智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

扩散模型在计算机视觉领域的研究现状

管凤旭, 张涵宇, 路斯棋, 赖海涛, 杜雪, 郑岩

引用本文:
管凤旭, 张涵宇, 路斯棋, 等. 扩散模型在计算机视觉领域的研究现状[J]. 智能系统学报, 2025, 20(2): 265-282.
GUAN Fengxu, ZHANG Hanyu, LU Siqi, et al. Research status of diffusion models in computer vision[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(2): 265-282.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202312041

DOI: 10.11992/tis.202312041
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250108.0933.013

扩散模型在计算机视觉领域的研究现状

管凤旭, 张涵宇, 路斯棋, 赖海涛, 杜雪, 郑岩
(哈尔滨工程大学 智能科学与工程学院, 黑龙江 哈尔滨 150001)

摘　要: 扩散模型是受分子热力学启发而来的一类新的生成模型, 具有训练稳定、对模型设置依赖性弱等优点。近年来, 扩散模型被广泛应用于各项任务, 并且取得了相比于以往生成模型更多样、更高质量的结果。目前, 扩散模型已成为计算机视觉领域热门的基准方法。为更好地促进扩散模型在计算机视觉领域的发展, 对扩散模型进行综述: 首先对比了扩散模型与其他生成模型的优劣, 介绍了扩散模型的数学原理; 随后, 从扩散模型存在的普遍问题出发, 介绍了相关学者近年来所做的改进工作, 以及扩散模型在多种视觉任务上的应用实例; 最后, 探讨了扩散模型存在的问题, 并提出了一些未来可能的发展趋势。

关键词: 扩散模型; 去噪扩散概率模型; 分数扩散模型; 深度学习; 计算机视觉; 图像生成; 生成模型; 生成对抗网络
中图分类号: TP18　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)02−0265−18

中文引用格式: 管凤旭, 张涵宇, 路斯棋, 等. 扩散模型在计算机视觉领域的研究现状 [J]. 智能系统学报, 2025, 20(2): 265–282.
英文引用格式: GUAN Fengxu, ZHANG Hanyu, LU Siqi, et al. Research status of diffusion models in computer vision[J]. CAAI transactions on intelligent systems, 2025, 20(2): 265–282.

Research status of diffusion models in computer vision

GUAN Fengxu, ZHANG Hanyu, LU Siqi, LAI Haitao, DU Xue, ZHENG Yan
(College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: The diffusion model is a new generative model inspired by molecular thermodynamics. This model offers stable training and low dependence on model settings, making it a popular benchmark in computer vision. In recent years, the diffusion model has been widely applied to various tasks, yielding diverse and high-quality results compared to traditional generative models. At present, the diffusion model is a prominent method in the field of computer vision. This paper provides a comprehensive overview of the diffusion model to further stimulate its development in this domain. First, the paper compares the advantages and disadvantages of diffusion models with other generative models and introduces the underlying mathematical principles. Then, the study presents recent efforts by researchers to improve diffusion models, starting with common challenges and highlighting application examples in various visual tasks. Finally, the study discusses existing issues with diffusion models and outlines potential future development trends.

Keywords: diffusion model; denoising diffusion probabilistic model; score-based generative model; deep learning; computer vision; image generation; generative model; generative adversarial network

生成模型是一种能够根据观测数据随机生成新样本的概率模型。在机器学习中, 通过学习数据的概率分布, 生成模型可以捕捉数据中的特征和结构, 从而生成类似于训练数据的新样本。早期的生成模型[1-3]主要是通过对图像像素和特征进行处理, 将其作为图像生成任务中的最小单元。然而, 由于存在计算机性能和建模条件复杂等限制, 这些传统方法在复杂场景条件下存在性能受限的情况。

近年来, 人工神经网络等技术的不断发展, 推动了生成模型与深度学习方法的结合。相比于传

收稿日期: 2023−12−27.　**网络出版日期**: 2025−01−08.
基金项目: 国家自然科学基金项目 (62101156).
通信作者: 管凤旭. E-mail: guanfengxu@hrbeu.edu.cn.

统方法，流模型 (flow-based models, Flow)[4]、变分自编码器 (variational auto-encoders, VAE)[5]、生成对抗网络 (generative adversarial networks, GANs)[6]等基于深度学习的生成模型，不仅可以降低数据获取和分布建模的难度、同时学习多个输出、完成更多类型的任务，还能获得更好的似然，从而提升数据质量。

2020 年，去噪扩散概率模型 (denoising diffusion probabilistic model, DDPM)[7] 的提出成为生成模型研究的重要进展。研究人员从分子热力学中获得启发，人为将输入分布以噪声扩散的方式逐步模糊，并利用神经网络进行分布重建，在图像生成任务上突破了新的技术水准。作为一类新的生成范式，扩散模型容易训练，生成质量高，在图像任务中取得了极为优秀的效果[8-10]，并迅速成为了计算机视觉领域的研究热点。

然而，相较于其他生成模型，扩散模型在该领域的综述性研究却相当匮乏[11-12]。为促进学者对扩散模型的深入理解，本文专注于扩散模型与计算机视觉的交叉领域，旨在为该领域的相关学者提供更深入的见解和更加充分的研究视角。综述内容从以下几个方面展开：

第 1 节从原理出发，介绍并对比了扩散模型与其他流行的生成模型之间的优缺点。

第 2 节从去噪扩散概率模型和基于分数的扩散模型 2 个重要分支出发，分别介绍二者的理论基础、模型训练与推理、目标函数训练等关键内容。

第 3 节从采样加速、似然优化和数据类型多样化 3 个维度，深入阐述了近年来扩散模型的改进。

第 4 节则聚焦于扩散模型在计算机视觉领域的应用，并呈现了近期工作的对比。

第 5 节总结了扩散模型目前仍然面临的问题，并从不同角度针对其未来的发展趋势提出了一定见解，旨在促进扩散模型在应用和技术层面的进一步发展。

1　多种生成模型的对比

从原理角度上看，生成模型的核心在于将先验分布的样本映射到数据分布的样本，不同生成模型的实现方式各有差异。

生成对抗网络[13-14] 采用对抗损失的方式进行分布映射，得益于无需对复杂数据分布进行建模，生成对抗网络被广泛应用于各类任务[15-18]。但损失函数的鞍点处理较为复杂，同时存在额外的判别器，导致模型训练困难，需要精心设计的正则化和优化技巧来避免这些问题的发生。变分自编码器[19-22] 通过对比真实样本与生成样本训练模型，但选择变分后验时，难以平衡模型的计算量和表达能力，导致生成图像的清晰度不足，质量较低。归一化流[23] 与变分自编码器类似，但该过程是通过一系列可逆变换来构建的，也因此使得流模型的推理能力受到限制，在高分辨率图像生成任务上，数据量激增的问题难以解决。

相较于其他生成模型，扩散模型 (diffusion models, DMs)[24] 不依赖对抗训练，不需要额外训练判别器，因此训练目标函数简单，不容易出现 GAN 训练时模式崩溃等问题；扩散模型具有良好的可解释性，并且经过坚实的数学推理证明，后验分布的似然值也明显高于 VAE 方法；此外，在高分辨率生成任务中，模型参数量不会成倍增长，训练成本明显低于 Flow 方法。然而，标准的扩散模型仍存在一些局限性，如推理速度慢、似然有待提高等。为了更直观地对比不同的生成模型，图 1 对比了模型间生成方式的差异、表 1 总结了不同模型的优缺点。

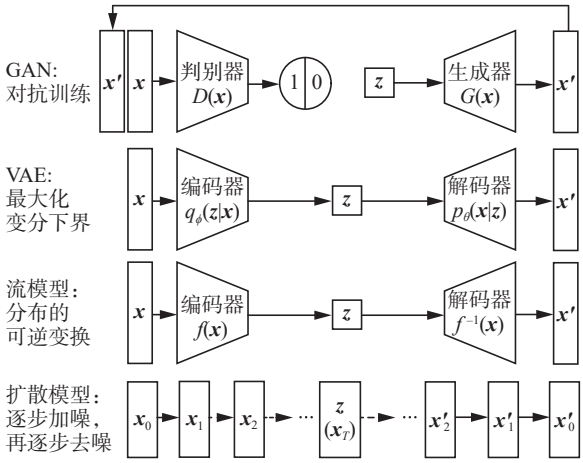


图 1　不同生成模型的流程图对比

Fig. 1　Comparison diagram of different generation models

表 1　不同生成模型的优劣对比

Table 1　Comparison of the advantages and disadvantages of different generative models

模型	年份	优势	劣势
变分自编码(VAE)	2013	可学习潜变量；是一种无监督学习方式；可处理离散及连续数据	难以选取合适的目标后验分布和先验分布，需要大量计算；生成结果不够精确

续表 1

模型	年份	优势	劣势
生成对抗网络(GAN)	2014	无需对复杂数据分布建模; 生成结果更清晰、真实; 模型推理速度快	不适合处理离散数据; 模型参数难以收敛, 需要精心设计正则化; 存在训练不稳定、梯度消失、模式崩溃的问题
流模型(Flow)	2014	直接面对生成模型的概率计算; 转换可逆	输入输出需在同一维度, 高分辨率下模型参数量会激增; 无法根据特定需求生成图像
扩散模型(DMs)	2020	训练稳定; 参数成本低; 数学推理证明坚实	推理速度慢; 似然性低; 数据类型单一

2　扩散模型的理论基础

本章节对去噪扩散概率模型 (DDPM) 和基于分数的扩散模型 (score-based generative models, SGM) 进行说明, 简要介绍它们的数学基础、模型的训练和推理、目标函数训练。

2.1　扩散模型介绍

扩散模型是一种利用噪声扰动和神经网络建模来学习目标数据分布的方法。总的来说, 扩散模型包括扩散过程和逆扩散过程 2 个阶段。在扩散过程中, 通过对初始数据分布施加若干次不同尺度的高斯噪声 (为方便叙述, 本章节提及噪声均为高斯噪声), 使数据逼近高斯分布, 通常来说, 该过程可控且不包含未知量。在逆扩散过程中, 从服从标准高斯分布的随机噪声出发, 使用神经网络预测并迭代计算后验分布, 并最终生成符合初始数据分布的样本。扩散模型流程如图 2 所示。

