平稳性,实践中容易出现振荡、收敛缓慢或对超参数高度敏感等现象。与此同时,模式崩塌仍是核心风险之一:生成器可能偏向少数高概率模式而牺牲分布覆盖,导致样本多样性不足。对原始GAN而言,当两分布支持集重叠较小时,基于JS散度的优化也可能出现有效梯度不足的问题[19]。

针对上述问题,代表性变体从目标函数、条件建模与结构设计3方面推进:WGAN以Wasserstein距离替代JS散度,显著改善了训练可解释性与稳定性[20];Conditional GAN将标签或语义条件显式注入生成器与判别器,实现可控生成[21];CycleGAN通过循环一致性约束实现非配对图像翻译,拓展了GAN在跨域映射任务中的适用性[22];StyleGAN系列则通过风格分层控制机制显著提升高分辨率图像的保真度与可编辑性,成为高质量图像合成的重要基线[23]。

1.4 扩散概率模型

扩散概率模型(Diffusion probabilistic model, DPM)由文献[24]最先提出,其目标是为复杂数据进

行建模,取得灵活性和可处理性之间的良好平衡。扩散概率模型构建原理是受到非平衡热力学物理的启发,通过前向扩散过程不断地在数据分布中添加噪声,逐步破坏数据结构,然后学习反向过程从噪声中恢复数据结构,其本质仍然是一个具有隐变量的概率图模型中的最大似然估计方法。

扩散概率模型中的前向轨迹相当于VAE中的编码过程,$\boldsymbol{x}^{(0)}$是观察数据,$\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\cdots,\boldsymbol{x}^{(T)}$是扩散过程中不断加入噪声后的数据,可视为与$\boldsymbol{x}^{(0)}$同维数的隐变量序列,这些变量序列组成马尔可夫序列,观察数据与隐变量的联合概率分布(相当于前文提到的完整数据的联合分布)为

$$q(\boldsymbol{x}^{(0,1,\cdots,T)})=q(\boldsymbol{x}^{(0)})\prod_{t=1}^{T}q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}) \tag{18}$$

式中$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$为高斯分布(连续)或者二项分布(离散)。

反向轨迹由如下概率模型描述

$$\begin{cases} p(\boldsymbol{x}^{(T)})=\pi(\boldsymbol{x}^{(T)}) \\ p(\boldsymbol{x}^{(0,1,\cdots,T)})=p(\boldsymbol{x}^{(T)})\prod_{t=1}^{T}p(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}) \end{cases} \tag{19}$$

式中$\pi(\boldsymbol{x}^{(T)})$为前向扩散过程最后的分布,通常为高斯或二项分布。当步长足够小,即轨迹足够长时,$p(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)})$与$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$是一样的分布。可以很容易地计算出数据的边缘分布$p(\boldsymbol{x}^{(0)})$,其可表示为前后向条件分布[24]。

扩散概率模型的训练就是最大化模型的对数似然函数L。对数似然可以定义为模型的对数概率密度(即$\ln p(\boldsymbol{x}^{(0)})$)关于数据分布(即$q(\boldsymbol{x}^{(0)})$)的期望。经过推导和简化,可得[24]

$$L\geqslant\int q(\boldsymbol{x}^{(0)})\left\{\int q(\boldsymbol{x}^{(1,2,\cdots,T)}|\boldsymbol{x}^{(0)})\ln\left[\frac{p(\boldsymbol{x}^{(0,1,\cdots,T)})}{q(\boldsymbol{x}^{(1,2,\cdots,T)}|\boldsymbol{x}^{(0)})}\right]\mathrm{d}\boldsymbol{x}^{(1,2,\cdots,T)}\right\}\mathrm{d}\boldsymbol{x}^{(0)}=$$

$$\int q(\boldsymbol{x}^{(1,2,\cdots,T)}|\boldsymbol{x}^{(0)})\ln\left[\frac{p(\boldsymbol{x}^{(0,1,\cdots,T)})}{q(\boldsymbol{x}^{(1,2,\cdots,T)}|\boldsymbol{x}^{(0)})}\right]\mathrm{d}\boldsymbol{x}^{(1,2,\cdots,T)} \tag{20}$$

式(20)与式(11)进行比较发现,$\boldsymbol{x}^{(0)}$对应于观察数据$\boldsymbol{X}$,$\boldsymbol{x}^{(1,2,\cdots,T)}$对应于隐变量$\boldsymbol{Z}$,前向过程的条件概率$q(\boldsymbol{x}^{(1,2,\cdots,T)}|\boldsymbol{x}^{(0)})$相当于隐变量的近似概率$q(\boldsymbol{Z})$。因此,式(20)的右边正是(11)和式(14)中的变分下界$L(q)$。这进一步说明了扩散概率模型与更一般的EM算法和变分推理一样,是一种最大似然变分推理框架。

进一步对扩散轨道,文献[24]证明了对数似然下界可以写成有解析式计算的熵和KL散度的形式,即

$$L\geqslant K=-\sum_{t=2}^{T}\int q(\boldsymbol{x}^{(0)},\boldsymbol{x}^{(t)})\mathrm{KL}(q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})\|p(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}))\mathrm{d}\boldsymbol{x}^{(0)}\mathrm{d}\boldsymbol{x}^{(t)}+$$

$$H_q(\boldsymbol{X}^{(T)}|\boldsymbol{X}^{(0)})-H_q(\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(0)})-H_p(\boldsymbol{X}^{(T)}) \tag{21}$$

式中$H(\cdot)$为熵。

扩散模型的前向概率$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$一般是预先给定的,训练是寻找反向马尔可夫转移概率$\hat{p}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)})$使对数似然下界最大,即

$$\hat{p}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)})=\arg\max_{p(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)})}K \tag{22}$$

对于高斯或伯努利分布,以上估计概率分布的训练目标简化为估计均值和协方差矩阵(或翻转概

率)的回归问题[24]。对高斯分布,用多层多尺度卷积神经网络学习均值和协方差;对伯努利分布,用多层感知机回归概率。文献[24]及后续研究通过大量的图像生成实验证明了扩散模型的良好生成性能,并与GAN等其他生成模型进行了对比。

从统一视角看,扩散模型首先构造一个前向扰动过程,通过持续注入噪声把数据分布逐步连接到易采样先验分布;随后学习一个反向生成过程,使噪声能够沿相反方向回到数据流形。该框架的关键优势在于训练目标稳定:无论采用变分形式还是之前介绍的分数匹配形式,本质都可转化为可优化的回归问题,从而避免了对抗训练中的训练目标不稳定性[6,16,24]。此外,扩散模型通常具备较强的分布覆盖能力,兼顾了样本多样性与保真度。然而扩散模型的主要代价是采样效率。标准反向过程往往需要多步迭代,推理时长与算力开销显著高于单步生成范式;同时,随机反向路径虽然有助于多样性,但在精细可控编辑和可逆映射需求下并不总是最优。Lai等[16]指出,扩散模型的性能边界往往由概率路径设计、训练目标选择与数值求解器效率三者共同决定,而非单一模块独立决定。

围绕“降本增效”与“增强可控”两条主线,扩散模型形成了清晰的演进路径。(1)潜空间路线以潜扩散模型(Latent diffusion model, LDM)为代表:通过VAE先压缩再扩散,把主要计算从高维像素空间迁移到低维潜空间,在保持质量的同时显著降低训练与采样成本[18]。(2)求解器路线以去噪扩散隐式模型及后续快速求解器为代表:在不改变训练目标的前提下,通过隐式或高阶离散化策略减少采样步数,直接改善推理效率[16,25-26]。(3)路径学习路线以修正流为代表:通过学习更直接的概率传输向量场,使少步采样在质量上更接近多步扩散模型[8,27]。(4)引导与对齐路线通过无分类器引导等技术把条件语义显式注入采样轨迹,提升文本-图像一致性与可控性,已成为大规模条件生成系统的标准组件[16,28]。

2 数据概率建模的分数匹配方法

2.1 分数匹配的解释和推广

分数匹配的数据概率建模方法最早在文献[2]中提出,主要规避最大似然中概率常数难以计算的问题。文献[29]对分数匹配进行了进一步解释,揭示了最大似然估计和分数匹配之间的深刻联系,分析表明分数匹配对噪声训练数据具有更强的鲁棒性,提出了一种分数匹配的推广,进一步将分数匹配推广到离散数据。文献[29]将d维数据$\boldsymbol{x}$的概率分布记为$p(\boldsymbol{x})$,将模型的概率分布记为$q_\theta(\boldsymbol{x})$,指出最大似然估计等价于最小化KL散度。这是因为:最大似然估计就是关于参数$\boldsymbol{\theta}$最大化对数似然函数$\int_x p(\boldsymbol{x})\ln q_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$,这等价于最小化它们之间的KL散度,即

$$\mathrm{KL}(p\|q_\theta)=\int_x p(\boldsymbol{x})\ln\frac{p(\boldsymbol{x})}{q_\theta(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}=\int_x p(\boldsymbol{x})\ln p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_x p(\boldsymbol{x})\ln q_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{23}$$

式(23)展开项中第1项与$\boldsymbol{\theta}$无关,第2项即为对数似然函数,这一结论也可从式(10)中得出。文献[29]的另一个结论是:分数匹配等价于最小化Fisher散度。Fisher散度定义为

$$D_\mathrm{F}(p\|q_\theta)=\int_x p(\boldsymbol{x})\left(\frac{\nabla_x p(\boldsymbol{x})}{p(\boldsymbol{x})}-\frac{\nabla_x q_\theta(\boldsymbol{x})}{q_\theta(\boldsymbol{x})}\right)^2\mathrm{d}\boldsymbol{x} \tag{24}$$

由于分数可定义为概率密度函数的对数关于数据的梯度$\nabla_x\ln p(\boldsymbol{x})$,因此Fisher散度即为两个概率密度函数的分数差的平方关于第1个分布的期望。Fisher散度还可以写成另外一种形式,即

$$D_\mathrm{F}(p\|q_\theta)=\int_x p(\boldsymbol{x})\left(\nabla_x\ln\frac{p(\boldsymbol{x})}{q_\theta(\boldsymbol{x})}\right)^2\mathrm{d}\boldsymbol{x} \tag{25}$$

