基于生成式人工智能的图像视频生成方法综述及展望

张璐瑶, 杨帅, 汪文靖, 高翔, 刘家瑛*

(北京大学王选计算机研究所 北京 100871)
(liujiaying@pku.edu.cn)

摘 要: 视觉生成在艺术、娱乐等创意领域, 以及医学影像和数字出版等关键领域起到越来越重要的作用, 生成式人工智能在视觉生成方面的发展有望彻底改变人们与视觉数据的交互方式. 文中首先介绍深度学习时代下经典的生成模型框架, 根据视觉生成输入条件的不同, 重点梳理了近年来几类重要的图像生成模型和方法, 包括无条件图像生成、类别生成图像、文字生成图像和图像转换方法, 并介绍了它们在图像编辑方面的应用; 然后根据训练数据要求的不同, 详细总结近年来以扩散模型为代表的视频生成和编辑模型及相应的优缺点; 再介绍目前经典的图像生成和视频生成数据集和常用的评价标准; 最后总结现阶段视觉生成面临的数据获取、推理效率、长视频生成、视频可控生成、安全等方面的挑战, 展望未来的研究方向.

关键词: 生成式人工智能; 扩散模型; 图像生成; 视频生成
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00281

AIGC-Based Image and Video Generation Method: A Review

Zhang Luyao, Yang Shuai, Wang Wenjing, Gao Xiang, and Liu Jiaying*

(*Wangxuan Institute of Computer Technology, Peking University, Beijing* 100871)

Abstract: Visual generation plays an increasingly important role across diverse fields, from creative domains such as art and entertainment to critical areas such as medical imaging and digital publishing. The development of AIGC in visual generation will potentially revolutionize our interactions with visual data. First, this paper introduces the classical generative models in the deep learning era. Then, based on different input conditions, several important image generation models developed in recent years including unconditional image generation, class-to-image generation, text-to-image generation and image-to-image translation are highlighted, along with their applications in image editing. Next, a detailed summary of video generation and editing models, especially video diffusion models, is provided. And their advantages and disadvantages based on the requirements of training data are outlined. Additionally, this paper reviews the classic datasets for image and video generation and the commonly used evaluation metrics. Finally, the paper summarizes the challenges faced in visual generation in terms of data collection, inference efficiency, long video generation, controllable video generation and security, and discusses potential future research directions.

Key words: artificial intelligence generated content; diffusion models; image generation; video generation

收稿日期: 2024-06-06; 修回日期: 2025-01-21. **基金项目**: 国家自然科学基金(62332010, 62471009); CCF-腾讯犀牛鸟基金(RAGR20240118). **张璐瑶**(1992—), 女, 博士研究生, 主要研究方向为图像生成、图像修复; **杨帅**(1991—), 男, 博士, 助理教授, 博士生导师, CCF 会员, 主要研究方向为图像生成、计算机视觉; **汪文靖**(1997—), 女, 博士, 主要研究方向为图像生成、图像增强; **高翔**(1992—), 男, 博士, 主要研究方向为图像生成、图像翻译; **刘家瑛**(1983—), 女, 博士, 副教授, 博士生导师, CCF 杰出会员, 论文通信作者, 主要研究方向为智能影像计算、计算机视觉.

1 视觉生成模型发展概览

随着人工智能以空前的速度不断发展, 生成式人工智能(artificial intelligence generated content, AIGC)重新定义了视觉内容的生成、制作和编辑过程, 彻底改变了我们感知和创造视觉内容的方式. 如图 1 所示, 早期算法通常使用图像匹配或人工设计的生成规则合成简单的纹理或结构; 随后, 深度学习时代提出变分自编码器(variational auto-encoder, VAE)[1]和生成对抗网络(generative adversarial network, GAN)[2], 它们能够学习如人像、室内场景等典型的图像分布; 当前, 涌现出了大型的扩散模型[3]来生成图像和更具挑战性的视频, 其生成的质量之高甚至是人也难以区分真假. 从传统算法到 VAE 和 GAN, 再到扩散模型, AIGC 正在向更高分辨率、更丰富内容和更可控的生成方向演变.

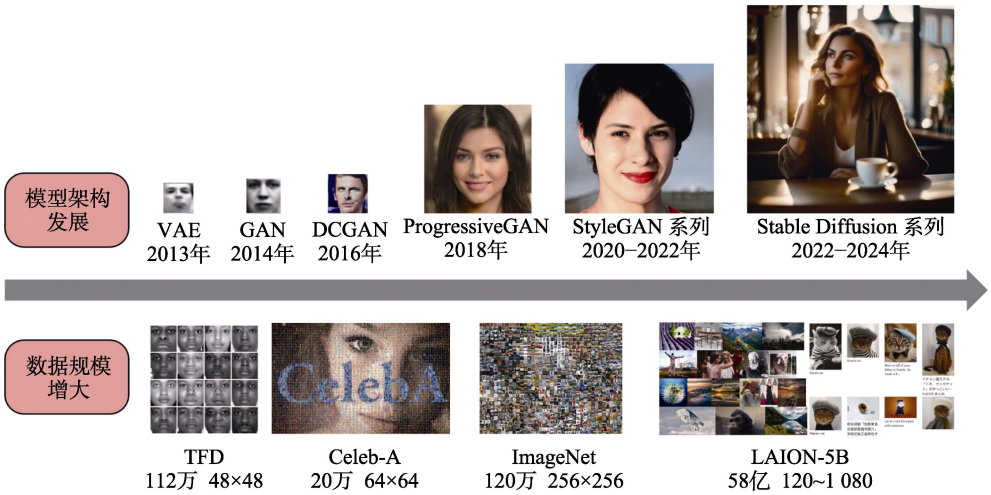

图 1 图像生成模型架构和图像生成数据集的发展概览

近年来, AIGC 在视觉生成方面的快速发展主要有 2 个因素:

(1) 行业内训练数据资源不断丰富, 从分辨率有限的 64×64 像素的 CelebA 数据集[4]的 20 万幅图像, 到分辨率为 256×256 像素的 ImageNet[5]的 120 万幅图像, 再到拥有惊人的 58 亿幅不同分辨率图像的 LAION-5B 数据集[6], 这些质量不断提高、规模不断增加的数据, 显著地增强了模型学习到的表征能力.