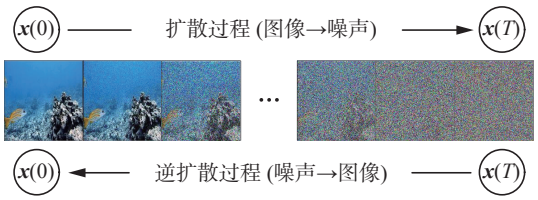


图 2　扩散模型的流程

Fig. 2　Process of diffusion model

下文介绍的去噪扩散概率模型和基于分数的扩散模型的流程均满足图 2, 不同的是它们的扩散与逆扩散过程所采用的处理方式存在差异。

2.2　去噪扩散概率模型

去噪扩散概率模型 (DDPM) 的流程如图 3 所示。简要来说, 去噪扩散概率模型可以被看作具有固定编码器的分层马尔可夫变分自编码器。具体而言, 去噪扩散概率模型的前向过程充当编码器, 并且该过程被构造为线性高斯模型。另一方面, 去噪扩散概率模型的反向过程对应于变分自编码器 (VAE) 的解码器, 该“解码器”在多个解码步骤之间共享[7], 其内在潜变量的大小都与样本数据的大小相同。

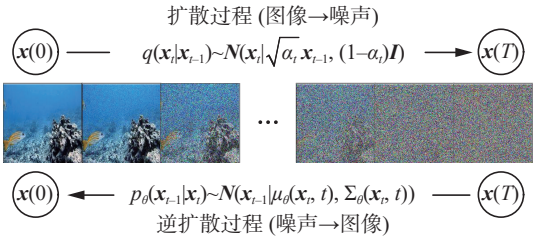


图 3　去噪扩散概率模型的流程

Fig. 3　Process of DDPM

2.2.1　去噪扩散概率模型的扩散过程

去噪扩散概率模型的扩散过程是一个迭代过程: 从初始数据分布中上采样出的图像 $\boldsymbol{x}_0$, 经过 T 次迭代计算后得到噪声图像 $\boldsymbol{x}_T$。去噪扩散概率模型的扩散过程满足马尔可夫链, 因此该过程后一时刻的结果仅与前一时刻有关, 在迭代过程中, 任意时刻的加噪结果可以由输入分布 $\boldsymbol{x}_0$ 直接计算得出。同时, 基于参数重整化技巧[19], 经过迭代计算, 最终得到 $\boldsymbol{x}_t$ 的递推公式为

$$\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}z,z\sim N(0,\boldsymbol{I}) \tag{1}$$

去噪扩散概率模型的扩散过程公式为

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0)\sim(\boldsymbol{x}_t;\sqrt{\bar{\alpha}}\boldsymbol{x}_0,(1-\bar{\alpha})\boldsymbol{I}) \tag{2}$$

式中: $\boldsymbol{x}_t$ 是对 $\boldsymbol{x}_0$ 加噪 t 次后的结果, $t\in[1,2,\cdots,T]$; α_t 是用于控制所加高斯噪声标准差的超参数, $\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$; $\boldsymbol{I}$ 是单位矩阵, $q(\cdot)$ 表示其数据分布。

从式 (2) 可以看出, 去噪扩散概率模型的扩散过程是一个不包含可学习参数的过程, 扩散过程中每一步骤的加噪结果均可由已知信息直接计算得出。随着迭代次数的增加, 初始的输入数据 $\boldsymbol{x}_0$ 逐步失去它的特征, 并最终趋向于符合式 (2) 的先验高斯分布。该分布在概率上表现为完全无序的噪声分布; 在图像上表现为一张纯高斯噪声的图片。

2.2.2　去噪扩散概率模型的逆扩散过程

逆扩散过程可以视为是扩散过程的反转, 仍然满足马尔可夫链的假设, 并且在每一时刻服从高斯分布[24]。然而, 由于无法通过迭代后验分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 得到初始输入分布 $q(\boldsymbol{x}_0|\boldsymbol{x}_1)$, Ho 等[7] 选择构建一个参数 θ 的网络 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, 对条件概率进行

估计，通过迭代最终得到近似分布为 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\sim N(\boldsymbol{x}_{t-1};\mu_\theta(\boldsymbol{x}_t,t),\Sigma_\theta(\boldsymbol{x}_t,t))$。

在给定 $\boldsymbol{x}_t$ 和 $\boldsymbol{x}_0$ 求取 $\boldsymbol{x}_{t-1}$ 时，根据式 (1) 和贝叶斯公式，可以得到 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 的均值和方差：

$$\mu_\theta(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\boldsymbol{z} \tag{3}$$

$$\Sigma_\theta(\boldsymbol{x}_t,t)=\frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t} \tag{4}$$

式中：$\beta_t=1-\alpha_t$，$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$，$\boldsymbol{z}\sim N(0,\boldsymbol{I})$。

由于式 (4) 是一个不含未知量的常数，因此通过预测随机噪声 $\boldsymbol{z}$ 进而计算后验分布中的均值便可以计算出近似的后验分布[7]。

2.2.3　去噪扩散概率模型的训练

采样成功的关键是训练 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 与正向马尔可夫链的实际时间反转 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 相匹配。Ho 等[7]通过构造 KL 散度 (Kullback-Leibler divergence) 并优化负似然函数的变分上界 L_{VLB}，从而等价地最小化损失函数，完成上述 2 个分布的拟合。经过化简，可以得到损失函数：

$$L=E_{t,x_0,\boldsymbol{\epsilon}}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t\right)\right\|^2 \tag{5}$$

因此，式 (3) 可以改写为

$$\mu_\theta(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t) \tag{6}$$

式中：E 为期望值，$t\in[1,T]$，$\boldsymbol{x}_0\sim p(\boldsymbol{x}_0)$，$\boldsymbol{\epsilon}\sim N(0,\boldsymbol{I})$ 为噪声，$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)$ 是预测噪声的神经网络。

从式 (5)、(6) 中可以看出，Ho 等[7]选择将网络建模为预测噪声，再代入计算均值，以此减小误差，从而实现更符合初始输入分布的样本生成。

综上所述，有关去噪扩散概率模型训练过程的算法步骤如下。

首先，输入数据分布 $\chi=\{\boldsymbol{x}^i\}_{i=1}^N$ 和超参数 α_t；

其次，设置迭代轮次并循环执行以下操作：

1）从输入数据集 χ 中采样 $\boldsymbol{x}_0$；

2）从均匀分布 $U(1,T)$ 中采样时间 t；

3）从标准高斯 $N(0,\boldsymbol{I})$ 分布中采样 $\boldsymbol{\epsilon}$；

4）执行梯度下降算法：

$$\nabla_\theta\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t\right)\right\|^2$$

直至收敛，停止。

有关去噪扩散概率模型推理过程的算法步骤如下：

首先，输入 α_t；

其次，从标准高斯分布 $N(0,\boldsymbol{I})$ 中采样得到 $\boldsymbol{x}_T$；

再次，设置时间步骤数 t 从 T 到 1 递减，并依次循环执行以下操作：

1）如果 $t>1$，从标准高斯 $N(0,\boldsymbol{I})$ 分布中采样噪声 $\boldsymbol{z}$，否则，令 $\boldsymbol{z}=0$；

2）对 $\boldsymbol{x}_{t-1}$ 参数重整化，得

$$\boldsymbol{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\right)+\sqrt{(1-\alpha_t)}\boldsymbol{z}$$

直至循环结束，输出 $\boldsymbol{x}_0$。

2.3　基于分数的扩散模型

作为扩散模型的另一重要分支，分数扩散模型[25]仍然遵循扩散模型的流程示意（图 2）。与去噪扩散概率模型不同的是，分数扩散模型的扩散过程利用随机微分方程 (stochastic differential equations, SDEs)，设计一个随时间演变的连续过程。

在逆扩散过程，分数扩散模型利用噪声条件分数网络 (noise conditional score network, NCSN)[26]估计与数据分布相关的梯度（分数 score）。该方法使用一组不同级别的噪声扰动数据，并训练单个网络同时估计所有噪声级别下的分数，以解决分数估计不准确、生成分布与实际偏差较大等问题。

在推理采样阶段，采用退火朗之万动力学的采样策略：将当前噪声级别下的采样结果作为下一噪声级别的初始样本，从而提高生成的样本质量。分数扩散模型的流程如图 4 所示。

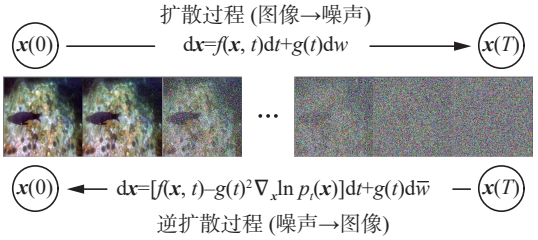


图 4　分数扩散模型的流程

Fig. 4　Process of SGM

2.3.1　分数扩散模型的扩散过程

分数扩散模型从随机微分方程的视角来设计扩散与逆扩散过程。与 DDPM 相似的是，分数扩散模型的扩散过程仍然是一个不依赖数据、也不包含可训练参数的过程：

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

式中：w 是个标准的维纳过程（又称布朗运动），具有增量独立性；$f(\boldsymbol{x},t)$ 和 $g(t)$ 是 $\boldsymbol{x}(t)$ 的漂移系数和扩散系数。随着扩散过程中噪声扰动的增加，先验分布 p_T 逐渐丧失了原始数据分布 p_0 的信息，最终导致分布趋向于 $N(0,\sigma^2)$。

2.3.2　分数扩散模型的逆扩散过程

分数扩散模型的逆扩散过程，通过求解仅与分数相关的反向时间 SDE[27-28]来进行样本生成：

$$\mathrm{d}\boldsymbol{x}=[f(\boldsymbol{x},t)-g(t)^2\nabla_x\ln p_t(\boldsymbol{x})]\mathrm{d}t+g(t)\mathrm{d}\bar{w} \tag{7}$$

式中: $\mathrm{d}\bar{w}$是逆时间方向的标准维纳过程; $\mathrm{d}t$是一个非常小的负时间步长; $p_t(\boldsymbol{x})$表示$\boldsymbol{x}(t)$的分布, 即$\boldsymbol{x}(t)\sim p_t, t\in[0,T]$。

在式 (7) 中, 未知参数仅有分数$\nabla_x\ln p_t(\boldsymbol{x})$, 因此, 文献 [25-26] 建议训练一个噪声条件分数网络$s_\theta(\boldsymbol{x},t)$来准确估计分数, 以此求解反向时间 SDE 并进行采样工作。

分数扩散模型使用预测–校正框架求解反向时间 SDE, 预测器采用离散化反向扩散采样器[25]; 校正器使用退火朗之万动力学方法[26], 分数扩散模型推理过程的算法步骤如下。

首先, 输入反向时间 SDE 的离散化步骤数 N 和校正器步骤数 M;

其次, 初始化参数, 从$p_T(\boldsymbol{x})$中采样得到$\boldsymbol{x}_N$;