式(25)与KL散度很类似,预示最大似然估计和分数匹配之间应该有比较紧密的关系。文献[29]给出了如下定理。

定理1[29] 设$\boldsymbol{y}=\boldsymbol{x}+\sqrt{t}\,\boldsymbol{w}$,$t\geqslant 0$,$\boldsymbol{w}$为零均值高斯白向量,当$\boldsymbol{x}$的分布分别是$p(\boldsymbol{x})$和$q(\boldsymbol{x})$时,$\boldsymbol{y}$的分布分别记为$\tilde{p}_t(\boldsymbol{y})$和$\tilde{q}_t(\boldsymbol{y})$,则当$\tilde{p}_t(\boldsymbol{y})$和$\tilde{q}_t(\boldsymbol{y})$是光滑且下降足够快时,有$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\tilde{p}_t(\boldsymbol{y})\|\tilde{q}_t(\boldsymbol{y}))=-\frac{1}{2}D_F(\tilde{p}_t(\boldsymbol{y})\|\tilde{q}_t(\boldsymbol{y}))$。由于$\tilde{p}_0(\boldsymbol{y})=p(\boldsymbol{x})$,$\tilde{q}_0(\boldsymbol{y})=q(\boldsymbol{x})$,进一步有$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\tilde{p}_t(\boldsymbol{y})\|\tilde{q}_t(\boldsymbol{y}))|_{t=0}=-\frac{1}{2}D_F(p(\boldsymbol{x})\|q(\boldsymbol{x}))$。

定理1说明在尺度因子t时两个概率密度的Fisher散度等于KL散度关于t的导数的负,由于Fisher散度总是非负的,因此KL散度不会随着尺度因子t的增加而增加。这是因为随着更强噪声的加入,信号的分布更接近于噪声,其分布也变得更相似。从定理1可以看出,最大似然估计旨在最小化KL散度,分数匹配寻求消除$t=0$尺度空间KL散度的导数。换言之,分数匹配力求稳定性,当训练数据中有噪声时,最优参数$\boldsymbol{\theta}$使得两个模型的KL散度变化最小,而最大似然估计追求KL散度的极端性(最小性)。最大似然方法对噪声数据敏感,分数匹配可能对小幅度的噪声更鲁棒。正因为如此,最大似然和分数匹配可能导致完全不同的解。文献[29]还说明,KL散度是由熵导出,而Fisher散度是由Fisher信息导出,熵关于尺度因子t的导数等于Fisher信息,其几何解释是熵与体积相关,而Fisher信息与表面积相关。

梯度算子是一个线性算子,当把Fisher散度中的梯度算子替换为一般的线性算子L时,得到广义Fisher散度D_L[29]为

$$D_L(p\|q_\theta)=\int_{\boldsymbol{x}}p(\boldsymbol{x})\left(\frac{L[p(\boldsymbol{x})]}{p(\boldsymbol{x})}-\frac{L[q_\theta(\boldsymbol{x})]}{q_\theta(\boldsymbol{x})}\right)^2\mathrm{d}\boldsymbol{x}\tag{26}$$

当以广义Fisher散度代替Fisher散度时,得到广义分数匹配方法。可以证明广义分数匹配也具有类似于分数匹配的性质,包括其展开后仅依赖于$q_\theta(\boldsymbol{x})$的线性算子和L的伴随算子。在广义分数匹配的框架下,文献[29]提出了另外一个线性算子:边缘化算子M。对1个概率密度函数$p(\boldsymbol{x})$,$M[p(\boldsymbol{x})]$为1个d维向量函数,其第i个分量为$\boldsymbol{x}$积分掉第i个分量后剩下的$d-1$个分量的边缘分布。基于这个线性算子,文献[29]提出了一种针对离散数据的分数匹配算法。

2.2 分数匹配与去噪编码器的联系

分数匹配与去噪自编码器(Denoising autoencoder,DAE)也有紧密联系[30]。DAE在训练数据$\boldsymbol{x}$中加入零均值协方差矩阵为$\sigma^2\boldsymbol{I}$的高斯噪声,得到含噪数据$\tilde{\boldsymbol{x}}=\boldsymbol{x}+\epsilon$,$\epsilon\sim N(0,\sigma^2\boldsymbol{I})$。这对应着条件概率$q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}\sigma^d}\mathrm{e}^{-\frac{1}{\sigma^2}\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|^2}$。噪声污染的信号$\tilde{\boldsymbol{x}}$经过一层仿射变换,接着用Sigmoid非线性变换进行编码,得到隐表示$\boldsymbol{h}=\mathrm{encode}(\tilde{\boldsymbol{x}})=\mathrm{sigmoid}(\boldsymbol{W}\tilde{\boldsymbol{x}}+\boldsymbol{b})$,其中矩阵、向量均有合适的维数。隐表示$\boldsymbol{h}$经过相同权重参数的仿射变换解码出1个重构数据$\boldsymbol{x}^{\mathrm{r}}=\mathrm{decode}(\boldsymbol{h})=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}+\boldsymbol{c}$。去噪自编码器DAE关于参数$\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}\}$最小化重构误差,即

$$J_{\mathrm{DAE}\sigma}(\theta)=E_{q_\sigma(\boldsymbol{x},\tilde{\boldsymbol{x}})}\left[\|\mathrm{decode}(\mathrm{encode}(\tilde{\boldsymbol{x}}))-\boldsymbol{x}\|^2\right]=E_{q_\sigma(\boldsymbol{x},\tilde{\boldsymbol{x}})}[\|\boldsymbol{W}^{\mathrm{T}}\mathrm{sigmoid}(\boldsymbol{W}\tilde{\boldsymbol{x}}+\boldsymbol{b})+\boldsymbol{c}-\boldsymbol{x}\|^2]\tag{27}$$

式中:$q_\sigma(\boldsymbol{x},\tilde{\boldsymbol{x}})=q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})q_0(\boldsymbol{x})$为联合概率分布,$q_0(\boldsymbol{x})=\frac{1}{N}\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n)$表示观察数据集的经验概率密度函数。

分数匹配的原始目标函数(6)在文献[30]中定义为$J_{\mathrm{ESM}q}(\boldsymbol{\theta})$,其中ESM表示显式分数匹配,$q$表示数据未知的真实分布。由于数据的真实分布一般是未知的,如果使用加噪数据的Parzen窗密度估计器

$q_\sigma(\tilde{\boldsymbol{x}})$作为匹配目标,则分数匹配为

$$J_{\mathrm{ESM}q_\sigma}(\boldsymbol{\theta})=E_{q_\sigma(\tilde{\boldsymbol{x}})}\left[\frac{1}{2}\left\|\boldsymbol{s}(\tilde{\boldsymbol{x}},\boldsymbol{\theta})-\nabla_{\tilde{\boldsymbol{x}}}\ln q_\sigma(\tilde{\boldsymbol{x}})\right\|^2\right]$$

对干净数据和加噪数据对$(\boldsymbol{x},\tilde{\boldsymbol{x}})$,文献[31]定义了一个去噪分数匹配(Denoising score matching, DSM)为

$$J_{\mathrm{DSM}q_\sigma}(\boldsymbol{\theta})=E_{q_\sigma(\boldsymbol{x},\tilde{\boldsymbol{x}})}\left[\frac{1}{2}\left\|\boldsymbol{s}(\tilde{\boldsymbol{x}},\boldsymbol{\theta})-\nabla_{\tilde{\boldsymbol{x}}}\ln q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})\right\|^2\right] \tag{28}$$

其基本思想是沿着加噪数据的对数密度梯度方向,理想情况下会移动到干净数据,且有

$$\nabla_{\tilde{\boldsymbol{x}}}\ln q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})=\frac{1}{\sigma^2}(\boldsymbol{x}-\tilde{\boldsymbol{x}}) \tag{29}$$

式(29)右边的方向正是从加噪数据移动回干净数据的方向,目标就是让模型的分数尽量匹配这个方向。文献[31]证明了$J_{\mathrm{ESM}q_\sigma}(\boldsymbol{\theta})$等价于$J_{\mathrm{DSM}q_\sigma}(\boldsymbol{\theta})$,且结论不依赖于概率分布$q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})$和$q(\boldsymbol{x})$的具体形式,只要条件概率$q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x})$可微。

如果定义模型的概率密度函数如下[31]

$$p(\boldsymbol{x},\boldsymbol{\theta})=\frac{1}{Z(\boldsymbol{\theta})}\exp(-E(\boldsymbol{x},\boldsymbol{\theta})) \tag{30}$$

式中能量函数为

$$E(\boldsymbol{x},\boldsymbol{\theta})=E(\boldsymbol{x};\boldsymbol{W},\boldsymbol{b},\boldsymbol{c})=-\frac{\langle\boldsymbol{c},\boldsymbol{x}\rangle-\frac{1}{2}\|\boldsymbol{x}\|^2+\sum_{j=1}^{d_h}\mathrm{softplus}\left(\langle\boldsymbol{W}_j,\boldsymbol{x}\rangle+\boldsymbol{b}_j\right)}{\sigma^2} \tag{31}$$

式中$\boldsymbol{W}_j$为矩阵$\boldsymbol{W}$的第j行分量。文献[30]证明了DSM等价于DAE,或者说去噪自编码器是去噪分数匹配的一种特例。最后的结论是:训练一个去噪自编码器等价于能量函数(31)与Parzen窗密度估计的分数匹配。

2.3 基于朗之万动力学的分数匹配

文献[32]针对数据一般位于高位空间的低维流形,在数据密度稀疏的地方分数估计不准确的问题,提出:在数据中添加噪声进行扰动,从而改善分数估计的准确性;利用神经网络学习不同噪声等级数据的分数,并利用朗之万动力学生成新数据。文献[2]中提出的分数匹配主要用于概率建模,用一个神经网络$s_\theta(\boldsymbol{x})$来估计分数,当进行隐式分数匹配时,需要计算其雅可比矩阵的迹$\mathrm{tr}(\nabla_x s_\theta(\boldsymbol{x}))$,在高维大规模情况下会带来计算困难。文献[33]提出了切片分数匹配方法,用分数网络的雅可比矩阵的随机投影近似迹的计算。