(2) 算力和模型架构及规模的发展也是提升模型表征能力的另一个重要因素. 早期的小模型只能建模简单和典型的图像分布, 如人脸、汽车; 之后, 设计精良的 BigGAN[7]支持 1 000 种不同的类别; 当前, 扩散模型的参数规模是 BigGAN 的几十到上百倍. 随着模型本身架构的进步和参数量的提升, 模型已经有能力学习更复杂的数据分布. 生成模型已经从每个模型只能生成单一图像主题进化为每个模型能生成任何主题, 甚至是涌现出真实世界不存在的全新视觉概念.

高质量数据集的提出和模型设计的改进共同促成了 AIGC 在视觉生成方面的跨越式发展, 新的模型层出不穷, 但目前尚缺乏对现有最新模型和方法沿革的详细梳理工作. 本文聚焦近几年扩散模型在图像和视频生成方面的前沿方法和应用, 并结合 VAE, GAN 和自回归模型[8]等经典模型架构, 总结生成模型的发展脉络, 探讨视觉生成的研究趋势.

本文从生成模型概览、图像生成模型、视频生成模型、数据集和评价标准 4 个方面展开综述. 在生成模型中, 介绍主流的 VAE, GAN, 自回归模型和扩散模型 4 类生成模型; 在图像生成模型中, 以输入条件为线索, 介绍无条件、标签引导、文字引导和图像引导的图像生成模型和方法; 在视频生成模型中, 按照训练时的数据需求, 介绍数据驱动、单样本和零样本的视频生成与编辑模型.

2 生成模型

生成模型的范式为基于训练集中的真实观测数据, 学习构建出由简单已知分布(如高斯分布)到目标数据分布的转换映射, 允许在推理阶段从已知分布中采样并生成符合目标数据分布的新样本. 当前, 主流的生成模型包括 VAE[1], GAN[2], 自回归模型[8]和扩散模型[3]. 图 2 所示为这 4 类模型的架构和生成流程的示意图.

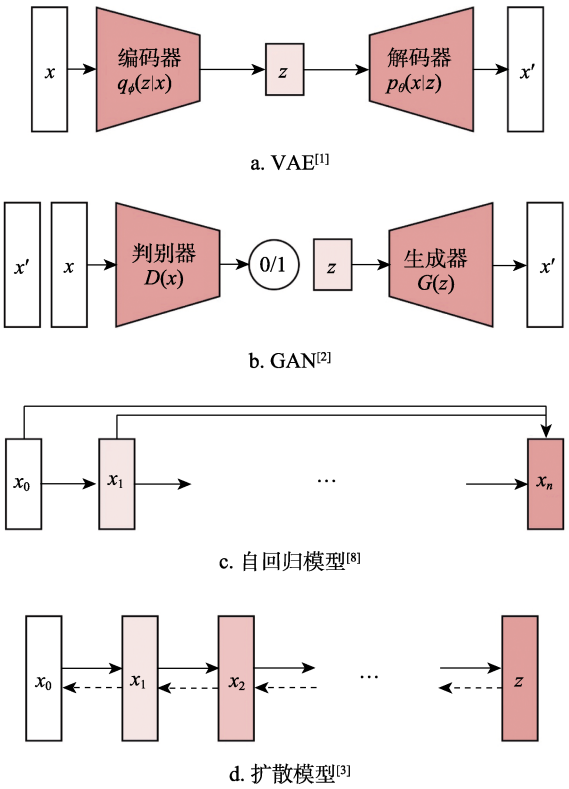

图 2　主流生成模型示意图

2.1　VAE

VAE[1]由处理数据降维问题的自编码器演变而来，在其基础上，通过对编码隐空间的分布进行显式约束实现生成功能. 以 x 表示真实数据样本，VAE 引入隐变量 z 表示似然函数 $\log p(x)$，并推导出最终表达形式为

$$\log p(x)=-D_{\mathrm{KL}}\left(p(z|x),p(z)\right)+E_{p(z|x)}\log\left(p(z|x)\right) \quad (1)$$

其中，$D_{\mathrm{KL}}\left(p_1,p_2\right)$ 表示 p_1 与 p_2 这 2 个分布之间的 KL 散度(Kullback-Leibler divergence). VAE 用编码器建模由数据样本 x 到隐变量 z 的映射，即条件分布 $p(z|x)$；用解码器建模由隐变量 z 到生成样本 $\hat{x}$ 的映射，即条件分布 $p(\hat{x}|z)$. 式(1)表明，极大化似然函数等价于最小化隐变量的后验分布 $p(x|z)$ 和先验分布 $p(z)$ 之间的KL距离(第1项)，并最大化生成样本 $\hat{x}$ 与数据样本 x 的一致性(第 2 项). 进一步，VAE 假设 $p(z)$ 为标准高斯分布 $\mathcal{N}(0,\mathcal{I})$，$p(z|x)$ 为各项独立的高斯分布 $\mathcal{N}\left(\mu,\sigma^2\mathcal{I}\right)$，并通过编码器预测出其均值 μ 和标准差 σ. 为实现以可导的方式从编码器输出的隐空间分布 $\mathcal{N}\left(\mu,\sigma^2\mathcal{I}\right)$ 中采样隐变量 z，VAE 使用了重参数化技巧