再次, 设置步骤数 i 从 N−1 到 0 递减, 并依次循环执行以下操作:

1)向预测器中输入$\boldsymbol{x}_{i+1}$, 输出$\boldsymbol{x}_i$;

2)设置步骤数 j 从 1 到 M 递增, 并依次循环执行以下操作:

3)向校正器中输入$\boldsymbol{x}_j$, 输出矫正后的$\boldsymbol{x}_j$;

循环执行上述操作, 直至结束, 返回$\boldsymbol{x}_0$。

2.3.3　分数扩散模型的训练

文献 [25-26] 建议训练一个依赖时间的基于分数的模型$s_\theta(\boldsymbol{x},t)$来准确估计分数$\nabla_x\ln p_t(\boldsymbol{x})$。模型的目标函数 L 可以表示为

$$L=E_t\Big\{\lambda(t)E_{\boldsymbol{x}(0)}E_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\Big[\big\|s_\theta(\boldsymbol{x}(t),t)-\nabla_{\boldsymbol{x}(t)}\ln p_{0|t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))\big\|_2^2\Big]\Big\} \tag{8}$$

式中: $\lambda(\cdot)\in[0,T]$是一个正加权函数, $t\sim U(0,T)$, $\boldsymbol{x}(0)\sim p_0(\boldsymbol{x})$, $\boldsymbol{x}(t)\sim p_{0|t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$。

值得注意的是, 式 (8) 中运用了去噪分数匹配 (denoising score matching, DSM) 策略[29], s_θ的输入量是$(\boldsymbol{x}(t),t)$。因此, 当数据和模型容量足够时, s_θ便可以精准预测$\nabla_x\ln p_t(\boldsymbol{x})$。

综上所述, 去噪扩散概率模型和分数扩散模型都是扩散模型的重要分支, 但它们在具体实现上存在一些差异: DDPM 直接利用网络预测噪声, 通过 T 次迭代的方法进行“显式”去噪, 而分数扩散模型则采用去噪分数匹配的方法, 通过预测分数实现了“隐式”去噪。2 种模型的共同点在于, 它们都可以被视为基于噪声生成样本: 通过选择式 (8) 中的正加权函数, 可以得到与去噪扩散概率模型相等价的目标函数。因此, 从广义上来说, DDPM 可以被视为分数扩散模型的一种特例[25]。

3　扩散模型的改进

根据前文的理论基础可知, 标准的扩散模型具有生成过程缓慢、数据类型单一等问题, 并且生成结果与输入分布的似然仍有进一步提升的空间。本章从推理加速、似然优化、数据结构多样化 3 个方面总结了近年来代表性的改进工作, 并将相关改进策略进行汇总。

3.1　推理加速

由于逆扩散过程需要若干次迭代计算, 导致扩散模型的推理速度较慢。对此, 研究人员从多个角度进行改进, 包括改进采样方式、知识蒸馏、与其他模型结合等。

一部分学者通过改进采样方式来提高模型的生成速度。去噪扩散隐式模型 (denoising diffusion implicit models, DDIM)[30] 是最先提出扩散模型采样加速方法的研究之一。该方法将去噪扩散概率模型扩展到非马尔可夫情况下, 可以表示为

$$\begin{cases}q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)=q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)\displaystyle\prod_{t=2}^{T}q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0),\ t>1\\ q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)\sim N(\sqrt{\bar{\alpha}_T}\boldsymbol{x}_0,(1-\bar{\alpha}_T)\boldsymbol{I})\\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\sim N(\mu_t,\sigma_t^2\boldsymbol{I})\end{cases}$$

其中,

$$\mu_t=\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{\sqrt{1-\bar{\alpha}_t}}$$

式中σ_t^2用于控制随机噪声的比率。

DDIM 通过学习一个马尔可夫链来逆转这个非马尔可夫扰动过程, 其采样过程相当于一个特殊离散化的概率流常微分方程[25]。DDIM 的采样是确定性的, 这也降低不同样本之间的差异。同时, DDIM 在采样时可以只针对其中一些步骤跳跃采样, 在保证样本质量的同时, 显著提高模型的生成速度。

对于分数扩散模型, 标准的数值随机微分方程求解器需要大量的分数网络评估, 导致这些模型生成数据的速度缓慢。对此, Jolicoeur-Martineau 等[28] 提出了一种更有效的求解器, 可以随时间动态地调整采样步长, 在保证生成质量的同时, 提高了分数模型的生成速度。Liu 等[31] 提出扩散模型的伪数值方法 (pseudo numerical methods for diffusion models, PNDM), 采用具备非线性部分的常微分方程 (ordinary differential equations, ODE) 求解器来生成样本, 在多数情况下表现出优秀的质量和速度性能。Song 等[32] 提出的一致性模型支持快速单步生成以及高质量迭代生成, 并且可以用于零样本 (zero-shot) 数据编辑任务。

该模型最显著的特性是：在同一轨道上，任意时刻的点$(\boldsymbol{x}_T, T)$都能映射到初始时刻$(\boldsymbol{x}_\epsilon, \boldsymbol{\epsilon})$上，映射方程为

$$f_\theta(\boldsymbol{x}, t) = c_{\text{skip}}(t)\boldsymbol{x} + c_{\text{out}}(t)F_\theta(\boldsymbol{x}, t)$$

式中：$c_{\text{skip}}(\boldsymbol{\epsilon}) = 1$，$c_{\text{out}}(\boldsymbol{\epsilon}) = 0$，$F_\theta(\cdot,\cdot)$为输出维度与$\boldsymbol{x}$相同的深度神经网络。

知识蒸馏是一种加速扩散模型采样的有效策略，一些学者利用这一技术来改善模型的推理时间。渐进式快速扩散模型 (progressive fast diffusion model, ProDiff)[33] 对初始数据添加噪声，使用 N 步 DDIM 教师模型生成的梅尔频谱作为采样目标，并将其提炼为 $N/2$ 步的新模型，减少了采样时间。

Meng 等[34] 提出了一种双阶段的蒸馏方法来解决无分类器指导模型采样效率低的问题。第一阶段利用 2 个教师模型指导学生模型$\boldsymbol{x}_{\eta_1}(z_t, \omega)$的优化；第二阶段将$\boldsymbol{x}_{\eta_1}(z_t, \omega)$转化为采样步数仅为 $N/2$ 的模型$\boldsymbol{x}_{\eta_2}(z_t, \omega)$。通过修改指导权重，模型能够参照 DDIM 的采样规则[30] 生成不同质量的图像。

Shang 等[35] 首次从无训练网络压缩角度研究扩散模型加速的工作。他们将预训练量化 (pre-training quantization, PTQ) 技术引入到扩散模型的加速中，提出即插即用的 PTQ4DM，可以在无需复杂处理的情况下应用到其他的扩散模型中，提高它们的采样速度。在 PTQ4DM 中，他们在最小化量化误差的基础上为预训练模型选择每层网络的量化参数，达到性能提升的目的。

SnapFusion[36] 是一种基于 Stable-Diffusion[10] 的文本到图像生成模型，通过改进 Efficient U-Net 网络、结合无分类器指导蒸馏与原始知识蒸馏的方法，成功将移动端设备上生成 512 像素×512 像素的高质量图像的时间压缩到 2 s 内。

扩散模型与其他生成模型结合，同样是重要的研究方向。与生成网络[37]、变分自编码器[38] 相结合，能够减少采样步数，提高扩散模型的生成速度；与归一化流 (normalizing flow)[23] 模型相结合，能够提高扩散模型的表达能力、节省计算成本、实现加速。

去噪扩散生成对抗网络 (denoising diffusion GAN)[37] 是首个将扩散模型与其他生成模型相结合的工作。它使用多模态 cGAN 对每个去噪步骤建模。该方法结合扩散模型和对抗生成网络的优点，处理速度更快，且样本多样性表现更好。Pandey 等[38] 将变分自编码器整合到扩散模型中，提出的 DiffuseVAE 利用 VAE 设计新的条件参数化方法，使模型能够从低维隐空间中进行高效、可控、高保真的生成。DiffFlow[23] 是基于 SDE 的扩散归一化流模型，它通过同时训练前向和逆向神经随机微分方程使它们的分布尽可能相似。Score-Flow[39] 生成去量化样本，并将数据投射到利用归一化流转化的去量化场上。不仅提高了模型的速度，而且解决了离散数据和连续密度之间的匹配问题，提高了模型的表达能力。

3.2　似然优化

从本质上说，无论是去噪扩散概率模型还是分数扩散模型，都是一种基于似然的生成模型，似然优化的目的在于使生成的伪分布与参考分布更加接近，从而使生成的样本具有更高的质量。优化证据下界 (evidence lower bound, ELBO)[40-42]、精确似然计算[25,40]、变分差优化[41] 等方法，都是提高扩散模型的生成效果和似然性的有效手段。

在去噪扩散概率模型中，逆扩散过程的方差是与β相关的常数，并且扩散过程施加的噪声是只与时间相关的线性参数，这大大限制了模型的似然能力。相关学者提出多种改进证据下界的方法，如：参数化方差、可学习噪声、重新设计训练目标等。Nichol 等[43] 提出一种线性插值的方法，将 DDPM 逆扩散过程中的方差参数化为

$$\sum_\theta(\boldsymbol{x}_t, t) = \exp(\theta \cdot \ln\beta_t + (1-\theta)\cdot\ln\tilde{\beta}_t)$$

式中：$\tilde{\beta}_t = \dfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t$，$\beta_t = 1-\alpha_t$。

$\tilde{\beta}_t$和θ共同训练，以实现更高的似然和更快的采样速度。同时，他们还针对扩散过程设计了一种余弦噪声：

$$\bar{\alpha}_t = \frac{h(t)}{h(0)}, h(t) = \cos\left(\frac{t/T+m}{1+m}\cdot\frac{\pi}{2}\right)^2$$

式中m是一个超参数，用于控制$t=0$时的噪声规模。结果表明以上方法使模型具备了更良好的对数似然性和样本生成能力。

Austin 等[42] 介绍了一种离散去噪扩散概率模型 (denoising diffusion models in discrete state-spaces, D3PM)，其与 Improved DDPM[43] 都引入了一个新的混合损失函数，该函数将变分下界与交叉熵损失加权融合。损失函数的形式为

$$L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{VLB}}$$

式中：L_{simple}是标准去噪扩散概率模型的损失函数，L_{VLB}是变分下界损失，λ为权重。

变分扩散模型 (variational diffusion models, VDM)[44] 通过重新设计噪声规模和相关参数的方法来优化证据下界。文献 [44] 证明了除端点外的变分下界 (variational lower bound, VLB)，都可以被简化为只取决于信号噪声比$M_{\text{SNR}}(t) = \alpha_t^2/\sigma_t^2$的形