数据生成就是进行采样,基于朗之万动力学的采样是一种经典的基于分数的采样方法[34]。对于概率密度函数$p(\boldsymbol{x})$,给定一个固定步长$\epsilon>0$和初始值$\boldsymbol{x}_0\sim\pi(\boldsymbol{x})$,其中$\pi(\boldsymbol{x})$是某个先验分布,朗之万采样方法表达式如下

$$\boldsymbol{x}_t=\boldsymbol{x}_{t-1}+\frac{\epsilon}{2}\nabla_x\ln p(\boldsymbol{x})+\sqrt{\epsilon}\,\boldsymbol{z}_t\quad t=1,2,\cdots,T \tag{32}$$

式中$\boldsymbol{z}_t\sim N(0,\boldsymbol{I})$。当$\epsilon\to 0,T\to\infty$时,$\boldsymbol{x}_T$的分布趋近于$p(\boldsymbol{x})$,$\boldsymbol{x}_T$就成为$p(\boldsymbol{x})$的一个样本[14,24]。因此,如果训练了1个分数网络$s_\theta(\boldsymbol{x})$,用其代替式(32)中的分数实现基于分数的生成建模。

设$\sigma_i(i=1,2,\cdots,L)$为递减几何数列,满足$\frac{\sigma_1}{\sigma_2}=\frac{\sigma_2}{\sigma_3}=\cdots=\frac{\sigma_{L-1}}{\sigma_L}>0$,加入高斯噪声后的数据分布为$q_\sigma(\boldsymbol{x})=\int p_{\mathrm{Data}}(\boldsymbol{y})N(y;0,\sigma^2\boldsymbol{I})\mathrm{d}\boldsymbol{y}$,文献[32]训练了一个称为噪声条件分数网络(Noise conditional

score network，NCSN）的神经网络模型，对所有噪声尺度估计加噪数据的分数：$s_\theta(\boldsymbol{x},\sigma)\approx\nabla_{\boldsymbol{x}}\ln q_\sigma(\boldsymbol{x})$。对每个尺度$\sigma_i$采用类似于式（28）的去噪分数匹配的损失函数$J_{\mathrm{DSM}q_{\sigma_i}}(\boldsymbol{\theta})$，将所有损失函数的加权和作为总的损失函数

$$J(\boldsymbol{\theta};\sigma_1,\sigma_2,\cdots,\sigma_L)=\frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)J_{\mathrm{DSM}q_{\sigma_i}}(\boldsymbol{\theta}) \tag{33}$$

文献[32]选择权重$\lambda(\sigma_i)=\sigma_i^2$，基于训练NCSN得到的分数估计$s_\theta(\boldsymbol{x},\sigma)$，代入式（32）生成期望分布的数据；在生成过程中提出了一种退火朗之万动力学采样，即步长ϵ按照一定的规律减小；最后与GAN等已有的生成方法进行比较，结果显示提出的基于分数的方法生成效果与SOTA方法相当。

2.4 去噪扩散概率模型

基于DPM[24]，文献[6]提出去噪扩散概率模型（Denoising diffusion probabilistic models，DDPM）。DDPM首次被应用于生成高质量图像，其图模型如图3所示[6]。

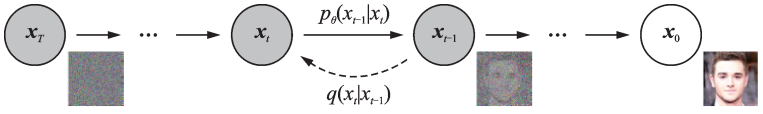


图3 DDPM有向图模型[6]

Fig.3 Directed graphical model of DDPM[6]

与DPM一样，DDPM的前向过程固定为1个马尔可夫链，不断地在图像中加入噪声，前向过程的后验概率为

$$q(\boldsymbol{x}_{1,2,\cdots,T}|\boldsymbol{x}_0)=\prod_{t=1}^{T}q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}),\ \ q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=N(\boldsymbol{x}_t;\sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1},\beta_t\boldsymbol{I}) \tag{34}$$

反向过程也是1个马尔科夫链，即

$$\begin{cases}p(\boldsymbol{x}_T)=N(\boldsymbol{x}_T;0,\boldsymbol{I}),p_\theta(\boldsymbol{x}_0)=\int p_\theta(\boldsymbol{x}_{0:T})\mathrm{d}\boldsymbol{x}_{1:T}\\ p_\theta(\boldsymbol{x}_{0:T})=p(\boldsymbol{x}_T)\prod_{t=1}^{T}p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t),\ \ p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)=N(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t,t))\end{cases} \tag{35}$$

训练过程是关于$\boldsymbol{\theta}$最小化负对数似然的变分界，即

$$E[-\ln p_\theta(\boldsymbol{x}_0)]\leqslant E_q\left[-\ln\frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]=E_q\left[-\ln p(\boldsymbol{x}_T)-\sum_{t\geqslant 1}\ln\frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]:=L \tag{36}$$

式（36）中第1个不等式后面的项就是式（11）中下界的负。因为这里$\boldsymbol{x}_0$为观测变量，$\boldsymbol{x}_{1:T}$为隐变量，$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$为隐变量后验概率的近似概率，$p_\theta(\boldsymbol{x}_{0:T})$为完整数据的模型概率，因此DDPM也是一个完全的最大似然估计框架。可以证明，前向过程的任意步t也服从高斯分布，即

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0)=N(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0,(1-\bar{\alpha}_t)\boldsymbol{I}) \tag{37}$$

式中：$\alpha_t=1-\beta_t$，$\bar{\alpha}_t=\prod_{s=1}^{t}\alpha_s$。

文献[6]进一步把变分界L分解成如下形式

$$E_q\left[\underbrace{\mathrm{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))}_{L_T}-\sum_{t>1}\underbrace{\mathrm{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))}_{L_{t-1}}-\underbrace{\ln p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}_{L_0}\right] \tag{38}$$

式中 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ 也是高斯分布,有 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)=N(\boldsymbol{x}_{t-1};\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0),\tilde{\beta}_t\boldsymbol{I})$,均值 $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0+\frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t,\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$。因此,式(38)中的所有KL散度都是高斯分布之间的散度,有闭式计算公式。

文献[6]推导出了扩散模型和DSM之间的联系。假设前向过程的参数固定,因此式(38)中的 L_T 为常数,与参数 $\boldsymbol{\theta}$ 无关。对式(38)的中间项,文献[6]证明了反向生成过程可用如下公式计算

$$\boldsymbol{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\right)+\sigma_t\boldsymbol{z}\quad\boldsymbol{z}\sim\boldsymbol{\mathcal{N}}(0,\boldsymbol{I}) \tag{39}$$

式(39)与朗之万动力学采样(式(32))很类似,$-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)$ 相当于学习的分数。目标函数 L_{t-1} 可简化为

$$E_{\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0+(1-\bar{\alpha}_t)\boldsymbol{\epsilon},t\right)\right\|^2\right] \tag{40}$$

式(40)目标函数类似于DSM的目标函数[30],这也说明了式(38)中的 L_{t-1} 等价于朗之万动力学中反向过程中的变分界[6],优化DSM等价于优化朗之万动力学采样链的变分推理。

前文所述DDPM是一个最大似然估计框架,这里又说明了DDPM等价于去噪分数匹配,但一般情况下最大似然估计与分数匹配不同。一种猜想是如文献[2]所述,对于高斯分布,最大似然估计与分数匹配的结果相同,而DDPM目标函数(38)中所有 L_{t-1} 项都是两个高斯分布之间的KL散度,KL散度最小化即是最大似然估计,而高斯之间的最大似然估计与分数匹配等价。式(38)中的最后一项解释为1个图像解码器,通过最后一步生成图像。

DDPM最终采用的是一个简化的目标函数,即

$$L_{\text{simple}}(\theta)=E_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0+(1-\bar{\alpha}_t)\boldsymbol{\epsilon},t\right)\right\|^2\right] \tag{41}$$

2.5 基于随机微分方程的分数生成模型

文献[11]提出用随机微分方程(Stochastic differential equation, SDE)通过注入噪声将复杂数据分布变换为已知先验分布(高斯分布),然后通过相应的反时SDE慢慢去除噪声将先验分布变换回数据分布的方法。反时SDE只依赖于扰动数据的分数,通过神经网络估计分数和SDE的数值积分生成数据,提出的框架统一了基于分数的生成建模和DDPM。在数据中加入不同尺度的噪声以扰动数据是去噪自编码器[30]、去噪分数匹配[30-31]、SMLD[32]、DDPM[6]等方法成功的关键理论。当噪声尺度变化连续且趋近于无穷小时,这个过程可以用SDE描述,即

$$\mathrm{d}\boldsymbol{x}=f(\boldsymbol{x},t)\mathrm{d}t+g(t)\mathrm{d}\boldsymbol{w} \tag{42}$$

式中:$\boldsymbol{w}$ 为标准维纳过程,$f(\cdot,t):\mathbf{R}^d\to\mathbf{R}^d$ 称为 $\boldsymbol{x}(t)$ 的漂移系数,$g(\cdot):\mathbf{R}\to\mathbf{R}$ 称为 $\boldsymbol{x}(t)$ 的扩散系数。下面用 $p_t(\boldsymbol{x})$ 表示 $\boldsymbol{x}(t)$ 的概率密度,用 $p_{st}(\boldsymbol{x}(t)|\boldsymbol{x}(s))$ 表示从 $\boldsymbol{x}(s)$ 到 $\boldsymbol{x}(t)$ 的转移核($0\leqslant s<t\leqslant T$)。式(42)描述的扩散过程初始数据为 $\boldsymbol{x}(0)$,其分布为 p_0,终态分布 p_T 一般为高斯分布。

反时SDE由如下方程描述[11,35]

$$\mathrm{d}\boldsymbol{x}=[f(\boldsymbol{x},t)-g(t)^2\nabla_x\ln p_t(\boldsymbol{x})]\mathrm{d}t+g(t)\mathrm{d}\bar{\boldsymbol{w}} \tag{43}$$