$$z=\mu+\sigma\varepsilon,\ \varepsilon\sim\mathcal{N}(0,\mathcal{I}).$$

将上述假设分布代入式(1)，可得出最终目标函数 L_{VAE}，其为 KL 散度损失 L_{KL} 和重构损失 L_{recon} 的加权和，即 $L_{\mathrm{VAE}}=L_{\mathrm{KL}}+CL_{\mathrm{recon}}$. 其中，

$$L_{\mathrm{KL}}=D_{\mathrm{KL}}\left(\mathcal{N}\left(\mu,\sigma^2\mathcal{I}\right),\mathcal{N}(0,\mathcal{I})\right)=\frac{1}{2}\left(-\log\sigma^2+\mu^2+\sigma^2-1\right);$$

$$L_{\mathrm{recon}}=\|\hat{x}-x\|_2;$$

常数 C 表示加权系数. 经模型训练，隐空间分布 $\mathcal{N}\left(\mu,\sigma^2\mathcal{I}\right)$ 不断趋近于标准高斯分布 $\mathcal{N}(0,\mathcal{I})$，从而允许推理阶段直接从标准高斯分布采样隐变量 z 解码出生成样本.

继标准 VAE 之后，一些相关变体对其做出了改进. 重要性加权自编码器[9]消除了对后验分布 $p(z|x)$ 服从高斯分布的假设的依赖，提高了模型的生成能力. 条件 VAE[10]和半监督 VAE[11]将 VAE 拓展到有条件生成和半监督学习研究中. InfoVAE[12]引入数据样本与隐变量之间的互信息，提出了更一般化的似然函数. 向量量化 VAE(vector quantized VAE, VQ-VAE)[13]将 VAE 推广为离散化的隐变量，为 VAE 建立了新的应用典范. 目前，VAE 类模型被广泛用于生成自然图像[14]、人脸图像[15]和自然语言文本[16]等.

2.2　GAN

与 VAE 显式地优化似然函数不同，GAN[2]通过构建相互博弈的生成器和判别器隐式地学习真实数据样本的分布. 其中，判别器试图区分真实样本和生成器伪造的假样本，生成器试图生成以假乱真的假样本来迷惑判别器，二者相互对抗、交替优化、互为促进、协同进化，最终使生成器生成数据的分布不断地接近真实数据的分布.

以 $G\left(z;\theta_g\right)$ 表示生成器，其中，$z\sim p_z$ 表示输入的随机噪声，θ_g 表示生成器参数；以 $D\left(x;\theta_d\right)$ 表示判别器，x 表示输入的真实数据样本或生成器伪造的假样本，θ_d 表示生成器参数. GAN 通过交替地优化 G 与 D 来不断地提升生成器的生成能力和判别器的鉴伪能力，其过程可由优化问题描述为

$$\min_G\max_D V(G,D)=E_{x\sim p_r}\left[\log D(x)\right]+E_{z\sim p_z}\left[\log\left(1-D(G(z))\right)\right] \quad (2)$$

其中，p_r 表示真实数据的分布；p_z 表示生成器生成数据的分布. 优化判别器时，固定生成器，训练判别器将真实数据 $x\sim p_r$ 分类为 1(给出高得分)、生成数据 $x=G(z)\sim p_g$ 分类为 0(给出低得分). 优化生成器时，固定判别器，训练生成器提高生成样

本的质量, 使判别器给出尽量高的得分.

式(2)中, 对判别器的优化目标 $\max\limits_{D} V(G,D)$ 关于 $D(x)$ 求导, 并令导数为 0, 可以求得判别器的理论最优解

$$D^*(x)=\frac{p_r(x)}{p_r(x)+p_g(x)} \tag{3}$$

而生成器的理论最优解应使生成数据的分布恰好为真实数据的分布, 即 $p_g=p_r$. 此时, 判别器恒有 $D^*(x)$=0.5, 即完全无法区分真伪数据样本, 双方达到纳什均衡.

然而, GAN的训练非常困难, 容易出现梯度消失和模式崩溃问题. 将式(3)的最优判别器形式代入式(2), 可以推导出

$$V(G,D^*)=2D_{\mathrm{JS}}(p_r,p_g)-\log 4 \tag{4}$$

其中, $D_{\mathrm{JS}}(p_r,p_g)$ 表示 p_r 与 p_g 这 2 个分布之间的 JS(Jensen-Shannon)散度. 式(4)表明, 判别器的优化本质是度量真实数据分布与生成数据分布之间的 JS 距离; 相应地, 生成器的优化过程则是最小化该 JS 距离, 进而实现 p_g 向 p_r 的对齐. 然而, JS 散度的定义决定了当 2 个分布没有重叠部分时将恒为常数 log2, 导致训练生成器出现梯度消失问题.

针对上述 GAN 训练的不稳定问题, LSGAN (least square GANs)[17]用最小二乘损失替换交叉熵损失, 规避了使用 Sigmoid 运算带来的梯度传递困难问题; WGAN(Wasserstein GANs)[18]通过权重裁剪技术对判别器施加 Lipschitz 约束, 将潜在优化目标中的JS散度替换为Wasserstein距离, 缓解了梯度消失和模式崩溃问题; WGAN-GP(WGAN with gradient penalty)[19]通过在优化目标中对判别器的梯度进行正则化实现 Lipschitz 约束, 进一步提升了 GAN 的收敛性能; SNGAN(spectral normalized GANs)[20]通过对判别器的参数进行谱归一化实现 Lipschitz 约束, 取得了更快的收敛速度.