式。因此, VDM 提出了一个新的训练目标:

$$L_{\text{Diffusion}}=\frac{1}{2}E_{\boldsymbol{x}_0,\boldsymbol{\epsilon}\sim N(0,\boldsymbol{I})}\int_{M_{\min}}^{M_{\max}}\|\boldsymbol{x}_0-\tilde{x}_\theta(\boldsymbol{x}_\upsilon,\upsilon)\|_2^2\mathrm{d}\upsilon$$

式中: 上下界 $M_{\min}=M_{\text{SNR}}(T)$, $M_{\max}=M_{\text{SNR}}(1)$; $\boldsymbol{x}_\upsilon=\alpha_\upsilon\boldsymbol{x}_0+\sigma_\upsilon\boldsymbol{\epsilon}$表示扩散过程$\boldsymbol{x}_0$在$\upsilon=M_{\text{SNR}}(t)$时刻的噪声数据点; $\tilde{x}_\theta$表示模型预测的去噪数据点。

因此, 变分下界仅受端点处$M_{\text{SNR}}(t)$函数的影响, 通常情况下, 噪声时间表σ_t^2只会影响变分下界的蒙特卡罗估计器的方差, 从而提高优化速度。

对于分数网络, 它们能够定义基于分数的常微分方程, 从而进行精确的似然计算。Lu 等[40]提出了一种精确似然的计算方法, 通过一阶、二阶和三阶的分数匹配误差来联合约束分数 ODE 的负似然。基于此约束, 他们提出一种最小化损失的高阶去噪分数匹配方法, 提高了模型的似然能力。

最小化变分差异也是最大化对数似然的一种方法。隐式非线性扩散模型 (implicit nonlinear diffusion model, INDM)[41]结合了扩散过程与归一化流的优化目标, 由网络决定漂移和扩散系数。通过将原始数据编码到隐空间中, 该方法改善了模型的学习曲线, 有效提高了模型似然性。

3.3　数据类型多样化

扩散模型的生成方式决定了它主要适用于对连续型数据进行建模, 难以直接处理离散变量或非数值型数据。为了解决这一问题, 研究人员在改进模型和处理数据 2 个主要方向上进行了大量工作, 目标是能够处理不同类型的数据, 如离散、不变结构、流形结构的数据等。

早期针对扩散模型的研究大都面向连续的数据域, 针对难以处理离散数据的问题, D3PM[42]提出通过离散化高斯核或构建具有吸收状态核的扩散过程, 使扩散模型能够处理离散数据。与 D3PM 类似, 自回归扩散模型 (autoregressive diffusion model, ARDM)[45]将分类扩散模型扩展到高维离散数据, 能够很好地应用于生成语言文本、分割图以及无损压缩。Campbell 等[46]首先提出用于处理离散数据的连续时间框架。该方法将扩散和逆扩散过程统一表述为连续时间马尔可夫链, 其性能超越离散数据的离散时间方法。

Gu 等[47]首次将扩散技术引入到矢量量化数据的处理中, 提出一种用于文本到图像生成的矢量量化扩散模型 (vector quantized diffusion model, VQ-Diffusion)。该方法利用掩码或随机替换操作代替扩散模型中前向过程的高斯噪声, 消除了单向偏差和误差积累问题。其前向过程的过渡核的形式为

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=v^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{Q}_t\boldsymbol{v}(\boldsymbol{x}_{t-1})$$

式中: $\boldsymbol{v}^{\mathrm{T}}$是一个独热码 (one-hot) 列向量, $\boldsymbol{Q}_t$被称为概率过渡矩阵, $\boldsymbol{x}_t$的分类分布由向量$\boldsymbol{Q}_t\boldsymbol{v}(\boldsymbol{x}_{t-1})$确定。在此基础上, Improved VQ-Diffusion[48]提出了一种更普遍、更有效的无分类指导实现, 和一种更高质量的采样策略。在文本到图像的生成任务中, 进一步提高了 VQ-Diffusion 的生成质量。

在计算机视觉任务中, 特征提取和匹配是基础步骤。对于具有不变结构的数据, 如图 (graph) 和点云 (point cloud) 等, 考虑其特征不变性是任务的关键。因此, 在利用扩散模型完成上述任务时, 赋予模型处理数据特征不变性的能力至关重要。

置换不变 (permutation-invariant) 图的生成是深度学习领域的研究重点。Niu 等[49]首先提出利用分数扩散模型生成置换不变图, 设计了边向密集预测图神经网络 (edgewise dense prediction graph neural network, EDP-GNN) 来替代分数网络以学习分数函数。Jo 等[50]提出了一种利用 SDE 系统的图扩散过程 (graph diffusion via the system of stochastic differential equations, GDSS), 该方法对节点和邻接边的联合分布进行建模, 保证了图的置换不变性。

点云生成任务同样受到学者们的广泛关注。Luo 等[51]率先提出将扩散模型应用于点云生成任务。他们将点云的点视为热力学系统中的粒子进行扩散, 点云的生成过程被视为逆扩散过程, 在隐空间转换点云数据, 从而生成所需要的点云形状。

Xu 等[52]提出一种用于分子构象预测的模型 (geometric diffusion model, GeoDiff)。GeoDiff 利用等变马尔可夫核演化的马尔可夫链, 诱导生成了一个旋转–平移不变的分布。Xu 等表明, 具备不变性的先验核和过渡核, 可以生成保持特性不变的分子构象。

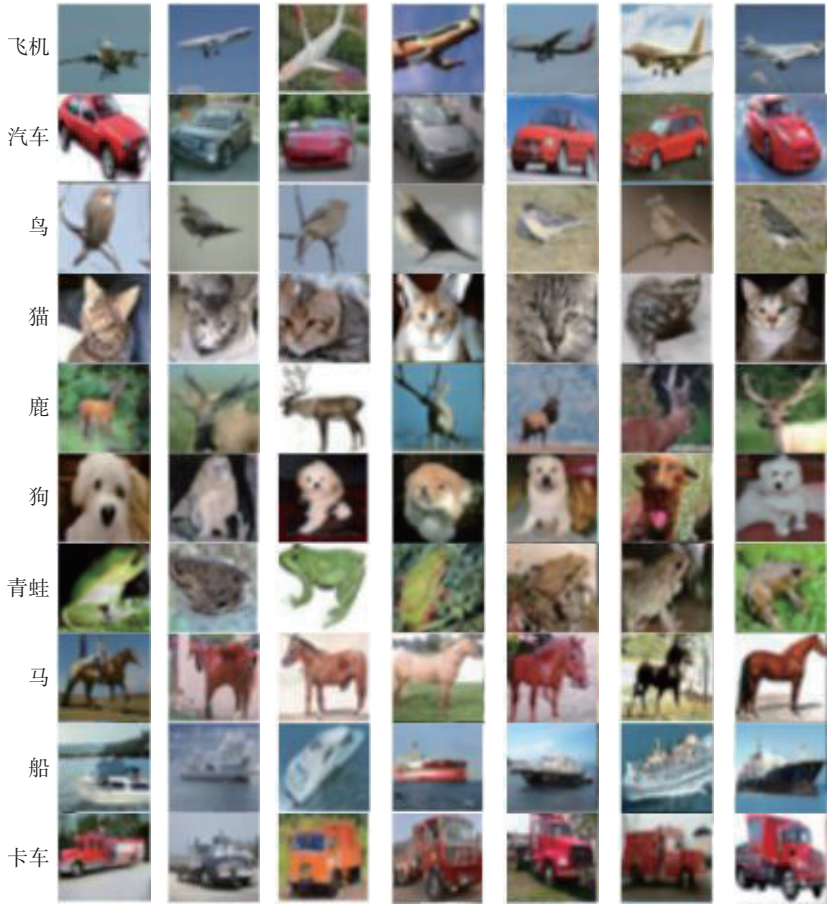
流形结构数据是地球和气候科学、蛋白质建模等学科常见的数据形式, 其分布通常由黎曼流形所描述[53]。黎曼扩散模型 (riemannian diffusion model, RDM)[54]和黎曼分数生成模型 (riemannian score-based generative model, RSGM)[53]分别将连续时间的扩散模型和分数扩散模型推广到黎曼流形的生成模型上。RDM 提出了一种变分框架, 通过最小化黎曼分数匹配, 来等同优化对数似然的变分下界。RSGM 在黎曼流形上直接定义扩散过

程，推导出流形值的反向过程，将数据的几何结构纳入模型中，使 RSGM 可以适应多种学科的流形数据。

将数据处理为较低维度的流形，并将扩散模型引入低维潜在空间中训练，可以有效地处理流形。潜在扩散模型 (latent diffusion models, LDM)[10] 和基于分数的潜在生成模型 (score-based generative modeling in latent space, LSGM)[55] 都是将扩散模型引入 VAE 产生的隐空间中进行训练。不同的是，LDM 分别训练 VAE 和扩散模型，而 LSGM 提出一种新的联合分数匹配目标，用于扩散模型和 VAE 的联合训练，推导出新的对数似然下界。Liu 等[31] 将 DDPM 视为一种流形上的微分方程，提出了扩散模型的伪数值方法 PNDM，从而在特定流形上生成样本。

扩散模型性能的量化指标可以按照采样加速、似然优化、可处理数据类型多样性等方面来分类：在采样加速方面的改进工作中，评价指标包括采样步骤数 (number of function evaluations, NFE) 和 FID(fréchet inception distance)，用以衡量加速效果和生成图像的质量及多样性，旨在展示改进工作在减少计算成本的同时，如何保持或提升图像生成的质量。在似然优化方面，评价指标为负对数似然 (negative log-likelihood, NLL)，等价于模型的训练目标，该指标衡量了模型在预测数据分布时的性能。在数据类型多样化方面，改进效果的评价指标同样包括负对数似然 (NLL) 以及可处理的数据类型范围，表明了模型的性能水平和适用范围。

为了直观地呈现典型改进工作的效果，图 5 对比了改进工作的生成效果，表 2 ~ 5 出了统一标准下基础扩散模型与典型改进工作的效果对比。表 2 给出了 3 种代表性的扩散模型基础框架的性能，表 3 ~ 5 则给出了相应改进工作的效果对比。由于部分模型无法处理图像数据或实验数据较少，未列出相关指标结果。所有改进及生成效果均基于 CIFAR-10 图像数据集，并根据改进内容选择了适当的评价指标。



(a) CD　(b) VDM　(c) DDPM　(d) PNDM　(e) LSGM　(f) iDDPM　(g) 文献[37]