式中 $\bar{\boldsymbol{w}}$ 是时间从 T 到0反向的标准维纳过程,如果任意时刻 t 的分数已知,则可以通过数值求解算法生成满足分布 p_0 的样本。这个过程如图4所示[11]。

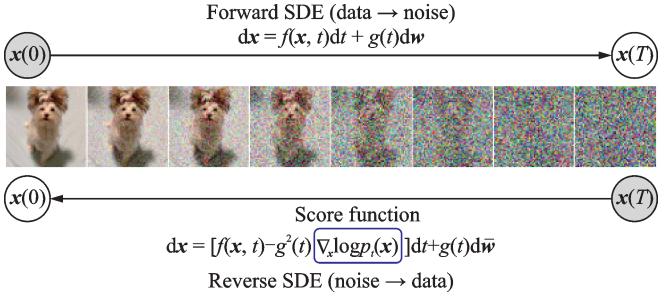


图4 求解反向SDE产生基于分数的生成模型[11]

Fig.4 Score-based generative modeling by solving the reverse-time SDE[11]

类似于NCSN[32],用1个神经网络来学习依赖于时间的分数网络$s_\theta(\boldsymbol{x},t)$,其目标函数为

$$E_t\left\{\lambda(t)E_{\boldsymbol{x}(0)}E_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\left[\left\|s_\theta(\boldsymbol{x}(t),t)-\nabla_{\boldsymbol{x}(t)}\ln p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))\right\|^2\right]\right\} \tag{44}$$

式中:t从$[0,T]$均匀采样,$\lambda(t)$为权值函数。

SMLD[32]和DDPM[6]可以看成两个不同SDE的离散化。SMLD[32]对应着如下马尔可夫链的分布

$$\boldsymbol{x}_i=\boldsymbol{x}_{i-1}+\sqrt{\sigma_i^2-\sigma_{i-1}^2}\,\boldsymbol{z}_{i-1}\quad i=1,2,\cdots,N;\ \boldsymbol{z}_{i-1}\sim\mathcal{N}(0,\boldsymbol{I}) \tag{45}$$

式中:$\sigma_0=0$,当$N\to\infty$,离散序列$\{\sigma_i\}$变成函数$\sigma(t)$,马尔科夫链$\{\boldsymbol{x}_i\}$变成由如下SDE描述的随机过程$\{\boldsymbol{x}(t)\}_{t=0}^1$,即

$$\mathrm{d}\boldsymbol{x}=\sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\,\mathrm{d}\boldsymbol{w} \tag{46}$$

类似地,DDPM[6]的马尔科夫链为

$$\boldsymbol{x}_i=\sqrt{1-\beta_i}\,\boldsymbol{x}_{i-1}+\sqrt{\beta_i}\,\boldsymbol{z}_{i-1} \tag{47}$$

其SDE为

$$\mathrm{d}\boldsymbol{x}=-\frac{1}{2}\beta(t)\boldsymbol{x}\mathrm{d}t+\sqrt{\beta(t)}\,\mathrm{d}\boldsymbol{w} \tag{48}$$

式(46)称为方差爆炸SDE,而式(48)称为方差保持SDE。当训练了分数网络s_θ,可用其代替反时SDE(式(43))中的分数函数,然后用SDE数值求解器生成符合p_0分布的数据。

2.6 分数匹配的优劣势与发展

从概率建模角度看,分数匹配的突出优势在于训练目标不依赖配分函数,天然适合非归一化密度建模;同时其目标可解释为Fisher散度最小化,在含噪数据条件下通常比纯最大似然更稳定,这也是后续去噪分数匹配与扩散训练目标能够统一的重要基础[2,16,29]。但是,其主要局限在于原始形式涉及高阶导数或雅可比迹项估计,高维数据下计算与方差控制比较困难,对网络参数化与噪声尺度设计也比较敏感[16,30,33]。近年来该方向已从经典分数匹配扩展到去噪分数匹配、多噪声层级学习与连续时间Score-SDE框架,实现了从静态密度估计到动态生成过程建模的过渡,并显著推动了高保真生成能力[6,11,16,30]。

3 基于流的数据概率建模方法

3.1 归一化流

归一化流本质上是一种最大似然的变分推理方法,最早在文献[7]中被提出,主要解决变分推理中

隐变量的后验概率选择的问题。如前文所述,概率建模总是在可处理性和灵活性之间权衡,平均场变分推理假设后验概率是可分解概率,限制过强,导致在解决实际复杂分布时变分推理的性能不佳。归一化流方法通过归一化流构造的分布来近似后验概率分布,简单的初始分布通过一系列可逆变换转化为期望分布。

考虑一个可逆光滑映射$f:\mathbf{R}^d\to\mathbf{R}^d$,其逆映射为$g=f^{-1}$。如果用这个映射将分布为$q(\boldsymbol{z})$的随机变量$\boldsymbol{z}$变换为随机变量$\boldsymbol{z}'$:$\boldsymbol{z}'=f(\boldsymbol{z})$,则$\boldsymbol{z}'$的分布为

$$q(\boldsymbol{z}')=q(\boldsymbol{z})\left|\det\frac{\partial f^{-1}}{\partial \boldsymbol{z}'}\right|=q(\boldsymbol{z})\left|\det\frac{\partial f}{\partial \boldsymbol{z}}\right|^{-1} \tag{49}$$

可以通过多个简单映射的复合反复地应用式(49)从简单分布出发得到任意复杂的分布。具有分布q_0的随机变量$\boldsymbol{z}_0$经过K个变换序列f_k的分布$q_K(\boldsymbol{z})$为

$$\boldsymbol{z}_K=f_K\circ\cdots\circ f_1(\boldsymbol{z}) \tag{50}$$

$$\ln q_K(\boldsymbol{z}_K)=\log q_0(\boldsymbol{z}_0)-\ln\left|\det\frac{\partial f_k}{\partial \boldsymbol{z}_{k-1}}\right| \tag{51}$$

初始分布为$q_0(\boldsymbol{z}_0)$随机变量序列$\boldsymbol{z}_k=f_k(\boldsymbol{z}_{k-1})$遍历的路径称为流,由分布$q_k$构成的路径称为归一化流。通过如上述一系列的变量变换,初始概率密度在可逆映射的作用下“流动”,最终得到期望的分布。式(51)描述的流是有限离散流。如果流的长度趋于无穷,则称为无穷小流[7],它由偏微分方程$\frac{\partial}{\partial t}q_t(\boldsymbol{z})=T_t[q_t(\boldsymbol{z})]$描述,刻画了初始概率密度$q_0(\boldsymbol{z})$如何随时间演化。文献[7]给出了两个无穷小流的例子:朗之万流和哈密顿流。

归一化流推理中的困难是计算雅可比矩阵的行列式,其计算复杂性是隐藏层维数的3次方。为简化计算,文献[7]提出了可逆线性变换流,表达式为

$$f(\boldsymbol{z})=\boldsymbol{z}+\boldsymbol{u}h(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}+b) \tag{52}$$

式中:$\boldsymbol{u}$、$\boldsymbol{w}$、b为合适维数的向量或数量参数,$h(\cdot)$为1个非线性函数,其导数为$h'(\cdot)$。式(52)的对数行列式可以线性时间复杂度计算为

$$\begin{cases}\psi(\boldsymbol{z})=h'(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}+b)\boldsymbol{w}\\ \left|\det\dfrac{\partial f}{\partial \boldsymbol{z}}\right|=\left|\det(\boldsymbol{I}+\boldsymbol{u}\psi(\boldsymbol{z})^{\mathrm{T}})\right|=|1+\boldsymbol{u}^{\mathrm{T}}\psi(\boldsymbol{z})|\\ \ln q_K(\boldsymbol{z}_K)=\ln q_0(\boldsymbol{z})-\log|1+\boldsymbol{u}_k^{\mathrm{T}}\psi(\boldsymbol{z}_{k-1})|\end{cases} \tag{53}$$

式(53)定义的流在与平面$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}+b=0$垂直的方向压缩或扩展初始概率$q_0(\boldsymbol{z})$,因此称为平面流。

归一化流的训练基于最大似然估计。最大化似然即最小化负对数似然,负对数似然的上界在文献[7]中称为自由能量$F(\boldsymbol{x})$,也即ELBO的负。对于归一化流,用流的最终分布$q_K(\boldsymbol{z}_K)$作为隐变量后验分布的近似:$q_\phi(\boldsymbol{z}|\boldsymbol{x})=q_K(\boldsymbol{z}_K)$,自由能量可以计算为关于初始分布$q_0(\boldsymbol{z})$的期望,即

$$\begin{aligned}F(\boldsymbol{x})&=E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\ln q_\phi(\boldsymbol{z}|\boldsymbol{x})-\ln p(\boldsymbol{x},\boldsymbol{z})]=E_{q_0(\boldsymbol{z}_0)}[\ln q_K(\boldsymbol{z}_K)-\ln p(\boldsymbol{x},\boldsymbol{z}_K)]=\\ &E_{q_0(\boldsymbol{z}_0)}[\ln q_0(\boldsymbol{z}_0)]-E_{q_0(\boldsymbol{z}_0)}[\ln p(\boldsymbol{x},\boldsymbol{z}_K)]-E_{q_0(\boldsymbol{z}_0)}\left[\sum_{k=1}^{K}\ln|1+\boldsymbol{u}_k^{\mathrm{T}}\psi(\boldsymbol{z}_{k-1})|\right]\end{aligned} \tag{54}$$

最小化如上目标函数,可以通过任何变分优化方法,包括广义EM算法。对于摊销变分推理,文献[7]构建了一个深度神经网络推理模型,将观察数据映射为初始分布的参数和归一化流的参数。

3.2 连续归一化流

文献[36]提出了连续归一化流(Continuous normalizing flow, CNF)。由于离散归一化流的主要计算瓶颈是变换映射雅可比矩阵的行列式,平面流将映射函数限定为线性函数,简化了计算但也限制了归一化流的表达能力,因此后来也有一些工作研究归一化流的表达能力和计算代价的折中[36]。文献[36]证明,当把离散归一化流的离散步长无限变小,表示归一化流的差分方程变为微分方程,此时变量替换的归一化常数变化的计算反而简单。