除了稳定性的改进, 大量方法对 GAN 在方法论上进行了完善. cGAN(conditional GANs)[21]将 GAN 推广到条件引导的数据生成任务; SAGAN (self-attention GANs)[22]引入自注意力机制, 提升了 GAN 的生成能力; LAPGAN(Laplacian pyramid GANs)[23]采用多尺度的层次化生成器网络架构提升图像生成质量; SinGAN[24]实现了基于单幅图像训练 GAN. 除了图像生成, GAN 还被广泛用于各类计算机视觉及数据挖掘任务, 表现出极为广泛的应用场景和巨大的应用潜力.

2.3 自回归模型

自回归模型以自回归的方式建模数据生成过程, 即根据时序序列中的历史数据预测生成当前时刻的数据, 递归地生成出完整的序列数据

$$p(x)=\prod_{i=1}^{n}p(x_i\mid x_{<i}).$$

自回归建模是自然语言生成模型、特别是大型语言模型的重要组成部分. 通过将图像转化为像素级或图像块级的序列数据, 自回归模型在图像生成任务也取得了成功的应用.

像素循环神经网络(pixel recurrent neural network, PixelRNN)[25]基于长短记忆网络(long short-term memory, LSTM)[26]以自回归的方式逐像素地生成图像, 包含 Row LSTM 和 Diagonal BiLSTM 这 2 种特殊的 LSTM 层. 前者通过构建三角上下文实现同行像素隐状态的并行计算, 降低了计算开销; 后者弥补前者带来的上下文不全的问题, 增强了数据建模能力.

考虑 LSTM 的复杂性和学习长期依赖的计算成本问题, PixelRNN[25]使用卷积神经网络处理图像数据的自回归生成或补全, 为卷积核设计掩码蒙版来保证对每个像素的生成只参考在其之前的所有像素的信息. 进一步, Gated PixelCNN[27]解决了 PixelRNN 存在的盲点问题, 并引入门控卷积层提高模型的性能. 受到在自然语言建模中大获成功的 Transformer[28]的启发, CogView[29]将图像序列化, 并用 Transformer 进行自回归图像生成, 实现了高质量的文生图内容创作. 由于自回归模型固有的推理时间开销大的弊端, 因此模型在视觉生成中的应用相对有限. 为了解决这一问题, VAM (visual autoregressive modeling)[30]采用多尺度机制模仿人感知图像的逻辑, 根据从整体到细节的多尺度顺序逐渐生成序列, 并在自回归的每步并行地预测当前尺度的信息, 在模型参数和图像尺寸相当的情况下, 大大减少了自回归生成图像的时间开销.

2.4 扩散模型

扩散模型[3]是当下最受关注的生成模型, 由一个逐级添加高斯噪声的前向扩散过程和一个逐级预测并消除噪声的反向去噪过程组成. 给定输入图像 $x_0\sim q(x_0)$, 前向扩散过程为一个 T 步马尔可夫链, 按照状态转移条件分布

$$q(x_t\mid x_{t-1})=\mathcal{N}\left(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)\mathcal{I}\right),$$

不断地向原始图像增添高斯噪声. 其中, $\alpha_t\in(0,1)$, $t=1,2,\cdots,T$ 为预设的噪声控制参数, 且有 $\alpha_t\geqslant\alpha_{t+1}$.

记 $\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$, 由高斯分布的性质可直接由 x_0 计算出 x_t 为

$$q(x_t|\ x_0)=\mathcal{N}\left(x_t;\sqrt{\bar{\alpha}_t}x_0,(1-\bar{\alpha}_t)\mathcal{I}\right).$$

通过重参数化技巧可得

$$x_t=\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon,\ \varepsilon\sim\mathcal{N}(0,\mathcal{I}) \tag{5}$$

当 $\bar{\alpha}_T\to 0$ 时, $q(x_T\mid x_0)\approx\mathcal{N}(0,\mathcal{I})$, 从而实现将原始数据映射到标准高斯噪声空间.

逆向去噪过程对应于按后验概率 $q(x_{t-1}\mid x_t)$ 逐级去噪, 构建由高斯噪声空间到图像空间的逆向映射. 由于 $q(x_{t-1}\mid x_t)$ 无法直接求解, 因此通过可求解的条件后验概率 $q(x_{t-1}\mid x_t,x_0)$ 进行逆向去噪, 即

$$q(x_{t-1}\mid x_t,x_0)=\mathcal{N}\left(x_{t-1};\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t\mathcal{I}\right) \tag{6}$$

其中,

$$\tilde{\mu}_t(x_t,x_0)=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0+\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t,$$

$$\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t;$$

$\beta_t=1-\alpha_t$. 去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)[3]将扩散模型应用于图像生成并获得了良好的效果. 为了解决式(6)中 x_0 在推理阶段不可知的问题, DDPM 在训练阶段构建噪声估计网络 ε_θ, 根据 x_t 和时间步 t 预测用于时间步 t 前向扩散的高斯噪声 ε, 即

$$L_\varepsilon=\left\|\varepsilon-\varepsilon_\theta(x_t,t)\right\|_2.$$

在推理阶段, DDPM 使用 $\varepsilon_\theta(x_t,t)$ 代替 ε, 并通过式(5)推得 x_0 的近似为

$$y_\theta(x_t,t)=\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t-\sqrt{1-\bar{\alpha}_t}\varepsilon_\theta(x_t,t)\right),$$

并用该近似值代替式(6)中的 x_0 进行逆向推理.