图 5　在 CIFAR-10 数据集下的生成效果对比

Fig. 5　Comparison of generation effects on CIFAR-10

表 2　基础扩散模型的性能

Table 2　Performance of diffusion models

基础框架	NFE↓	FID↓	NLL↓	处理数据类型	年份
DDPM[7]	**1 000**	3.17	3.72	图像	2020
SDE[25]	2 000	**2.41**	**3.13**	图像	2020
ODE[25]	2 000	3.17	3.75	图像	2020

注: 粗体表示每一列指标最优值。

表 3　采样加速改进效果对比

Table 3　Comparison of improvement effects of sampling acceleration

具体手段	基础框架	改进工作	NFE	FID	年份
改进采样方式	DDPM	DDIM[30]	100	4.16	2021
	SDE	Gotta Go Fast[28]	1 000	2.94	2021
	ODE	PNDM[31]	125	3.46	2022
		CD[32]	**2**	2.93	2023
改进训练方式	DDPM	ProDiff[33]	16	—	2022
		PTQ4DM[35]	250	11.66	2023
		文献[34]	16	**2.78**	2023
		SnapFusion[36]	8	—	2023
与其他模型结合	SDE	DiffFlow[23]	100	14.14	2021
		Score-Flow[39]	1 000	2.86	2021
	DDPM	DiffuseVAE[38]	100	11.71	2022
		文献[37]	4	3.75	2022

注: 粗体表示每一列指标最优值。

表 4　似然优化改进效果对比

Table 4　Comparison of improvement effects of likelihood optimization

具体手段	基础框架	相关工作	NLL	年份
优化证据下界	SDE	Score-Flow[39]	2.83	2021
	DDPM	Improved DDPM[43]	2.94	2021
		D3PM[42]	3.44	2023
		VDM[44]	**2.65**	2023
精确似然计算	SDE	Score SDE[25]	2.99	2020
	ODE	文献[40]	3.27	2022
变分差优化	SDE	INDM[41]	3.09	2022

注: 粗体表示每一列指标最优值。

表 5　数据类型多样化改进效果对比

Table 5　Comparison of improvement effects of data generalization

具体手段	基础框架	相关工作	NLL	处理数据类型	年份
离散数据	DDPM	ARDM[45]	**2.68**	文本、图像	2022
		文献[46]	3.44	音频、图像	2022
		VQ-Diffusion[47]	—	矢量量化数据	2022
		Improved VQ-Diff[48]	—	矢量量化数据	2023
		D3PM[42]	3.44	文本、图像	2023
具有不变结构的数据	DDPM	文献[51]	—	点云数据	2021
		GeoDiff[52]	—	分子构象	2022
	SDE	GDSS[50]	—	图结构数据	2022
流型结构数据	SDE	RSGM[53]	—	黎曼流形	2022
		RDM[54]	—	黎曼流形	2022
	DDPM	LSGM[55]	2.87	隐空间编码	2022
	ODE	PNDM[31]	—	流形、图像	2022

注: 粗体表示每一列指标最优值。

从图 5 和表 2 ~ 5 中可以看出，在采样速度层面的对比中，多种改进策略维持了基础框架的生成质量，并将采样步骤减少了 2 ~ 3 个数量级，极大地提升了模型的可用性；在似然优化层面的对比中，研究人员的改进工作均提高了模型的生成能力；从可处理数据类型多样性层面来看，多种改进工作在优化似然能力的同时，成功将扩散模型推广到其他领域，拓展了模型的应用范围。

综上所述，研究人员们运用知识蒸馏、数据处理等通用方法，流形结构、结合其他模型等交叉方法，以及变分差优化、构建转换核、采样加速等改进模型的方法，解决了原始扩散模型存在的推理速度慢、似然性低、泛化能力差等问题，使模型的应用价值得以增强，为后续的相关研究工作提供了更可信的基础。

4　扩散模型与计算机视觉

自扩散模型提出以来，它在计算机视觉 (computer vision, CV) 领域迅速发展，利用过程的随机性，扩散模型能够在多种视觉任务上实现比 VAE 模型和 GAN 模型更为精确和细致的结果。为更好地介绍扩散模型在计算机视觉领域上的应用，本章分别从图像生成、视频生成等多种视觉任务的角度出发，介绍扩散模型的相关工作和成果。

4.1　基于扩散模型的图像生成

由于扩散过程和逆扩散过程的特殊性，模型最先被应用于图像生成任务上，相比于大规模生成对抗网络，扩散模型及其变体在图像生成方面表现出更好的质量和多样性[56-57]。Nichol 等[43] 和 Dhariwal 等[57] 发现，在使用相同的架构和训练算法时，扩散模型的生成结果比生成对抗网络更加清晰和真实。近期的优秀工作，如 DG(discriminator guidance)[58]、MDT(masked diffusion transformer)[59]，进一步提升了生成样本的质量。

此外，嵌入 (embedding) 技术与扩散模型的结合，使扩散模型在基于文本引导的图像生成任务中取得了重要进展。通过将输入的文本信息作为先验嵌入到逆扩散过程中，扩散模型能够生成符合条件的图像。

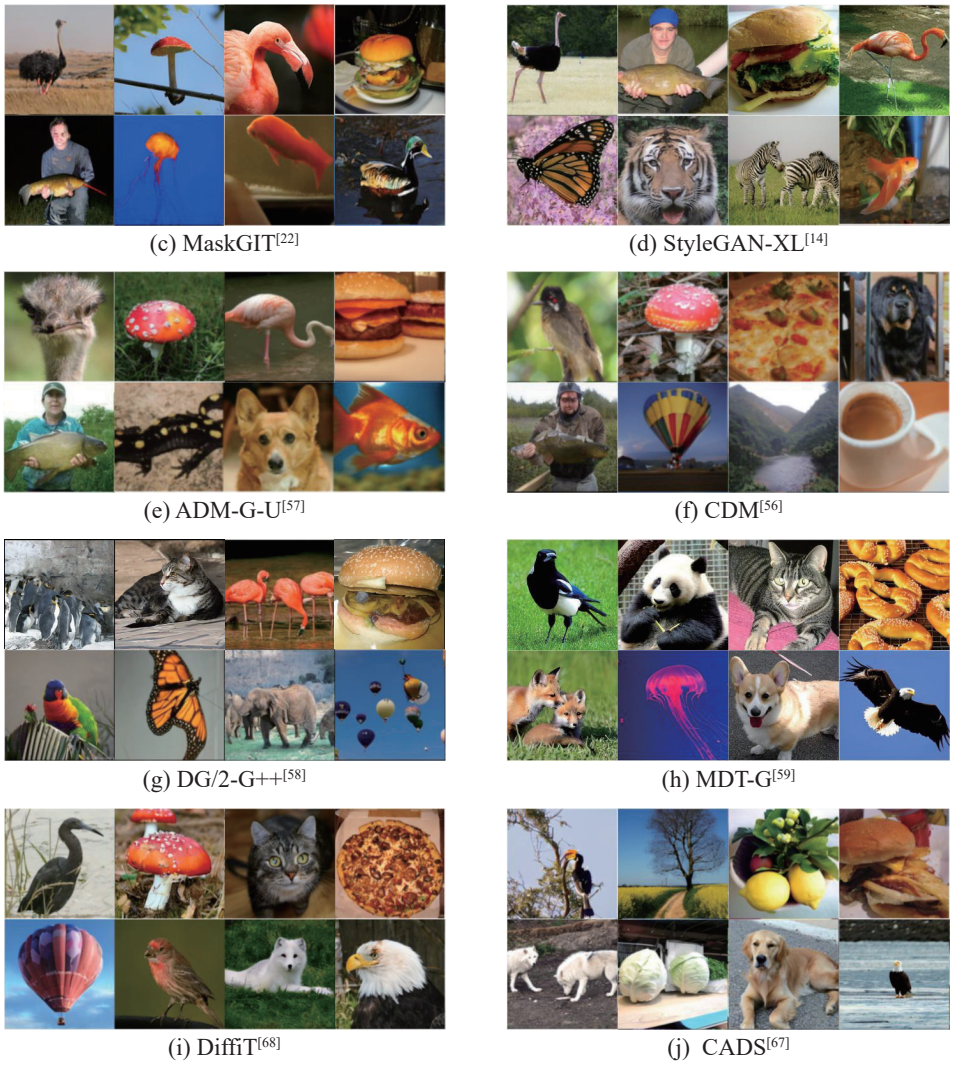
Batzolis 等[60] 证明了扩散模型在条件图像生成领域的可行性，完成了条件去噪估计器的原理性验证，为后续的理论研究提供了有效支撑。Nichol 等[61] 提出的无分类器引导策略，在评价指标和视觉效果上均超越了当时领先的 CLIP (contrastive language-image pretraining) 模型。Imagen[62] 改进了 U-Net 结构和采样方法，生成的图像更真实、细节更加丰富，取得了与 LDM[10]、VQGAN(vector quantized generative adversarial networks)-CLIP[63] 和 DALL-E2[64] 具有竞争力的效果。Imagic[65] 利用一个文本到图像的预训练扩散模型生成与输入图像和目标文本相一致的文本嵌入。通过微调扩散模型，可以改变图像中一个或多个物体的姿势和组成，并保留其原始特征[33]。

然而在上述任务中，模型在处理复杂对象和关系方面仍然存在挑战，并且需要更精细地设置引导词，这表明模型的语义理解能力仍有待提升。

除文本嵌入外，从场景图生成图像也是一项重要且具有挑战性的任务。传统方法主要是通过预测类似图像的布局来生成图像。场景图扩散模型 (diffusion-based scene graph model, SGDiff)[66] 是首个专门用于从场景图生成图像的扩散模型，它通过学习连续的场景图像嵌入来调节潜在扩散模型。相比非扩散方法，SGDiff 生成的图像能更好地表达场景图像中密集和复杂关系。此外，还有许多其他生成方法取得了最新的 SOTA(state of the arts) 效果，如：CADS(diffusion models with condition-annealed sampling)[67]、DiffiT(diffusion vision transformers)[68] 等。为了验证扩散模型的先进性，图 6 对比了不同方法的生成质量及多样性，图 7 呈现了多种模型在图像生成任务中不同评价指标下的表现，其中蓝色为 VAE 模型变体、绿色为 GAN 模型变体、黑色为扩散模型变体。

(a) VQ-VAE-2[21]

(b) BigGAN-deep[7]



(c) MaskGIT[22]　(d) StyleGAN-XL[14]