定理2(瞬时变量替换定理)[36] 假设$\boldsymbol{z}(t)$为一个有限连续随机变量,其概率分布$p(\boldsymbol{z}(t))$依赖于时间t。假设$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t}=f(\boldsymbol{z}(t),t)$是描述$\boldsymbol{z}(t)$随时间变化的微分方程,且$f$关于$\boldsymbol{z}$一致Lipschitz连续,关于$t$连续,则对数概率的变化满足如下微分方程

$$\frac{\partial \ln p(\boldsymbol{z}(t))}{\partial t}=-\mathrm{tr}\left(\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{z}(t)}\right) \tag{55}$$

不像离散情况下需要计算行列式,这里只需要计算迹。另外,微分方程f不必要求是双射。考虑连续平面流

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t}=\boldsymbol{u}h(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}(t)+b),\quad \frac{\partial \ln p(\boldsymbol{z}(t))}{\partial t}=-\boldsymbol{u}^{\mathrm{T}}\frac{\partial h}{\partial \boldsymbol{z}(t)} \tag{56}$$

可以通过解这个微分方程组来计算概率密度。

求矩阵的行列式不是线性算子,而求迹是线性算子,因此如果动态系统可以表示为函数的和,则对数密度也可以表示为函数的和,即

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t}=\sum_{n=1}^{M}f_n(\boldsymbol{z}(t)),\quad \frac{\mathrm{d}\ln p(\boldsymbol{z}(t))}{\mathrm{d}t}=-\mathrm{tr}\left(\frac{\mathrm{d}f_n}{\mathrm{d}\boldsymbol{z}}\right) \tag{57}$$

式(57)表明可以计算由很多隐节点的流模型,而复杂度关于节点个数M是线性的。还可以引入门控神经网络$\sigma_n(t)\in(0,1)$来控制流模型的隐节点:$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t}=\sum_{n=1}^{M}\sigma_n(t)f_n(\boldsymbol{z})$。这些模型都称为连续归一化流。文献[36]用最大似然估计方法训练CNF来进行密度估计。

3.3 流匹配

流匹配[8]是一种基于CNF的生成模型,与将噪声变换为数据的一般高斯路径兼容,包含常见的扩散路径。用流匹配训练扩散模型更鲁棒和稳定,而用流匹配训练CNF可以生成比扩散模型方法更好的效果。

1个概率密度路径$p:[0,1]\times\mathbf{R}^d\to\mathbf{R}_{>0}$是1个依赖于时间的概率密度函数,1个流$\boldsymbol{\phi}:[0,1]\times\mathbf{R}^d\to\mathbf{R}^d$由1个依赖于时间的向量场$\boldsymbol{v}:[0,1]\times\mathbf{R}^d\to\mathbf{R}^d$通过微分方程定义

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_t(\boldsymbol{x})=\boldsymbol{v}_t(\boldsymbol{\phi}_t(\boldsymbol{x})),\quad \boldsymbol{\phi}_0(\boldsymbol{x})=\boldsymbol{x} \tag{58}$$

式中$\boldsymbol{\phi}_t(\boldsymbol{x})$就是一个CNF,其将1个简单的初始分布$p_0$通过如下的前推方程变换为1个复杂的分布$p_1$

$$p_t=[\boldsymbol{\phi}_t]_*p_0 \tag{59}$$

前向过程算子$[\cdot]_*$,即变量替换,定义为

$$[\boldsymbol{\phi}_t]_*p_0(\boldsymbol{x})=p_0(\boldsymbol{\phi}_t^{-1}(\boldsymbol{x}))\left[\det\frac{\partial\boldsymbol{\phi}_t^{-1}(\boldsymbol{x})}{\partial\boldsymbol{x}}\right] \tag{60}$$

称向量场$\boldsymbol{v}_t$生成了概率密度路径p_t,如果它的流$\boldsymbol{\phi}_t$满足式(59)。假设$\boldsymbol{x}_1$为一个随机变量,其分布

$q(\boldsymbol{x}_1)$未知,假设p_t为一个概率路径,$p_0=p$为一个简单初始分布,典型情况为标准高斯分布,p_1近似等于$q(\boldsymbol{x}_1)$。给定一个概率密度路径$p_t(\boldsymbol{x})$及其相应的向量场$u_t(\boldsymbol{x})$,即$u_t(\boldsymbol{x})$生成了$p_t(\boldsymbol{x})$。流匹配(Flow matching, FM)的目标函数定义如下

$$L_{\mathrm{FM}}(\boldsymbol{\theta})=E_{t,\boldsymbol{x}\sim p_t(\boldsymbol{x})}\left[\left\|v_t(\boldsymbol{x};\boldsymbol{\theta})-u_t(\boldsymbol{x})\right\|^2\right] \tag{61}$$

式中:CNF向量场$v_t(\boldsymbol{x};\boldsymbol{\theta})$可用1个神经网络来近似[36],$\boldsymbol{\theta}$表示网络的可学习参数,$t\sim U[0,1]$。式(61)即用神经网络$v_t$回归目标向量场$u_t(\boldsymbol{x})$,收敛后,学习的CNF模型即可生成$p_t(\boldsymbol{x})$。

但是流匹配模型是不可解的,因为实际应用中不知道p_t和u_t。文献[8]提出利用在观察数据点上的条件概率路径简化流匹配问题。假设有一个观察数据点$\boldsymbol{x}_1$,其定义了条件概率路径$p_t(\boldsymbol{x}|\boldsymbol{x}_1)$和条件向量场$u_t(\cdot|\boldsymbol{x}_1)$,将难于处理的边缘向量场转化为易于处理的条件向量场,因此不可处理的流匹配(式(61))可以转化为以下条件流匹配(Conditional flow matching, CFM)形式

$$L_{\mathrm{CFM}}(\boldsymbol{\theta})=E_{t,q(\boldsymbol{x}_1),p_t(\boldsymbol{x}|\boldsymbol{x}_1)}\left[\left\|v_t(\boldsymbol{x};\boldsymbol{\theta})-u_t(\boldsymbol{x}|\boldsymbol{x}_1)\right\|^2\right] \tag{62}$$

不同于式(61),只要能够高效地采样条件概率$p_t(\boldsymbol{x}|\boldsymbol{x}_1)$并计算$u_t(\boldsymbol{x}|\boldsymbol{x}_1)$,式(62)就易于估计。而且可以证明FM和CFM有相同的梯度,因此优化式(61)等价于优化式(62)[8]。

文献[8]考虑一个特例:高斯条件概率路径和条件流,表达式分别为

$$p_t(\boldsymbol{x}|\boldsymbol{x}_1)=N(\boldsymbol{x}|\mu_t(\boldsymbol{x}_1),\sigma_t(\boldsymbol{x}_1)^2\boldsymbol{I}) \tag{63}$$

$$\psi_t(\boldsymbol{x})=\sigma_t(\boldsymbol{x}_1)x+\mu_t(\boldsymbol{x}_1) \tag{64}$$

文献[8]说明了方差爆炸扩散模型和方差保持扩散模型都是高斯条件路径的特例。通过实验比较了DDPM、分数匹配方法与流匹配方法,表明流匹配方法取得了更好的效果。

流匹配将生成问题表述为沿概率路径学习速度场。相较以随机扰动为核心的SDE路线,这类方法通常对应直接的向量场回归目标,优化更稳定、采样轨迹更可控,并在低函数评估次数(Number of function evaluation, NFE)场景下展现出明显潜力[8,16,26-27]。其不足主要体现在路径设计依赖较强:不同插值路径与条件策略会显著影响传输几何、收敛速度和最终质量;在复杂多模态数据上,如何在少步采样下兼顾分布覆盖与细节保真仍是开放问题[8,16,37]。从发展趋势看,Rectified Flow、Consistency/Flow-Map类中训练加速方法与高阶求解器正在逐步融合,流匹配正由“可行替代”走向“统一框架中的核心参数化选项”[16,26-27]。

3.4 扩散模型与流匹配的比较讨论

扩散模型与流匹配可分别理解为“分数场建模”与“速度场建模”的两种代表路径。扩散模型通过前向扰动与反向去噪构造生成过程,训练目标与变分推断、分数学习具有稳定对应;流匹配则在预设概率路径上直接学习速度场,用常微分方程统一刻画从先验到数据分布的传输[6,8,16]。两者的工程权衡主要体现在路径设计、采样效率与系统成熟度3个维度。扩散模型依托成熟噪声调度与数值求解器,在大规模条件生成任务中表现稳健;流匹配通过更平直的传输路径在低函数评估次数区间具有显著潜力,特别是在少步采样场景下优势明显[26-27,37]。因此,扩散模型与流匹配更适合作为统一框架下的两类参数化选择,而非简单替代关系。面向高质量优先场景,扩散路线通常更稳妥;面向低时延与少步采样场景,流匹配路线往往更具吸引力。

4 结构化比较

在时间线梳理之外,本节采用“似然是否显式可计算”与“生成轨迹是否显式建模”两条轴线对概率建模方法进行结构化分类,3种路径分别为:显式似然/变分路径(VAE及其层次化扩展)[10,38-39],隐式似

然/对抗路径(GAN家族)[40-41],以及分数与流场路径(DDPM、EDM(Elucidating diffusion model)、Rectified flow、Flow matching及随机插值框架)[8,27,42-44]。这种分类方式更便于比较不同方法在可解释性、训练稳定性、采样效率及可控性上的系统差异。

从统一理论看,变分目标、分数目标与速度场目标在条件化形式下具有深层对应关系:它们在优化层面往往可归约为等价的回归问题,仅在路径参数化与离散化实现上呈现差异。该观点为理解"同一模型族在不同训练表述下为何性能接近"提供了理论支撑,也解释了扩散、分数模型与流匹配之间日益增强的互通性[16]。

从综合实证结果可见,不同方法在质量、效率、稳定性与可控性之间呈现系统性权衡。结合第3.4节的机制比较,更可行的实践策略不是在扩散与流匹配之间做单一路线替代,而是在统一框架下按任务目标选择路径参数化、训练目标与求解器组合,并配合蒸馏与引导策略实现整体优化[16,26,37,43,45-46]。