扩散模型以简洁的优化目标和稳定的训练过程迅速获得大量关注. 分类器梯度引导技术[31]的提出, 使扩散模型在图像生成质量上首次超越 GAN, 成为生成模型家族中的主流范式; 随后提出的无分类器引导技术[32], 在保证生成质量的同时极大地精简了扩散模型的训练流程; 去噪扩散隐概率模型(denoising diffusion implicit model, DDIM)[33], DPM-Solver[34]等采样优化器将原始 DDPM 的采样步数减少到原来的十分之一, 大幅提升了扩散模型的推理速度; Palette[35]构建条件扩散模型基本框架, 弥补了扩散模型在有条件生成任务的空白. 这些工作逐步完善扩散模型的方法论体系, 促进了其在多种生成类视觉任务(如超分辨率[36]、图像修补[37]、图像翻译[38]、图像着色[39])上的应用与突破. 随着图文多模态技术的进步和 Transformer 的发展与成熟, 涌现出以 DALL-E2[40], GLIDE[41]和 Imagen[42]为代表的大规模文生图扩散模型, 它们允许用户通过自然语言提示创作出高分辨率、高质量且富有创意的图像, 将 AI 内容创作推向了全新的高度. 特别地, 潜在扩散模型(latent diffusion model, LDM)[43]的提出将扩散模型由像素空间迁移到低维特征空间, 大幅提高了文生图大模型的训练效率和推理速度. 以 SDXL[44]和 DiT(diffusion transformer)[45]为代表的研究工作改进扩散模型去噪网络的基础架构, 进一步提高了图像生成的质量. 当前, 关于扩散模型基础架构与应用推广的研究已成为计算机视觉领域的关注焦点, 并不断引领 AIGC 方法与技术的快速迭代.

2.5　发展脉络

视觉生成模型的发展经历多个关键的技术突破, 形成了从 VAE[1]到 GAN[2]再到扩散模型[3]的演变过程. (1) VAE 作为早期的生成模型, 将图像生成问题转化为概率建模问题, 尽管在生成质量上存在模糊的问题, 但为后续的生成技术奠定了基础. (2) 2014 年, GAN 通过生成器与判别器的对抗训练提升图像的质量, 但其训练具有不稳定性, 给高分辨率的图像生成带来挑战. VAE 和 GAN 都是学习从高斯噪声分布到图像分布的映射关系并具有一定的互补性, 它们的思想常常被用于生成和编辑模型的设计和训练中. (3) 自回归模型则是显式地建模图像的概率分布, 逐像素生成图像. 通过图像编解码技术降低自回归的搜索空间, Transformer[28]架构的提出更好地建模像素点之间的上下文关系, 使得自回归模型的图像生成质量得到有效的提升, 但其生成速度较慢, 限制了实际应用. (4) 随着扩散模型的出现, 图像生成进入了一个新的阶段. 2022 年迎来了扩散模型一个成功的模型实践 LDM[43], 通过逐步去噪和文本引导生成高质量图像; 2024 年是基于纯 Transformer 的 DiT[45]大放异彩的一年. 与 LDM 相比, DiT 更善于捕捉图像中的上下文关系, 同时具有更强的可伸缩性, 使得大规模的数据和更多的参数能带来更高的生成质量. 同时 DiT 能够更好地捕捉视频中的复杂时序关系, 给视频生成应用带来新的机遇, 尤其在长视频生成和细节一致性方面表现突出. 综上所述, 视觉生成技术的演变从 VAE 的概率建模到 GAN 的对抗训练, 再到扩散模型的去噪过程, 推动了图像

视频生成质量的不断提升.

3 图像生成

本节以输入条件为线索, 介绍无条件图像生成、标签引导的图像生成、文字引导的图像生成(文生图)和图像引导图像生成(图像转换、图生图)的代表性模型和方法. 图像生成与编辑模型分类如图 3 所示.

3.1 无条件图像生成

无条件图像生成任务不依赖于任何特定条件, 通常仅通过随机噪声生成图像, 是后续条件图像生成的基础. 从早期仅能生成简单数据, 到当前能够生成高清图像, GAN 在无条件图像生成的发展中发挥了至关重要的作用. 下面介绍 DCGAN(deep convolutional GANs)[46], ProgressiveGAN[47]和 StyleGAN[48]系列 3 个具有代表性的 GAN 模型, 它们分别从基础网络结构、训练范式和模型设计的角度, 显著地提升了图像生成效果.

3.1.1 DCGAN

与最早提出的 GAN 使用多层感知机构建生成器和判别器[1]相比, DCGAN[46]将全卷积神经网络结构应用到 GAN 中. DCGAN 的判别器由卷积层、批标准化层和 LeakyReLU 激活层组成, 不使用池化层; 生成器由转置卷积层、批标准化层和 ReLU 激活层组成. 与多层感知机相比, 卷积神经网络具有更加强大的表征能力并且参数量更小, 受益于此, DCGAN 能在更高分辨率的图像上训练, 取得更优的生成质量.