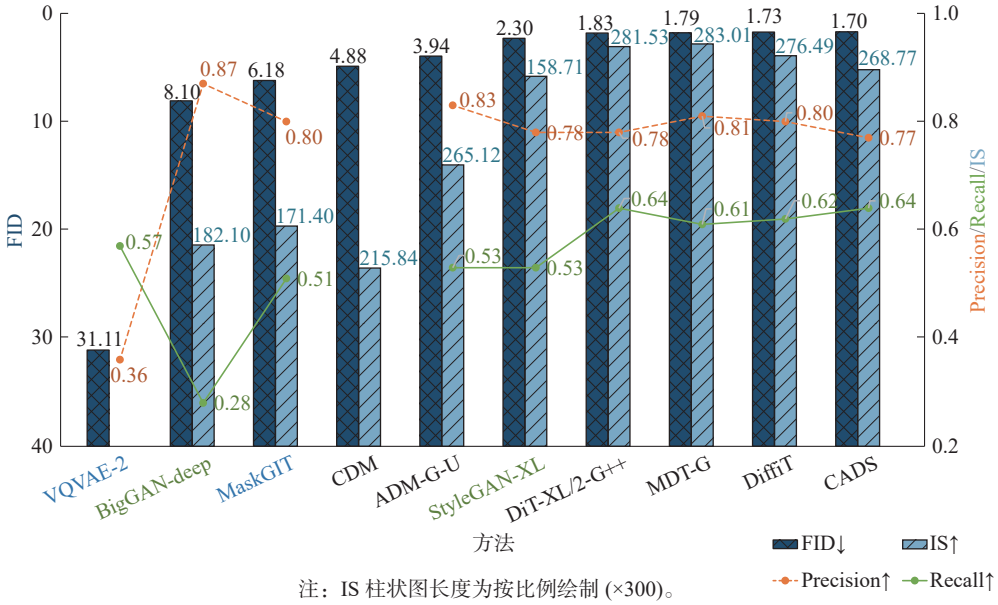
(e) ADM-G-U[57]　(f) CDM[56]

(g) DG/2-G++[58]　(h) MDT-G[59]

(i) DiffiT[68]　(j) CADS[67]

图 6　图像生成工作效果对比

Fig. 6　Comparison of image generation works effect



注：IS 柱状图长度为按比例绘制 (×300)。

图 7　图像生成工作评价指标对比

Fig. 7　Comparison of evaluation indicators for image generation works

图 6 给出了基于不同基准模型的 10 种生成方法在 ImageNet 数据集 256 像素×256 像素规模下的生成效果图。可以看出，生成式 VAE 的变体模型如 VQ(vector quantized)-VAE-2[21]、MaskGIT (masked generative image transformer)[22] 所生成的样本，在生成质量和多样性方面都仍需提升。此外，当目标图像具有人脸时，GAN 方法 (BigGAN[7]、StyleGAN[14]) 生成的样本质量较差。当样本具有近似的视觉感知质量时，Diffusion 模型比 GAN 模型包含更多的生成模式，例如放大的鸵鸟头部、站在水面上的鹰、更多的热气球等。

此外，根据图 7 的对比结果，使用扩散模型变体的生成方法在 FID、IS(inception score) 和准确率 (precision) 这 3 个指标上表现最优，并在召回率 (recall) 上达到次优水平。

综合而言，由于扩散模型本身的随机性和强大的拟合能力，其生成过程更加稳定，画面偏差较少，生成的样本也更符合现实世界的客观规律。

4.2　基于扩散模型的视频生成

在视频生成任务中，存在数据维度高和时间动态复杂性大等挑战，因此生成高质量和时间连贯的视频仍然是一个挑战。

Ho 等[69] 首次将扩散模型应用到视频生成领域。基于改进的 U-Net 架构，提出了一种应用于无条件视频生成的梯度条件生成法，并联合训练图像和视频以实现更好的帧数和分辨率。Ni 等[70] 提出了一种从图像到视频生成的潜流扩散模型 (latent flow diffusion models, LFDM)，通过在潜流空间中生成光流序列扭曲图像，生成的视频具备更好的时间动态和空间细节。Luo 等[71] 提出了一种分解扩散模型 VideoFusion。该模型将每一帧的噪声分解为基础噪声 (在所有帧之间共享) 和残余噪声 (沿时间轴变化)，并联合训练 2 个网络实现去噪。Yu 等[72] 提出一种视频扩散模型 (video probabilistic diffusion models, PVDM)。可以将视频投影为潜空间中的二维向量，从而减少生成高分辨率视频所需的资源。

然而，在视频生成任务中，扩散模型对生成帧之间的背景信息关注较少，因此常会出现帧间背景不连续的情况。

4.3　扩散模型与低级视觉任务

低级视觉任务关注从视网膜图像中提取特征并进行处理[73]。常见的任务包括：图像编辑 (editing)、翻译 (translation)、修复 (inpainting)、超分辨率重建 (super-resolution) 等。

SDEdit[8] 以风格化的图像为条件，利用扩散模型进行图像转换，并通过 SDE 先验对图像去噪。

Palette[74] 是一种基于条件扩散的统一框架。在图像修复、着色、扩充、去噪的任务上超越 GANs 模型，其生成结果具备更好的多样性和生成质量。

SynDiff[75] 是一种基于对抗性扩散模型的新方法，该方法设计了一个扩散和非扩散模块循环一致的架构，利用对抗性投影加大了采样间隔，有效地提高了医学图像转换的性能。

在图像超分辨率和增强任务中，基于扩散模型的方法表现出色。Li 等[9] 提出的 SRDiff(super-resolution with diffusion probabilistic model) 方法避免了包含 GAN 在内的标准方法的缺陷，在图像超分辨率任务中取得了优秀的视觉效果。Saharia 等[76] 提出的 SR3(image super-resolution via iterative refinement) 与朗之万动力学类似，通过一系列细化步骤将标准正态分布转换为经验分布。结合针对性改进的 U-Net 网络，使扩散模型取得了比 GAN 等方法更具真实感的生成结果。SR3+[77] 在此基础上改进，提出了一种能够利用退化自监督训练的盲超分辨率模型，并在训练和测试时进行噪声补偿增强，使得 SR3+在性能上有了更进一步的提升。

Gao 等[78] 提出的隐性扩散模型 (implicit diffusion models, IDM) 整合了隐性神经表征和去噪扩散模型，用于高保真连续图像超分辨率。在解码过程中，IDM 采用隐性神经表征学习连续分辨率表征，并引入调节网络和缩放因子的调节机制，实现了对于不同分辨率需求的平滑过渡。Rombach 等[10] 采用预训练 VAE 产生的低维潜在空间训练扩散模型，保留了模型的灵活性和生成质量，并大大降低了模型的计算要求。

在图像修复和重建任务中，扩散模型可以在保留细节信息的同时提升图像的质量，有效地去除图像中的噪声和掩码。RePaint 方法[79] 通过调节生成过程进行图像修复工作。该方法适用于任何掩码，在极端遮蔽的情况下仍然表现良好。Xie 等[80] 提出了一个基于文本和形状引导的对象修复扩散模型，通过学习局部文本描述，该模型能根据不同精度的对象掩码修复图像并保留背景信息。Wang 等[81] 提出了一种基于扩散的鲁棒性退化模型 (diffusion-based robust degradation remover, DR2)，使用预训练的扩散模型来消除退化，提高了对复杂退化的鲁棒性。结合增强模块，在双阶段的盲面部修复任务上超越了 SOTA 工作的修复质量。Fei 等[82] 提出的 GDP(generative diffusion prior) 是一种适用于多种图像任务的无监督方法。该方法在每个采样步骤下直接预测

最终样本, 并利用预测结果引导下一步采样。在图像修复和低照度增强等任务上, GDP 超过领先的无监督方法。

然而, 在低级视觉任务中, 扩散模型仍具有一定的局限性, 如采样速度慢、无法实时应用。特别是在图像修复和重建任务中, 模型的泛化能力以及适应不同类型掩码的能力需要进一步加强。

4.4　扩散模型与高级视觉任务

高级视觉任务主要关注的是知觉组织的日常功能[73]。常见的高级视觉任务有: 检测 (detection)、分割 (segmentation) 等任务。

分割任务的目的是将图像中的每个像素分配给其所属的对象类别。Baranchuk 等[83] 成功地利用扩散模型完成了语义分割任务。Brempong 等[84] 提出了基于去噪的预训练方法 DDeP(denoising pretraining for semantic segmentation), 将扩散模型与自编码器相结合, 在标签有效的语义分割任务上取得了出色的表现。Amit 等[85] 提出一种无需预训练的图像分割方法 SegDiff, 将输入图像编码后的特征图与条件图像的特征图相加合并, 形成最终的分割图像。

检测任务的目的是识别图像中的指定物体并给出标记或边界。Chen 人[86] 提出的 DiffusionDet 首次将扩散模型应用到目标检测任务。该方法将目标检测过程表述为从噪声到目标的逆扩散过程, 将随机生成的目标范围细化为准确的输出结果。

4.5　扩散模型与其他视觉任务

近年来扩散模型在视觉领域的应用日益广泛, 除了在自然图像处理中的成功应用外, 扩散模型还在点云生成、医学成像、3D 模型生成等任务中展示了巨大潜力。

点云生成任务旨在从给定的输入空间中生成一组三维点。利用扩散模型能够推断点云的缺失部分, 提高点云的完整性。Luo 等[51] 将点云的点视为热力学系统中的粒子, PVD(point-voxel diffusion)[87] 将去噪扩散模型与三维的点–体素相结合, 二者都将点云的生成过程视为逆扩散过程, 实现了从随机点云生成所需的点云形状。此外, PDR(point diffusion-refinement)[88] 是一种点扩散细化方法, 结合再融合网络细化条件 DDPM 的粗略生成, 从而提高了点云质量, 并且实现了生成点云和真实数据的逐点映射。

完整的点云信息对三维重建等任务也有着积极作用。三维重建是从二维图像或其他数据中构建出三维结构信息的过程。Wang 等[89] 提出了三维生成模型 Rodin, 该方法将三维空间特征转化为在隐空间中的连续二维特征向量。Rodin 可以利用文本或图片生成 3D 头像, 极大地提高了生成效率。Lyu 等[90] 提出了一种用于网格生成的稀疏潜点扩散模型。该方法将点云编码为具备语义特征的稀疏潜点, 并分别学习潜点的位置和特征, 该方法实现了采样速度、建模质量和可控性的全面提高。

扩散模型在医学领域也展现出了卓越的成果。GeoDiff[52] 利用扩散模型生成具有旋转–平移不变特性的分布, 并用其成功预测分子构象。Chung 等[91] 结合迭代重建思想, 使用二维 CT 图像预训练模型。该方法成功降低了数据需求量级, 并在数据部分缺失的情况下也能重建出高质量的 3D 医学图像。另外, Song 等[92] 针对预训练的分数扩散模型提出了无监督采样方法, 很好地解决了医学成像中逆向生成的问题。在从部分测量值重建医学图像时, 比其他监督学习方法的性能更好。