5 性能比较

为保证横向比较具备可复现性,本文在性能对比中统一采用CIFAR-10、ImageNet 64×64和ImageNet 512×512三类基准数据集:第1类用于无条件生成比较,后两类用于类条件生成比较[47-49]。

指标方面,NFE用于衡量推理阶段网络函数评估次数,之所以采用NFE作为效率主指标,是因为生成推理中的主要计算消耗通常来自神经网络前向评估;FID(Fréchet inception distance)衡量生成分布与真实分布在特征空间中的距离[50-52],可以综合衡量生成保真度和多样性,是文献中用于比较生成效果的主要指标。

为简洁且覆盖主要结论,表1~3比较主要以FID为例,并考虑NFE作为计算消耗[49]。关于"为何仍需较多步数",需要区分"采样步数"与"NFE"。前者是离散时间网格上的积分步,后者是一次采样中

表1 无条件CIFAR-10生成样本质量对比[49,53]

Table 1 Comparison of sample quality on unconditional CIFAR-10[49,53]

Model	Method	NFE (↓)	FID (↓)
Diffusion model	LSGM[53]	138[54]	2.10
	Score SDE (deep)[11]	2 000	2.20
	EDM[43]	35	2.01
Flow model	Flow Matching[8]	142	6.35
	OT-CFM[55]	1 000	3.57
	2-Rectified Flow[27]	1	4.85
	Consistency-FM[56]	2	5.34
GAN	Diffusion GAN[57]	4	3.75
	Diffusion StyleGAN[58]	1	3.19
	StyleGAN-XL[59]	1	1.52
VAE	NVAE[10]	1	23.49
	VAEBM[38]	1	12.19
	NCP-VAE[60]	1	24.08
	DC-VAE[61]	1	17.90
Diffusion model + advanced numerical solver	DPM-Solver[62]	10	4.70
	DPM-Solver++[63]	10	2.91
	DPM-Solver-v3[64]	10	2.51

注:↓表示数值越小越好。

表2 类条件ImageNet 64×64样本质量对比[49]

Table 2 Comparison of sample quality on class-conditional ImageNet 64×64[49]

Model	Method	NFE (↓)	FID (↓)
Diffusion model	ADM[45]	250	2.07
	RIN[65]	1 000	1.23
GAN	StyleGAN-XL[59]	1	1.52
Diffusion model + advanced numerical solver	DPM-Solver[62]	20	3.42
	EDM (Heun)[43]	79	2.44
	EDM2 (Heun)[66]	63	1.33

注:↓表示数值越小越好。

表3 类条件ImageNet 512×512样本质量对比[49]

Table 3 Comparison of sample quality on class-conditional ImageNet 512×512[49]

Model	Method	NFE (↓)	FID (↓)	Params/10^6
Diffusion model	ADM-G[45]	250×2	7.72	559
	RIN[65]	1 000	3.95	320
	U-ViT-H/4[67]	250×2	4.05	501
	DiT-XL/2[68]	250×2	3.04	675
	SimDiff[69]	512×2	3.02	2 000
	VDM++[70]	512×2	2.65	2 000
	DiffiT[71]	250×2	2.67	561
	DiMR-XL/3R[72]	250×2	2.89	525
	DIFFUSSM-XL[73]	250×2	3.41	673
	DiM-H[74]	250×2	3.78	860
	U-DiT[75]	250	15.39	204
	SiT-XL (flow-interpolant)[46]	250×2	2.62	675
	Large-DiT[76]	250×2	2.52	3 000
	MaskDiT[77]	79×2	2.50	736
	DiS-H/2[78]	250×2	2.88	900
	DRWKV-H/2[79]	250×2	2.95	779
	EDM2-S[66]	63×2	2.23	280
	EDM2-M[66]	63×2	2.01	498
	EDM2-L[66]	63×2	1.88	778
	EDM2-XL[66]	63×2	1.85	1 100
	EDM2-XXL[66]	63×2	1.81	1 500
GAN + VQ-tokenizer + masked model	BigGAN[41]	1	8.43	160
	StyleGAN-XL[59]	1×2	2.41	168
	MaskGIT[80]	12	7.32	227
	MAGVIT-v2[81]	64×2	1.91	307
	MAR[82]	64×2	1.73	481
	VAR-d36-s[83]	10×2	2.63	2 300
	VQGAN[84]	1 024	26.52	227

注:↓表示数值越小越好。

网络函数被调用的总次数。高分辨率类条件生成通常需要沿概率路径进行多次迭代修正,因此基础NFE偏高;而表3中“2×a”通常表示每个采样步采用条件/无条件双分支引导(例如无分类器引导),总函数评估次数约为2a。本文沿用这一设置以保证不同方法在效率与质量上的可比性[28,49]。

为对应正文中讨论的代表性变种,表1~3已显式覆盖Flow matching与Rectified flow(表1)、DPM-Solver系列(表1与表2)、StyleGAN/Diffusion-GAN路线(表1~3),并在ImageNet 512×512比较中补充了VQ-tokenizer与掩码生成路线(VQGAN、MaskGIT、MAGVIT-v2、MAR、VAR),以反映“潜空间离散表征+生成器”一类方法的工程表现[8,27,49,57,59,62-64,80-84]。WGAN、CGAN、CycleGAN、beta-VAE和NVAE等方法在任务设定上与表1~3的统一评测设置不完全一致,因此保留在机制对比部分,不做直接数值横向对比[10,17,20-22]。

需要强调的是,VQGAN并非标准VAE,它采用VQ-VAE式离散潜变量编码,并结合对抗训练提升感知质量,不并入传统VAE类别[14,84]。表1中VAE子类的FID数据补充自文献[53]的表2。考虑到标准VAE与beta-VAE更侧重表征解耦、归一化流更侧重似然估计,本文在不改变表1~3的FID对比设置前提下补充表4,汇总VAE与归一化流的可核查量化结果[10,17,85]。

表4 VAE与归一化流补充量化结果(非FID设置)[10,17,85]

Table 4 Supplementary quantitative results of VAE and normalizing flow (non-FID setting)[10,17,85]

Model	Method	Dataset	Metric	Value
VAE + Normalizing flow (non-FID protocol)	VAE (beta=1)[17]	2D Shapes	Factor-CLS average accuracy (↑)	21.77
	beta-VAE (beta=4)[17]	2D Shapes	Factor-CLS average accuracy (↑)	43.04
	NVAE w/o flow[10]	CIFAR-10 32×32	Bits/dimension (↓)	2.93
	NVAE w/ flow[10]	CIFAR-10 32×32	Bits/dimension (↓)	2.91
Normalizing flow (CIFAR-10, likelihood protocol)	FFJORD[86]	CIFAR-10 32×32	Negative Log-Likehood (↓)	3.40
	Flow++[87]	CIFAR-10 32×32	Negative Log-Likehood (↓)	3.08
	VFlow[88]	CIFAR-10 32×32	Negative Log-Likehood (↓)	2.98
	ANF[89]	CIFAR-10 32×32	Negative Log-Likehood (↓)	3.05

注:↑表示数值越大越好,↓表示数值越小越好。

6 应用领域简介

数据的概率模型在机器学习、深度学习及人工智能中处于核心地位,应用范围覆盖表示学习、可控生成、反问题重建和科学计算等方向[16,90]。下面按照“表示学习与生成建模—视觉反问题—多模态与科学控制”3条主线简介其应用。

在表示学习与图像生成场景中,VAE依托显式潜变量建模,广泛用于解耦表示学习、压缩表征与可控生成[3,10,17]。GAN在高保真图像生成、图像翻译与风格迁移等任务中仍具优势[19,22-23]。

在计算机视觉反问题(如去噪、去模糊、超分辨与重建等)中,生成模型可将观测一致性与数据先验结合,如:DPS用于一般噪声逆问题的后验采样[91];伪逆引导扩散模型通过伪逆引导提升线性逆问题稳定性[92];扩散零空间模型在零样本设定下实现图像恢复[93]。这些结果表明,扩散先验已从“样本生成”进一步拓展到“生成驱动求解”[91-93]。

概率建模的核心假设是:观测样本来自未知但客观存在的底层概率分布。样本可视为对该分布的独立采样,模型通过学习分数函数或速度场等关键函数,在概率路径上完成从噪声到数据的反演生成[8,11,16]。当数据与条件信息成对出现时,可进一步建模条件分布$p(\boldsymbol{x}|c)$,其中$\boldsymbol{x}$为目标样本,c为条件变量,从而把生成模型从“无条件拟合”推进到“可控生成”[90]。

在多模态领域,生成任务常可写为条件分布$p(\boldsymbol{x}_{\text{img}}|c_{\text{text}})$,其中$c_{\text{text}}$为文本指令,$\boldsymbol{x}_{\text{img}}$为图像样本。Stable Diffusion/SDXL将扩散过程放入潜空间,并通过跨注意力注入文本条件,实现了质量与成本之间的工程平衡[18,94]。DALL·E3通过高质量再标注提升复杂提示词遵循能力,说明训练数据中的文本-图像条件质量会直接影响条件分布的学习效果[95]。FLUX.1 Kontext与Stable Diffusion 3 Medium进一步把流匹配和长提示词理解用于可编辑生成,推动商业系统从“生成图像”向“可控编辑”演进[96-97]。

在视频与音频方向,建模对象可分别写为$p(\boldsymbol{x}_{1:T}|c_{\text{text}}, c_{\text{motion}}, c_{\text{ref}})$和$p(\boldsymbol{x}_{\text{audio}}|c_{\text{text}}, c_{\text{speaker}}, c_{\text{style}})$,其中$\boldsymbol{x}_{1:T}$为视频序列,$\boldsymbol{x}_{\text{audio}}$为音频样本,$c_{\text{text}}$为文本条件,$c_{\text{motion}}$为运动条件,$c_{\text{ref}}$为参考上下文,$c_{\text{speaker}}$为说话人条件,$c_{\text{style}}$为风格条件等。MovieGen与Gen-3 Alpha的实践表明,关键帧、运动条件与上下文重用能够显著提升长序列一致性[98-99]。SpeechFlow与Audiobox则说明同一概率路径框架可以迁移到语音预训练与统一音频生成,条件变量可细化到音素、说话人和风格控制[100-101]。从方法实现看,这些系统仍遵循“条件注入 + 迭代去噪/流场积分”的共同范式[90]。