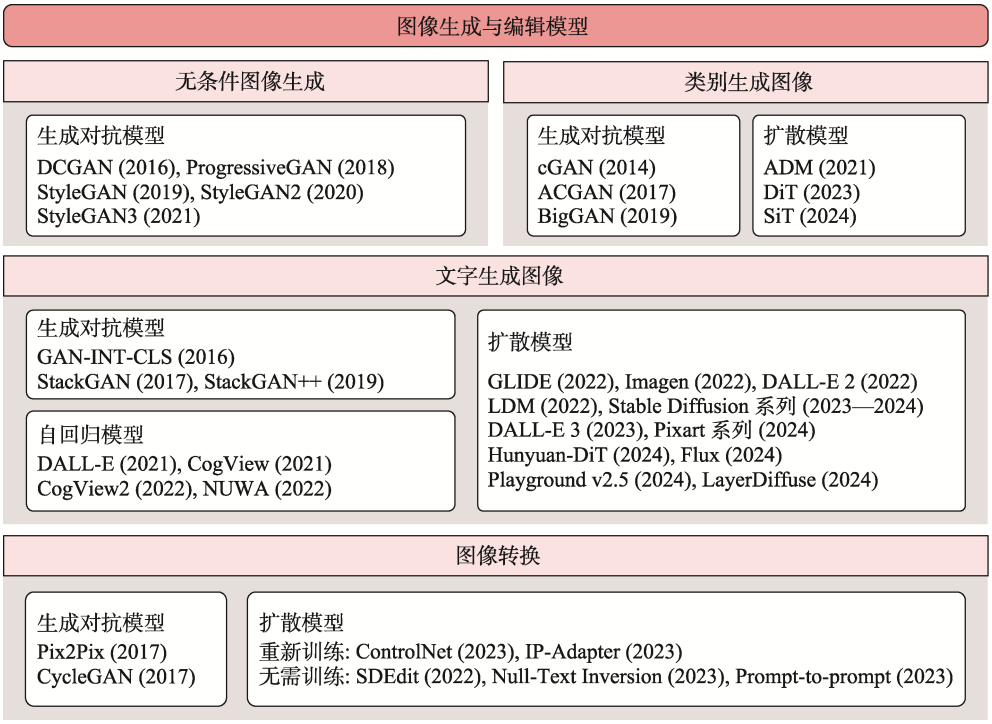

图 3 图像生成与编辑模型分类

3.1.2 ProgressiveGAN

ProgressiveGAN[47]通过逐步加大用于训练生成器和判别器的图像分辨率, 得到更稳定的训练过程和更多元的生成结果. ProgressiveGAN 从 4×4 分辨率的隐变量开始训练生成器和判别器, 并逐步提高分辨率和添加可学习模块, 直到分辨率达到 1 024×1 024 像素; 生成器和判别器采用同样的结构并同时增加可学习模块. 这种思路可以理解为先让网络学习生成粗粒度特征的能力, 再通过逐步加大分辨率, 使得网络逐步学习生成细粒度细节的能力, 这种训练范式不仅能缩短训练时间, 还可以提高训练稳定性.

除此之外, ProgressiveGAN 还提出小批量标准差技术, 以提升生成质量和多元性. 在训练过程中, ProgressiveGAN 首先抽取一个训练批次中的小批量数据并计算其平均标准差; 然后将该标准差复制并派分到每一个小批量数据的位置上, 形成一个单通道的特征; 再通过将该特征与判别器最后一个模块的起始层原始特征拼接起来, 判别器除了判断图像的质量外, 还判断图像之间的差异,

促使生成器生成多样化的图像. ProgressiveGAN 还使用正则化抑制判别器和生成器的病态竞争问题.

3.1.3　StyleGAN 系列

继 ProgressiveGAN 之后, 2019 年提出的 StyleGAN[48]进一步提升了图像生成质量. 在网络的每层, StyleGAN 通过自适应实例归一化(adaptive instance normalization, AdaIN)[49]将可学习的高层视觉信息先验注入到生成器中, 创建可解释的隐变量空间, 提供了风格融合的交互式生成方式, 为后续众多图像生成工作奠定了坚实的基础.

StyleGAN 的生成器包含 4 个模块: (1) 可学习的风格隐变量输入; (2) 将风格编码映射到可注入的高层视觉信息先验隐变量空间的映射网络; (3) AdaIN 风格融合模块; (4) 负责保持生成多样性的随机噪声输入. StyleGAN 以 4×4 分辨率的固定特征为第 1 层进行图像的生成, 每经过一个网络模块会上采样一倍.

传统的 GAN 训练生成器将隐变量空间直接映射到图像空间, 导致训练数据集的风格信息隐变量只是全体风格信息隐变量空间的稀疏子集, 在此基础上, 难以拥有可融合、可调制的隐变量空间供交互式生成. StyleGAN 采用一个由 8 层多层感知机组成的映射网络, 首先对固定分布的风格信息隐变量进行变换, 将其映射到可学习的分布稠密的中间隐变量空间中; 中间隐变量空间的风格信息可以通过线性融合的方式进行特征融合, 不会造成风格信息的损失, 有效地促进了风格信息之间的解耦合.

受到风格迁移相关工作的启发, StyleGAN 使用 AdaIN 进行风格信息的注入, 即利用上述中间隐变量空间的风格信息作为输入, 学习一个仿射变换的参数, 并对生成器的中间特征进行变换, 对生成结果进行显式的控制. 负责保持生成结果多样性的随机噪声通过额外的注入分支加入到生成器中, 该随机噪声只会影响生成结果的高频细节, 而大体风格信息则由上述中间隐变量空间的风格信息确定.

在 StyleGAN 问世 1 年之后, StyleGAN2[50]被提出, 进一步解决了 StyleGAN 中由于使用 AdaIN 而产生的伪影问题. AdaIN 在仿射变换前对特征进行均值和方差的标准化, 以减弱特征之间的相对差异性, 而这种差异性度量了特征之间的语义信息. 为了保持这些语义信息, StyeGAN 学习在局部区域生成极值来保留均值之间的差异, 但同时也会在该区域生成伪影. 为了解决该问题, StyleGAN2 对 AdaIN 风格注入模块进行了如下修改: (1) 移除对均值的标准化, 只对方差进行标准化; (2) 将随机噪声的加入移动至风格注入之后.