总的来说, 扩散模型能够适应多种类型的视觉任务, 并展现出了独特的优势。它在图像生成任务上取得了最新的 SOTA 结果, 在图像修复、超分辨率以及医学领域[52,75,91-92] 等实际任务中同样具备出色的表现。本章介绍了扩散模型在计算机视觉领域的发展, 并将相关工作汇总在表 6。

表 6　扩散模型在不同视觉任务上的应用

Table 6　Application of diffusion models to various vision tasks

具体任务	相关工作	年份
图像生成	CDM[56]、Imagen[62]、SGDiff[66]、GLIDE[61]、SR3[76]	2022
	ADM[57]、DG[58]、MDT[59]、Imagic[65]、CADS[67]、DiffiT[68]	2023
视频生成	文献[69]	2022
	LFDM[70]、VideoFusion[71]、PVDM[72]	2023
低级视觉任务 (图像编辑、修复、翻译等)	SDEdit[8]	2021
	SRDiff[9]、LDM[10]、Palette[74]、RePaint[79]、SR3[76]	2022
	SynDiff[75]、SR3+[77]、IDM[78]、DR2[81]、GDP[82]、文献[80]	2023

续表 6

具体任务	相关工作	年份
高级视觉任务(检测、分割等)	文献[83]、DDeP[84]、SegDiff[85]、DiffusionDet[86]	2022
其他视觉任务 (点云、医学、3D模型生成等)	文献[51]、PVD[87]、文献[92]	2021
	PDR[88]	2022
	Rodin[89]、文献[90]、文献[91]	2023

5　扩散模型存在的问题与展望

尽管近年来，得益于强大的分布拟合能力和模型随机性，扩散模型在相关领域得到了充分的发展，但由于其生成方式特殊，在现阶段的发展中，扩散模型仍然存在一些问题：

1) 推理速度慢[12]。与其他生成模型相比，扩散模型的生成过程需要进行多次迭代计算，推理的时间成本较高。如在自然语言处理任务中，相较于其他生成模型，扩散模型在处理大规模数据时表现欠佳，这也是扩散模型与其他任务相结合时亟待解决的问题。

2) 似然性低。扩散模型的采样过程从高斯噪声出发最终到达目标样本，在此过程中产生的信息损失和误差累积等问题，使得生成样本的质量下降，且不同样本之间存在较大的差异。应用噪声调度优化、改进 ELBO、反向方差学习、变分差优化等方法，可以提高扩散模型的生成效果和似然性。

3) 缺乏统一扩散框架。目前，扩散模型在许多独立的任务上取得了优秀的研究进展。但是，文献 [74] 指出，提出统一的扩散模型框架来解决更丰富的、相似的下游任务，是扩散模型在应用层面需要研究的问题。

4) 泛化能力有限。原始的扩散模型在训练过程中不断从高斯噪声进行采样，可能会导致模型泛化能力降低，过度适应噪声。同时，扩散模型的训练较为依赖数据集，如果数据不够充分，模型的泛化性可能会进一步变差。

除上述问题需要关注之外，在未来的研究中，或许会从以下方向改进和发展扩散模型：

1) 在更多复杂领域上的应用。尽管相关学者已经成功利用扩散模型解决诸多领域的实际问题，如金融风险预测、人工智能对话、医学图像分析等，但是在更多复杂领域的应用仍有待探索。如：自动驾驶、卫星图像处理、目标跟踪、工业应用等。

2) 学术价值与工程价值的统一。目前，针对扩散模型所做的大量的研究仅停留在理论研究阶段，这些理论成果亟待在工程上落地，体现它们的工程价值。

3) 深入理解扩散模型思想。目前，扩散模型在许多任务上取得了优秀的研究进展。然而，部分扩散模型变体要么严格遵循原始扩散模型的假设，要么尝试引入深度学习领域中广泛使用的改进技术，使其发展为另一个生成对抗网络。深入理解模型背后的物理学和统计学思想，可能有助于克服模型发展中的局限性。

4) 聚焦于扩散模型本身的研究。例如，在扩散过程中，扩散模型并不总是需要抹除数据的全部信息，且该方法生成的分布可能并不始终等效于先验分布。研究并改进扩散模型本身，可能会进一步提升模型的泛化能力和推理速度。

5) 与其他模型的结合应用。近年来，世界各国纷纷致力于发展通用人工智能 (artificial general intelligence, AGI)，其中以 ChatGPT 为代表的许多大规模语言模型在推动该领域的进步方面起到了积极的作用，国内外的高科技组织也已经结合扩散模型推出 AI 产品。因此，将扩散模型与其他模型相结合，或许能够更好地为人类社会做出贡献。

不可否认的是，扩散模型是一类优秀的生成模型。随着研究人员的持续改进，其在学术和工业领域的发展空间和应用前景将进一步被拓宽。

6　结束语

作为一类新的生成模型范式，扩散模型为图像生成任务及其相关领域提供了一种全新的研究思路。它通过可控的扩散过程和合理的逆扩散过程，对随机噪声进行建模，并生成具有更高质量和更强泛化能力的建模结果。不仅如此，由于其出色的性能指标，扩散模型在学术端和应用端备受关注。其多种变体不仅对广大用户和创作者产生了重要影响，也为人工智能产业带来了全新的商业化机会，推动了深度学习研究的可持续性发展。

为帮助学者更好地了解扩散模型，对扩散模型进行了综述。首先对比了多种生成模型的优劣，介绍了去噪扩散概率模型和基于分数的扩散模型的数学原理。随后，从扩散模型面临的挑战

出发，介绍了近年来学者在这些方面的工作，同时也介绍了扩散模型在计算机视觉领域上的应用。最后，探讨了扩散模型所存在的问题，并提出了未来可能的发展趋势，旨在促进扩散模型在应用和技术发展中的进一步提升。

参考文献：

[1] CRIMINISI A, PEREZ P, TOYAMA K. Object removal by exemplar-based inpainting[C]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison: IEEE, 2003: II.

[2] WANG Zhou, YU Yinglin, ZHANG D. Best neighborhood matching: an information loss restoration technique for block-based image coding systems[J]. IEEE transactions on image processing, 1998, 7(7): 1056−1061.

[3] TURK M A, PENTLAND A P. Face recognition using eigenfaces[C]//Proceedings of 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Maui: IEEE, 1991: 586−591.

[4] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015: 1530−1538.

[5] KINGMA D P, WELLING M. Auto-encoding variational Bayes[EB/OL]. (2013−12−20)[2023−12−27]. https://arxiv.org/abs/1312.6114.

[6] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139−144.

[7] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 6840–6851.

[8] MENG Chenlin, HE Yutong, SONG Yang, et al. SDEdit: guided image synthesis and editing with stochastic differential equations[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2022: 1−33.

[9] LI Haoying, YANG Yifan, CHANG Meng, et al. SRDiff: Single image super-resolution with diffusion probabilistic models[J]. Neurocomputing, 2022, 479: 47−59.

[10] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10674−10685.

[11] 闫志浩, 周长兵, 李小翠. 生成扩散模型研究综述[J]. 计算机科学, 2024, 51(1): 273−283.
YAN Zhihao, ZHOU Zhangbing, LI Xiaocui. Survey on generative diffusion model[J]. Computer science, 2024, 51(1): 273−283.

[12] YANG Ling, ZHANG Zhilong, SONG Yang, et al. Diffusion models: a comprehensive survey of methods and applications[J]. ACM computing surveys, 2024, 56(4): 1−39.

[13] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[C]//International Conference on Learning Representations. New Orleans: OpenReview.net, 2019: 1–35.

[14] SAUER A, SCHWARZ K, GEIGER A. StyleGAN-XL: scaling StyleGAN to large diverse datasets[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. Vancouver: ACM, 2022: 1–10.

[15] 曹锦纲, 李金华, 郑顾平. 基于生成式对抗网络的道路交通模糊图像增强[J]. 智能系统学报, 2020, 15(3): 491−498.
CAO Jingang, LI Jinhua, ZHENG Guping. Enhancement of blurred road-traffic images based on generative adversarial network[J]. CAAI transactions on intelligent systems, 2020, 15(3): 491−498.

[16] 严浙平, 曲思瑜, 邢文. 水下图像增强方法研究综述[J]. 智能系统学报, 2022, 17(5): 860−873.
YAN Zheping, QU Siyu, XING Wen. An overview of underwater image enhancement methods[J]. CAAI transactions on intelligent systems, 2022, 17(5): 860−873.

[17] 姜义, 吕荣镇, 刘明珠, 等. 基于生成对抗网络的人脸口罩图像合成[J]. 智能系统学报, 2021, 16(6): 1073−1080.
JIANG Yi, LYU Rongzhen, LIU Mingzhu, et al. Masked face image synthesis based on a generative adversarial network[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1073−1080.

[18] 毕晓君, 潘梦迪. 基于生成对抗网络的机载遥感图像超分辨率重建[J]. 智能系统学报, 2020, 15(1): 74−83.
BI Xiaojun, PAN Mengdi. Super-resolution reconstruction of airborne remote sensing images based on the generative adversarial networks[J]. CAAI transactions on intelligent systems, 2020, 15(1): 74−83.

[19] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing: PMLR, 2014: 3057−3070.

[20] 张冀, 曹艺, 王亚茹, 等. 融合 VAE 和 StackGAN 的零样本图像分类方法[J]. 智能系统学报, 2022, 17(3): 593−601.
ZHANG Ji, CAO Yi, WANG Yaru, et al. Zero-shot image classification method combining VAE and StackGAN[J]. CAAI transactions on intelligent systems, 2022, 17(3): 593−601.

[21] RAZAVI A, OORD A V D, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: 1–15.

[22] CHANG Huiwen, ZHANG Han, JIANG Lu, et al. MaskGIT: masked generative image transformer[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11305−11315.

[23] ZHANG Qinsheng, CHEN Yongxin. Diffusion normaliz-

ing flow[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 16280–16291.

[24] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: JMLR, 2015: 2246–2255.

[25] SONG Yang, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2020: 1–23.

[26] SONG Yang, ERMON S. Generative modeling by estimating gradients of the data distribution[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: 1–36.

[27] ANDERSON B D O. Reverse-time diffusion equation models[J]. Stochastic processes and their applications, 1982, 12(3): 313−326.