在科学建模与决策控制方向,如蛋白质设计、机器人动作控制,任务可写为$p(\boldsymbol{x}_{\text{struct}}|c_{\text{target}})$和$p(a_{1:T}|o_{1:T}, g)$,其中$\boldsymbol{x}_{\text{struct}}$为结构样本,$c_{\text{target}}$为目标条件,$a_{1:T}$为动作序列,$o_{1:T}$为时序观测,$g$为任务指令。SE(3)-Stochastic Flow Matching将几何等变先验并入概率路径学习,已用于蛋白骨架生成[102]。具身智能模型π_0进一步把流模型拓展到视觉-语言-动作一体化控制,说明该路线可以从“样本生成”延展到“生成+控制”[103]。这与扩散模型跨学科应用的总体趋势一致,即把领域先验与条件采样耦合以提升可行性和任务有效性[90]。

尽管有上述进展,应用的具体落实仍存在一些限制,可归纳为3点。(1)概率建模通常通过基于期望的损失最小化进行参数估计,效果高度依赖大规模高质量数据[16];在多类应用中,高质量配对数据仍然稀缺,数据质量与覆盖度直接限制模型上限,这也制约了在数据受限场景下的落地[90]。(2)模型对数据特征的刻画主要以隐分布参数化方式存在,缺少可直接操控的显式语义表示;与VAE、GAN相比,扩散与流匹配模型在潜表示可操作性上相对较弱,且潜空间常与数据空间同维,影响表示学习与采样效率[90]。(3)当前主流分数匹配与流匹配方法的高质量采样仍需多步迭代,推理阶段需要大量函数评估,相比VAE、GAN等一步生成模型在时延敏感场景下仍存在效率瓶颈[16,49,90]。

7 结束语

本文从数据统一概率建模视角梳理了其从传统到现代的方法演进:在最大似然估计主线下连接高斯/混合高斯、EM、变分推理、VAE与GAN;在分数匹配主线下连接DSM、DDPM与Score-SDE;在流方法主线下连接归一化流、连续归一化流与流匹配[1-2,6-8,36]。3条主线共同指向“在概率路径上学习关键场”的统一范式,即分别学习似然、分数场或速度场,并在条件化建模中呈现互通[16]。结合结构化比较

与性能对比可见,不同方法的关键差异不在于是否彼此替代,而在于路径参数化、训练目标与数值求解器的组合方式及其工程化成熟度[49]。面向高质量优先任务,分数匹配路线通常更稳健;面向低时延和少步采样任务,最大似然估计、流匹配路线更具潜力[26,27,37]。从数学原理上看,3条路线本质上具有相通性。应用情况则表明,可控生成能力正从图像扩展到视频、音频与科学控制场景,但统一框架下的跨任务迁移规律仍需系统化总结[90]。下一阶段,将围绕“路径、目标、高效求解、引导”4层协同推进,并正面回应应用落地的3类瓶颈:数据质量与规模依赖、潜表示可操作性不足、多步采样时延压力[16,49,90]。与此同时,离散数据与多模态统一建模,以及“效果-效率-安全-合规”四维一体的评价治理闭环,将是生成模型从研究原型走向大规模可信应用的关键[104-107]。

参考文献:

[1] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[2] HYVÄRINEN A. Estimation of non-normalized statistical models by score matching[J]. Journal of Machine Learning Research, 2005(6): 695-709.

[3] CHEN Y, LIU J, PENG L, et al. Auto-encoding variational Bayes[J]. Cambridge Explorations in Arts and Sciences, 2024, 2(1): 1-14.

[4] DOERSCH C. Tutorial on variational autoencoders[EB/OL]. (2016-06-19). https://arxiv.org/abs/1606.05908.

[5] KINGMA D P, WELLING M. An introduction to variational autoencoders[EB/OL]. (2019-06-06). https://arxiv.org/abs/1906.02691.

[6] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. (2020-06-19). https://arxiv.org/abs/2006.11239.

[7] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[EB/OL]. (2015-05-21). https://arxiv.org/abs/1505.05770.

[8] LIPMAN Y, CHEN R T Q, BEN-HAMU H, et al. Flow matching for generative modeling[EB/OL]. (2022-10-10). https://arxiv.org/abs/2210.02747.

[9] JOHNSON M J, DUVENAUD D, WILTSCHKO A B, et al. Composing graphical models with neural networks for structured representations and fast inference[EB/OL]. (2016-03-20). https://arxiv.org/abs/1603.06277.

[10] VAHDAT A, KAUTZ J. NVAE: A deep hierarchical variational autoencoder[EB/OL]. (2020-07-10). https://arxiv.org/abs/2007.03898.

[11] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[EB/OL]. (2020-11-03). https://arxiv.org/abs/2011.13456.

[12] GREGOR K, DANIHELKA I, MNIH A, et al. Deep AutoRegressive networks[EB/OL]. (2013-10-31). https://arxiv.org/abs/1310.8499.

[13] VAN DEN OORD A, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2017.

[14] RAZAVI A, VAN DEN OORD A, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2019.

[15] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric [C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. [S.l.]: ACM, 2016.

[16] LAI C H, SONG Y, KIM D, et al. The principles of diffusion models[EB/OL]. (2025-10-24). https://arxiv.org/abs/2510.21890.

[17] HIGGINS I, MATTHEY L, PAL A, et al. Beta-VAE: Learning basic visual concepts with a constrained variational framework[C]//Porceedings of International Conference on Learning Representations. [S.l.]: [s.n.], 2016.

[18] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Porceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 10674-10685.

[19] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2014.

[20] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: [s.n.], 2017.

[21] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-03). https://arxiv.org/abs/1411.1784.

[22] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 2242-2251.

[23] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 4396-4405.

[24] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[EB/OL]. (2015-03-11). https://arxiv.org/abs/1503.03585.

[25] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[EB/OL]. (2020-10-07). https://arxiv. org/abs/2010.02502.

[26] HU Z, LAI C H, MITSUFUJI Y, et al. CMT: Mid-training for efficient learning of consistency, mean flow, and flow map models[EB/OL]. (2025-09-11). https://arxiv.org/abs/2509.24526.

[27] LIU X, GONG C, LIU Q. Flow straight and fast: Learning to generate and transfer data with rectified flow[EB/OL]. (2022-09-21). https://arxiv.org/abs/2209.03003.

[28] HO J, SALIMANS T. Classifier-free diffusion guidance[C]//Proceedings of NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications. [S.l.]: [s.n.], 2021.

[29] LYU S. Interpretation and generalization of score matching[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Montreal, Quebec, Canada: ACM, 2009: 359-366.

[30] VINCENT P. A connection between score matching and denoising autoencoders[J]. Neural Computation, 2011, 23(7): 1661-1674.

[31] KINGMA D P, KINGMA D P, LECUN Y, et al. Regularized estimation of image statistics by score matching[C]//Proceedings of the 24th International Conference on Neural Information Processing Systems—Volume 1. Vancouver, British Columbia, Canada: ACM, 2010: 1126-1134.

[32] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2019: 11895-11907.

[33] SONG Y, GARG S, SHI J, et al. Sliced score matching: A scalable approach to density and score estimation[C]//Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence. Tel Aviv, Israel: [s.n.], 2019.

[34] WELLING M, WELLING M, TEH Y W, et al. Bayesian learning via stochastic gradient Langevin dynamics[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, Washington, USA: ACM, 2011: 681-688.

[35] ANDERSON B D O. Reverse-time diffusion equation models[J]. Stochastic Processes and Their Applications, 1982, 12(3): 313-326.

[36] CHEN RI T Q, RUBANOVA Y, BETTENCOURT J, et al. Neural ordinary differential equation[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2018.

[37] DENG M, LI H, LI T, et al. Generative modeling via drifting[EB/OL]. (2026-02-24). https://arxiv.org/abs/2602.04770.

[38] XIAO Z, KREIS K, KAUTZ J, et al. VAEBM: A symbiosis between variational autoencoders and energy-based models[EB/OL]. (2020-10-30). https://arxiv.org/abs/2010.00654.

[39] CHILD R. Very deep VAEs generalize autoregressive models and can outperform them on images[C]//Proceedings of International Conference on Learning Representations. Vienna, Austria: ICML, 2021.

[40] KARRAS T, AITTALA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[EB/OL]. (2020-06-11). https://arxiv.org/abs/2006.06676.

[41] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[EB/OL]. (2018-09-28). https://arxiv.org/abs/1809.11096.

[42] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. (2020-06-19). https://arxiv. org/abs/2006.11239.

[43] KARRAS T, AITTALA M, AILA T, et al. Elucidating the design space of diffusion-based generative models[EB/OL]. (2022-06-01). https://arxiv.org/abs/2206.00364.

[44] ALBERGO M S, VANDEN-EIJNDEN E. Building normalizing flows with stochastic interpolants[EB/OL]. (2022-09-30). https://arxiv.org/abs/2209.15571.

[45] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[EB/OL]. (2021-05-11). https://arxiv.org/abs/2105.05233.

[46] MA N, GOLDSTEIN M, ALBERGO M S, et al. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers[C]//Proceedings of ECCV 2024. Cham: Springer Nature Switzerland, 2024: 23-40.

[47] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.

[48] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248-255.

[49] LU C, SONG Y. Simplifying, stabilizing and scaling continuous-time consistency models[EB/OL]. (2024-10-14). https://arxiv.org/abs/2410.11081.

[50] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing System. [S.l.]: ACM, 2016.

[51] HEUSEL M, HEUSEL M, RAMSAUER H, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: ACM, 2017: 6629-6640.

[52] KYNKAANNIEMI T, KARRAS T, LAINE S, et al. Improved precision and recall metric for assessing generative models [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2019.

[53] VAHDAT A, KREIS K, KAUTZ J. Score-based generative modeling in latent space[C]//Proceedings of Neural Information Processing Systems. [S.l.]: [s.n.], 2021.

[54] PARK D, LEE S, KIM S, et al. Constant acceleration flow[EB/OL]. (2024-09-26). https://openreview. net/forum id= hsgNvC5YM9.