StyleGAN2 不仅改进网络结构, 还提出创新的生成高分辨率图像的方式, 并指出, 简单地逐级增加生成图像分辨率的方式会造成细节纹理的局部固化, 降低视觉效果; 此外提出残差生成的方式, 即将不同的低分辨率生成图像经过上采样后与高分辨率生成图像相加, 逐级迭代以提高分辨率, 这种做法不需要在训练过程中改变网络结构, 即可逐步生成高分辨率的图像.

2021 年被提出的 StyleGAN3[51], 进一步解决了 StyleGAN 和 StyleGAN2 的生成结果中高频细节不完善的问题. 由于 StyleGAN 和 StyleGAN2 在生成人脸图像时生成的高频细节主要依据周边像素信息而不是人脸表面信息, 对于移动中的人脸生成造成困难, 因此 StyleGAN3 进一步重新设计了生成器中的上采样和 ReLU 激活部分, 以确保生成结果的连续等变性.

3.2　类别生成图像

通常, 随机生成图像在实际应用中缺少实用性. cGAN[21]引入物体类别标签等额外条件控制生成过程, 除增加可控性外, 引入附加信息还可以降低训练难度, 显著地提高生成样本的质量. cGAN 向生成器 G 与判别器 D 中引入了条件 y, 即

$$\min_G\max_D E_{x,y\sim p_r}\left[\log D(x,y)\right]+ E_{z\sim p_z,y\sim p_y}\left[\log\left(1-D\left(G(z,y),y\right)\right)\right].$$

ACGAN(auxiliary classifier GANs)[52]在判别器中也考虑了条件 y, 其中, 损失函数分为无条件和有条件 2 部分, 前者采用式(2); 后者衡量给定类别下的概率分布

$$\min_G\max_D E_{x,y\sim p_r}\left[\log P\left(C=y\mid x\right)\right]+ E_{z\sim p_z,y\sim p_y}\left[\log P\left(C=y\mid G(z,y)\right)\right].$$

BigGAN[7]提出一个综合的 GAN 框架, 能够生成分辨率为 256×256 像素的 1 000 类 ImageNet 图像. BigGAN 增大了训练批大小和模型通道数, 不仅向输入层也向网络中间各层加入噪声和生成条件控制; 此外, 在判别器中计算条件向量和最后一层特征向量的内积, 显著地提升了条件生成质量.

Transformer[28]被证明具有良好的可扩展性. DiT[45]使用 Transformer 替换 U-Net. 由于 Transformer 需要标记(Token)作为输入, 因此 DiT 将输入映射为块并应用位置编码, 然后这些 Token 被一组 Transformer 模块处理, 重新由块解码回输入形状.

为了引入类别标签、扩散模型时间步等控制信息, DiT 中提出自适应层归一化(adaptive layer normalization, AdaLN)机制, 自适应地根据条件信息预测层归一化中的均值和方差. 与交叉注意力机制或直接将生成条件与输入合并相比, AdaLN 在计算效率上更具优势. 还有的研究使用 Mamba[53], RWKV[54]等架构替代 U-Net 和 Transformer, 构建扩散模型的主干网络.

扩散生成过程可以描述为基于流的插值过程

$$x_t = a_t x_* + \sigma_t \varepsilon.$$

其中, x_* 表示数据, ε 表示噪声, a_t 随 t 递减, σ_t 随 t 递增.

与扩散模型相比, 插值框架能够灵活地连接 2 个数据分布. SiT[55]通过深入研究时间步选取的离散与连续策略、优化对象、插值参数选择, 以及采样采用常微分方程(ordinary differential equations, ODE)[56]还是随机微分方程(stochastic differential equations, SDE)[57], 提出了线性插值的方案, 实现了在相同模型架构条件下更优秀的生成效果.

3.3 文字生成图像

自然语言是人类最常用的表达方式, 利用文字生成图像非常符合用户的使用习惯和需求. 如图 4 所示, 只需要输入一段英文描述, 模型就能生成对应的图像, 并且最新的模型生成的内容细节和分辨率都得到显著提升. 然而, 要实现这一技术, 需要克服自然语言建模和图像生成 2 个方面的挑战.

图 4 不同模型在文生图任务上的结果

早期的文生图工作主要围绕 GAN 展开. 基于 cGAN, GAN-INT-CLS[58]根据文本描述生成分辨率为 64×64 像素的鸟类和花卉图像. 后续工作引入物体位置坐标作为额外的输入条件, 能够生成分辨率为 128×128 像素的图像. 为了无须额外输入条件也能生成分辨率为 256×256 像素的高清图像, StackGAN[59]采用两阶段策略: 首先生成 64×64 像素的低分辨率图像; 然后利用第 1 阶段的结果作为第 2 阶段的输入条件, 生成分辨率为 256×256 像素的图像. 在第 1 阶段, StackGAN 预处理文本, 利用全连接层将文本映射为均值与方差, 并根据此均值和方差从正态分布中采样. 这是因为训练集中的文本数量较少, 如果直接将文本作为生成条件, 文本隐变量会过于稀疏. 由于第 1 阶段已经引入随机性, 第 2 阶段不需要使用随机噪声向量作为输入. 在 StackGAN 的基础上, StackGAN++[60]进一步采用多阶段框架, 包含树状结构的多个生成器和判别器, 分别对应不同尺度的图像; 同时进行了有条件和无条件的生成训练.