[28] JOLICOEUR-MARTINEAU A, LI Ke, PICHÉ-TAILLEFER R, et al. Gotta go fast when generating data with score-based models[EB/OL]. (2021−05−28)[2023−12−27]. https://arxiv.org/abs/2105.14080.

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

[30] SONG Jiaming, MENG Chenlin, ERMON S. Denoising diffusion implicit models[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2021: 1–22.

[31] LIU Luping, REN Yi, LIN Zhijie, et al. Pseudo numerical methods for diffusion models on manifolds[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2022: 1–24.

[32] HARRISON G. Consistency models[M]//Next Generation Databases. Berkeley: Apress, 2015: 127−144.

[33] HUANG Rongjie, ZHAO Zhou, LIU Huadai, et al. ProDiff: progressive fast diffusion model for high-quality text-to-speech[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 2595–2605.

[34] MENG Chenlin, ROMBACH R, GAO Ruiqi, et al. On distillation of guided diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 14297−14306.

[35] SHANG Yuzhang, YUAN Zhihang, XIE Bin, et al. Post-training quantization on diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 1972−1981.

[36] LI Yanyu, WANG Huan, JIN Qing, et al. SnapFusion: text-to-image diffusion model on mobile devices within two seconds[EB/OL]. (2023−06−01)[2023−12−27]. https://arxiv.org/abs/2306.00980.

[37] XIAO Zhisheng, KREIS K, VAHDAT A. Tackling the generative learning trilemma with denoising diffusion GANs[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2022: 1–28.

[38] PANDEY K, MUKHERJEE A, RAI P, et al. DiffuseVAE : efficient, controllable and high-fidelity generation from low-dimensional latents[J]. Transactions on machine learning research, 2022: 1–39.

[39] SONG Yang, DURKAN C, MURRAY I, et al. Maximum likelihood training of score-based diffusion models[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 1415–1428.

[40] LU Cheng, ZHENG Kaiwen, BAO Fan, et al. Maximum likelihood training for score-based diffusion ODEs by high-order denoising score matching[C]//International Conference on Machine Learning. New York: PMLR, 2022: 14429–14460.

[41] KIM D, NA B, KWON S J, et al. Maximum likelihood training of implicit nonlinear diffusion models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 32270–32284.

[42] AUSTIN J, JOHNSON D D, HO J, et al. Structured denoising diffusion models in discrete state-spaces[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 17981–17993.

[43] NICHOL A, DHARIWAL P. Improved denoising diffusion probabilistic models[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8162–8171.

[44] KINGMA D P, SALIMANS T, POOLE B, et al. Variational diffusion models[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 21696–21707.

[45] HOOGEBOOM E, GRITSENKO A A, BASTINGS J, et al. Autoregressive diffusion models[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2022: 1–23.

[46] CAMPBELL A, BENTON J, DELIGIANNIDIS G, et al. A continuous time framework for discrete denoising models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 28266–28279.

[47] GU Shuyang, CHEN Dong, BAO Jianmin, et al. Vector quantized diffusion model for text-to-image synthesis [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10686−10696.

[48] TANG Zhicong, GU Shuyang, BAO Jianmin, et al. Improved vector quantized diffusion models[EB/OL]. (2022−05−31)[2023−12−27]. https://arxiv.org/abs/2205.16007.

[49] NIU Chenhao, SONG Yang, SONG Jiaming, et al. Per-

mutation invariant graph generation via score-based generative modeling[C]//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistic. Palermo: PMLR, 2020: 4474–4484.

[50] JO J, LEE S, HWANG S J. Score-based generative modeling of graphs via the system of stochastic differential equations[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 10362–10383.

[51] LUO Shitong, HU Wei. Diffusion probabilistic models for 3D point cloud generation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2836−2844.

[52] XU Minkai, YU Lantao, SONG Yang, et al. GeoDiff: a geometric diffusion model for molecular conformation generation[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2022: 1–19.

[53] BORTOLI V De, MATHIEU É, HUTCHINSON M, et al. Riemannian score-based generative modelling[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 2406–2422.

[54] HUANG C W, AGHAJOHARI M, BOSE A J, et al. Riemannian diffusion models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 2750–2761.

[55] VAHDAT A, KREIS K, KAUTZ J. Score-based generative modeling in latent space[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 11287–11302.

[56] HO J, SAHARIA C, CHAN W, et al. Cascaded diffusion models for high fidelity image generation[J]. The journal of machine learning research, 2022, 23(1): 2249−2281.

[57] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 8780–8794.

[58] KIM D, KIM Y, KWON S J, et al. Refining generative process with discriminator guidance in score-based diffusion models[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: JMLR, 2023: 16567–16598.

[59] GAO Shanghua, ZHOU Pan, CHENG Mingming, et al. MDTv2: masked diffusion transformer is a strong image synthesizer[EB/OL]. (2023−03−25)[2023−12−27]. https://arxiv.org/abs/2303.14389.

[60] BATZOLIS G, STANCZUK J, SCHÖNLIEB C B, et al. Conditional image generation with score-based diffusion models[EB/OL]. (2021−11−26)[2023−12−27]. https://arxiv.org/abs/2111.13606.

[61] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 16784−16804.

[62] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 36479−36494.

[63] CROWSON K, BIDERMAN S, KORNIS D, et al. VQGAN-clip: open domain image generation and Editing with Natural language guidance[C]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 88−105.

[64] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022−04−13)[2023−12−27]. https://arxiv.org/abs/2204.06125.

[65] KAWAR B, ZADA S, LANG O, et al. Imagic: text-based real image editing with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 6007−6017.

[66] YANG Ling, HUANG Zhilin, SONG Yang, et al. Diffusion-based scene graph to image generation with masked contrastive pre-training[EB/OL]. (2022−11−21)[2023−12−27]. https://arxiv.org/abs/2211.11138.

[67] SADAT S, BUHMANN J, BRADLEY D, et al. CADS: unleashing the diversity of diffusion models through condition-annealed sampling[EB/OL]. (2023−10−26)[2023−12−27]. https://arxiv.org/abs/2310.17347.

[68] HATAMIZADEH A, SONG Jiaming, LIU Guilin, et al. DiffiT: diffusion vision transformers for image generation[EB/OL]. (2023−12−04)[2023−12−27]. https://arxiv.org/abs/2312.02139.

[69] HO J, SALIMANS T, GRITSENKO A, et al. Video diffusion models[EB/OL]. (2022−04−07)[2023−12−27]. https://arxiv.org/abs/2204.03458.

[70] NI Haomiao, SHI Changhao, LI Kai, et al. Conditional image-to-video generation with latent flow diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 18444−18455.

[71] LUO Zhengxiong, CHEN Dayou, ZHANG Yingya, et al. VideoFusion: decomposed diffusion models for high-quality video generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 10209–10218.

[72] YU S, SOHN K, KIM S, et al. Video probabilistic diffusion models in projected latent space[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 18456−18466.

[73] VAN DER HELM P A. Simplicity in vision: a multidisciplinary account of perceptual organization[M]. Cambridge: Cambridge University Press, 2014: 1–8.

[74] SAHARIA C, CHAN W, CHANG Huiwen, et al. Palette: image-to-image diffusion models[C]//Special Interest Group on Computer Graphics and Interactive Techniques

Conference Proceedings. Vancouver: ACM, 2022: 1–10.
[75] OZBEY M, DALMAZ O, DAR S U H, et al. Unsupervised medical image translation with adversarial diffusion models[J]. IEEE transactions on medical imaging, 2023, 42(12): 3524−3539.
[76] SAHARIA C, HO J, CHAN W, et al. Image super-resolution via iterative refinement[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(4): 4713−4726.
[77] SAHAK H, WATSON D, SAHARIA C, et al. Denoising diffusion probabilistic models for robust image super-resolution in the wild[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 6840−6851.
[78] GAO Sicheng, LIU Xuhui, ZENG Bohan, et al. Implicit diffusion models for continuous super-resolution[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 10021−10030.
[79] LUGMAYR A, DANELLJAN M, ROMERO A, et al. RePaint: inpainting using denoising diffusion probabilistic models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11451−11461.
[80] XIE Shaoan, ZHANG Zhifei, LIN Zhe, et al. SmartBrush: text and shape guided object inpainting with diffusion model[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 22428−22437.
[81] WANG Zhixin, ZHANG Ziying, ZHANG Xiaoyun, et al. DR2: diffusion-based robust degradation remover for blind face restoration[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 1704−1713.
[82] FEI Ben, LYU Zhaoyang, PAN Liang, et al. Generative diffusion prior for unified image restoration and enhancement[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 9935−9946.
[83] BARANCHUK D, RUBACHEV I, VOYNOV A, et al. Label-efficient semantic segmentation with diffusion models[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2021: 1–15.
[84] BREMPONG E A, KORNBLITH S, CHEN Ting, et al. Denoising pretraining for semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022: 4174−4185.
[85] AMIT T, NACHMANI E, SHAHARBANY T, et al. SegDiff: image segmentation with diffusion probabilistic models[EB/OL]. (2021−12−01)[2023−12−27]. https://arxiv.org/abs/2112.00390.
[86] CHEN Shoufa, SUN Peize, SONG Yibing, et al. DiffusionDet: diffusion model for object detection[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 19773−19786.
[87] ZHOU Linqi, DU Yilun, WU Jiajun. 3D shape generation and completion through point-voxel diffusion[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 5806−5815.
[88] LYU Zhaoyang, KONG Zhifeng, XU Xudong, et al. A conditional point diffusion-refinement paradigm for 3D point cloud completion[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2021: 1–24.
[89] WANG Tengfei, ZHANG Bo, ZHANG Ting, et al. RODIN: a generative model for sculpting 3D digital avatars using diffusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 4563−4573.
[90] LYU Zhaoyang, WANG Jinyi, AN Yuwei, et al. Controllable mesh generation through sparse latent point diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 271−280.
[91] CHUNG H, RYU D, MCCANN M T, et al. Solving 3D inverse problems using pre-trained 2D diffusion models [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 22542−22551.
[92] SONG Yang, SHEN Liyue, XING Lei, et al. Solving inverse problems in medical imaging with score-based generative models[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2021: 1–18.

作者简介：

管凤旭，副教授，博士，主要研究方向为无人系统自主控制、机器视觉目标检测与跟踪、计算机控制及应用。获授权发明专利近 20 项，发表学术论文 40 余篇，出版教材 5 部。E-mail：guanfengxu@hrbeu.edu.cn。

张涵宇，硕士研究生，主要研究方向为图像去雾、计算机视觉。E-mail：zhy875329435@163.com。

路斯棋，硕士研究生，主要研究方向为水下图像处理、计算机视觉。E-mail：lusiqi9803@163.com。