[55] TONG A, FATRAS K, MALKIN N, et al. Improving and generalizing flow-based generative models with minibatch optimal transport[EB/OL]. (2023-02-13). https://arxiv.org/abs/2302.00482.

[56] YANG L, ZHANG Z, ZHANG Z, et al. Consistency flow matching: Defining straight flows with velocity consistency[EB/OL]. (2024-07-31). https://arxiv.org/abs/2407.02398.

[57] XIAO Z, KREIS K, VAHDAT A. Tackling the generative learning trilemma with denoising diffusion GANs[EB/OL]. (2021-12-30). https://arxiv.org/abs/2112.07804.

[58] WANG Z, ZHENG H, HE P, et al. Diffusion-GAN: Training GANs with diffusion[EB/OL]. (2022-06-11). https://arxiv.org/abs/2206.02262.

[59] SAUER A, SAUER A, SCHWARZ K, et al. StyleGAN-XL: Scaling StyleGAN to large diverse datasets[C]//Proceedings of ACM SIGGRAPH 2022 Conference Proceedings. Vancouver, BC, Canada: ACM, 2022: 1-10.

[60] ANEJA J, SCHWING A, KAUTZ J, et al. NCP-VAE: Variational autoencoders with noise contrastive priors[EB/OL]. (2020-10-06). https://arxiv.org/abs/2010.02917v1.

[61] PARMAR G, LI D, LEE K, et al. Dual contradistinctive generative autoencoder[EB/OL]. (2020-11-21). https://arxiv.org/abs/2011.10063.

[62] BAO F, CHEN J, LI C, et al. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, Louisiana, USA: ACM, 2022.

[63] LU C, ZHOU Y, BAO F, et al. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models[EB/OL]. (2022-11-29). https://arxiv.org/abs/2211.01095.

[64] ZHENG K, LU C, CHEN J, et al. DPM-solver-v3: Improved diffusion ODE solver with empirical model statistics[EB/OL]. (2023-10-11). https://arxiv.org/abs/2310.13268.

[65] JABRI A, FLEET D, CHEN T. Scalable adaptive computation for iterative generation[EB/OL].(2022-12-31). https://arxiv.org/abs/2212.11972.

[66] KARRAS T, AITTALA M, LEHTINEN J, et al. Analyzing and improving the training dynamics of diffusion models[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 24174-24184.

[67] BAO F, NIE S, XUE K, et al. All are worth words: A ViT backbone for diffusion models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 22669-22679.

[68] PEEBLES W, XIE S. Scalable diffusion models with transformers[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 4172-4182.

[69] HOOGEBOOM E, HEEK J, SALIMANS T. Simple diffusion: End-to-end diffusion for high resolution images[C]//Proceedings of International Conference on Machine Learning. [S.l.]: ACM, 2023.

[70] KINGMA D P, GAO R. Understanding diffusion objectives as the ELBO with simple data augmentation[EB/OL]. (2023-03-21). https://arxiv.org/abs/2303.00848.

[71] HATAMIZADEH A, SONG J, LIU G, et al. DiffiT: Diffusion vision transformers for image generation[EB/OL]. (2023-12-21). https://arxiv.org/abs/2312.02139.

[72] LIU Q, ZENG Z, HE J, et al. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization[EB/OL]. (2024-06-09). https://arxiv.org/abs/2406.09416.

[73] YAN J N, GU J, RUSH A M. Diffusion models without attention[EB/OL]. (2023-11-29). https://arxiv.org/abs/2311.18257.

[74] TENG Y, WU Y, SHI H, et al. DiM: Diffusion mamba for efficient high-resolution image synthesis[EB/OL]. (2024-05-15). https://arxiv.org/abs/2405.14224.

[75] TIAN Y, TU Z, CHEN H, et al. U-DiTs: Downsample tokens in U-shaped diffusion transformers[EB/OL]. (2024-05-12). https://arxiv.org/abs/2405.02730.

[76] Alpha-VLLM. Large-DiT-ImageNet[EB/OL]. (2024-01-01) [2026-02-28]. https://github.com/Alpha-VLLM/LLaMA2-Accessory/tree/f7fe19834b23e38f333403b91bb0330afe19f79e/Large-DiT-ImageNet.

[77] ZHENG H, NIE W, VAHDAT A, et al. Fast training of diffusion models with masked transformers[EB/OL]. (2023-06-21). https://arxiv.org/abs/2306.09305.

[78] FEI Z, FAN M, YU C, et al. Scalable diffusion models with state space backbone[EB/OL]. (2024-02-17). https://arxiv.org/abs/2402.05608.

[79] FEI Z, FAN M, YU C, et al. Diffusion-RWKV: Scaling RWKV-like architectures for diffusion models[EB/OL]. (2024-04-01). https://arxiv.org/abs/2404.04478.

[80] CHANG H, ZHANG H, JIANG L, et al. MaskGIT: Masked generative image transformer[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 11305-11315.

[81] YU L, LEZAMA J, GUNDAVARAPU N B, et al. Language model beats diffusion: Tokenizer is key to visual generation [EB/OL]. (2023-10-11). https://arxiv.org/abs/2310.05737.

[82] LI T, TIAN Y, LI H, et al. Autoregressive image generation without vector quantization[EB/OL]. (2024-06-19). https://

arxiv.org/abs/2406.11838.

[83] TIAN K, JIANG Y, YUAN Z, et al. Visual autoregressive modeling: Scalable image generation via next-scale prediction[EB/OL]. (2024-04-18). https://arxiv.org/abs/2404.02905.

[84] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 12868-12878.

[85] SONG Y, DURKAN C, MURRAY I, et al. Maximum likelihood training of score-based diffusion models[EB/OL]. (2021-01-10). https://arxiv.org/abs/2101.09258.

[86] GRATHWOHL W, CHEN R T Q, BETTENCOURT J, et al. FFJORD: Free-form continuous dynamics for scalable reversible generative models[EB/OL]. (2018-10-11). https://arxiv.org/abs/1810.01367.

[87] HO J, CHEN X, SRINIVAS A, et al. Flow++: Improving flow-based generative models with variational dequantization and architecture design[EB/OL]. (2019-02-13). https://arxiv.org/abs/1902.00275.

[88] CHEN J, LU C, CHENLI B, et al. VFlow: More expressive generative flows with variational data augmentation[C]//Proceedings of the 37th International Conference on Machine Learning. [S.l.]: ACM, 2020.

[89] HUANG C W, DINH L, COURVILLE A. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models[EB/OL]. (2020-02-24). https://arxiv.org/abs/2002.07101.

[90] YANG L, ZHANG Z, SONG Y, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2024, 56(4): 1-39.

[91] CHUNG H, KIM J, MCCANN M T, et al. Diffusion posterior sampling for general noisy inverse problems[EB/OL]. (2022-09-01). https://arxiv.org/abs/2209.14687.

[92] SONG J M, VAHDAT A, MARDANI M, et al. Pseudoinverse-guided diffusion models for inverse problems[C]//Proceedings of International Conference on Learning Representations. [S.l.]: [s.n.], 2023.

[93] WANG Y, YU J, ZHANG J. Zero-shot image restoration using denoising diffusion null-space model[EB/OL]. (2022-12-21). https://arxiv.org/abs/2212.00490.

[94] PODELL D, ENGLISH Z, LACEY K, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis [EB/OL]. (2023-07-08). https://arxiv.org/abs/2307.01952.

[95] BETKER J, GOH G, JING L, et al. Improving image generation with better captions[EB/OL]. (2023-01-01)[2026-02-28]. https://cdn.openai.com/papers/dall-e-3.pdf.

[96] LABS B F, BATIFOL S, BLATTMANN A, et al. FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space[EB/OL]. (2025-06-11). https://arxiv.org/abs/2506.15742.

[97] STABILITY AI. Stable diffusion 3 medium[EB/OL]. (2024-01-01)[2026-02-28]. https://stability.ai/news/stable-diffusion-3-medium.

[98] POLYAK A, ZOHAR A, BROWN A, et al. Movie Gen: A cast of media foundation models[EB/OL]. (2024-10-17). https://arxiv.org/abs/2410.13720.

[99] Runway. Introducing Gen-3 Alpha: A new frontier for video generation[EB/OL]. (2024-01-01)[2026-02-28]. https://runwayml.com/research/introducing-gen-3-alpha.

[100] LIU A H, LE M, VYAS A, et al. Generative pre-training for speech with flow matching[EB/OL]. (2023-10-15). https://arxiv.org/abs/2310.16338.

[101] VYAS A, SHI B, LE M, et al. Audiobox: Unified audio generation with natural language prompts[EB/OL]. (2023-10-25). https://arxiv.org/abs/2312.15821.

[102] BOSE A, AKHOUND-SADEGH T, HUGUET G, et al. SE(3)-Stochastic flow matching for protein backbone generation [EB/OL]. (2023-10-03). https://arxiv.org/abs/2310.02391.

[103] BLACK K, BROWN N, DRIESS D, et al. π_0: A vision-language-action flow model for general robot control[J]. Robotics: Science and Systems Foundation, 2025. DOI:10.15607/rss.2025.xxi.010.

[104] AUSTIN J, JOHNSON D D, HO J, et al. Structured denoising diffusion models in discrete state-spaces[EB/OL]. (2021-07-07). https://arxiv.org/abs/2107.03006.

[105] LI X L, THICKSTUN J, GULRAJANI I, et al. Diffusion-LM improves controllable text generation[EB/OL]. (2022-05-10). https://arxiv.org/abs/2205.14217.

[106] CARLINI N, HAYES J, NASR M, et al. Extracting training data from diffusion models[EB/OL]. (2023-01-28). https://arxiv.org/abs/2301.13188.

[107] FERNANDEZ P, COUAIRON G, JÉGOU H, et al. The stable signature: Rooting watermarks in latent diffusion models [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 22409-22420.

作者简介:

卢宏涛(1967-),**通信作者**,男,教授,研究方向:机器学习、计算机视觉,E-mail:htlu@sjtu.edu.cn。

胡宇庭(1998-),男,博士研究生,研究方向:扩散模型、流匹配模型,E-mail:yutinghu2024@sjtu.edu.cn。

(编辑:张黄群)