随着 Transformer 在自然语言处理领域取得巨大成功, 如何将这些模型应用于开放域的文生图任务吸引了研究人员的广泛关注. DALL-E[61]是基于 Transformer 的文生图自回归模型, 它将文本和图像编码为统一的 Token 流进行处理; 为了降低计算复杂度, 采用离散 VAE 将分辨率为 256×256 像素的图像压缩为 32×32 像素的 Token 表示. CogView[29]进一步构建了支持中文的文生图自回归模型. CogView2[62]中引入分层设计, 提升了自回归模型的生成速度. NÜWA[63]中提出一个通用的 3D Transformer 框架, 同时涵盖语言、图像和视频, 可用于多种输入与条件组合的视觉生成任务.

GLIDE[41]首次将扩散模型应用于文生图任务, 其中比较了 2 种不同的文本引导策略: 基于文本编码器的分类器梯度引导, 以及无分类器引导, 实验结果表明, 无分类器引导可以获得更加逼真并符合文本描述的结果. GLIDE 在训练文生图的同时联合训练文本编码器, Imagen[42]则采用预训练且参数冻结的大型语言模型作为文本编码器, 减少了训练代价, 并且文本编码器可以预先在图像-文本数据或纯文本语料库上进行训练. 通常, 纯文本语料库的规模远大于图像-文本对数据, 使得大型语言模型能够接触到更加丰富和分布广泛的文本, 获得更强的自然语言建模能力. 为了生成分辨率为 1 024×1 024 像素的高清图像, Imagen 采用级联的模型架构, 首先生成 64×64 像素的图像, 然后将其上采样至 256×256 像素和 1 024×1 024 像素分辨率.

对齐文本与图像表征是文生图等跨模态应用的基础之一. 多模态感知模型 CLIP(contrastive language-image pre-training)[64]经过大量文本-图像配对数据的训练, 利用对比学习方法实现了文本与图像在统一表征空间中的对齐; 在此基础上, DALL-E2[40]提出一种逆向操作 CLIP 图像编码器的方法, 首先依据 CLIP 的文本编码生成相应的图像编码, 然后通过图像解码器将图像编码还原为实际图像. 其中, 生成图像编码的过程可借助自回

归模型或扩散模型完成; 图像解码器则采用三阶段的扩散模型, 首先将图像编码转换为 64×64 像素的图像, 然后逐步上采样至 256×256 像素和 1 024×1 024 像素分辨率.

针对扩散模型在像素空间的训练与推理复杂度极高的问题, LDM[43]采用编码器将 256×256×3 的图像由像素压缩为 32×32×4 隐变量, 并在隐空间进行扩散模型的训练与推理, 生成隐变量后通过解码器恢复为图像, 大幅降低了扩散模型的计算开销; 此外, 引入交叉注意力作为将文本等控制信息融入模型引导图像生成的方式.

基于 LDM, 诞生了知名的开源文生图模型 Stable Diffusion 系列[43]. Stable Diffusion 1.x 版本能够生成分辨率为 512×512 像素的图像, 而 Stable Diffusion 2.x 版本支持的分辨率上升到了 768×768 像素. SDXL[44]使用的参数量为 Stable Diffusion 1 和 Stable Diffusion 2 的 3 倍, 其针对模型训练时数据随机裁切造成的画面主体偏移, 以及图像不同宽高比难以适配深度学习架构并行计算的问题, 提出将图像裁切位置与原宽高比作为生成条件提供给模型. SDXL turbo[65]结合模型蒸馏与对抗学习, 将大型生成模型的知识浓缩到更小的模型中. Stable Diffusion 3[66]采用多模态 DiT 架构, 通过 AdaLN 和输入 Token 叠加 2 种方式引入控制条件, 由于图像和文本特征是完全不同的, 对这 2 种模态采用 2 组独立的权重, 即图像和文字有其独立的 Transformer 层, 但同时进行注意力操作互相感知.

得益于 Transformer 在大规模视觉任务上更好的扩展性, DiT 架构被广泛使用. PixArt-α[67]设计了三阶段的训练, 依次进行像素依赖学习、文本-图像对齐学习, 以及高分辨率与美学图像生成, 逐步优化模型的生成能力; 在此基础上, PixArt-Σ[68]通过整合更高质量的数据提升生成效果, 并提出一种压缩注意力层键和值 Token 的机制, 显著地提高了效率; PixArt-δ[69]引入隐变量一致性模型(latent consistency model, LCM)提高推理速度; 面向细粒度英语与中文理解, Hunyuan-DiT[70]构建多模态大模型对图像标签进行精细调整; Flux.1①将模型的参数量拓展到 12 字节, 并引入旋转位置嵌入和并行注意力层提升模型性能. 其中, 非商用开源版本 Flux.1 dev 对引导进行蒸馏, 推理时不需要负向提示词, 只使用正向提示词以及一个引导蒸馏超参, 降低了推理代价; 可商用开源版本 Flux.1 schnell 进一步实施步数蒸馏, 在小步数时也可以生成高质量的结果, 显著地提升了图像的生成速度.

一些工作侧重于优化图像生成的某一个方面. DALL-E 3[71]重点关注文本对图像内容的控制. 针对在互联网上爬取的数据中, 文本描述和图像数据的噪声过多, 文本描述经常包括不完整或与图像内容无关的信息的问题, DALL-E 3 构建了一个图像描述模型, 对训练数据进行重新标注, 不仅生成描述图像主体的短标签, 还生成描述图像背景、风格、颜色、内嵌文字等信息的长标签; 此外, 还能与 ChatGPT[72]结合, 将图像生成融入对话交互中.

Playground v2.5[73]关注生成结果的美学质量, 在训练时, 采用 EDM 噪声机制增强生成结果的色彩与对比度, 同时增加不同宽高比的数据并采用更加平衡的宽高比采样机制; 此外, 还使用类似于大型语言模型中的监督微调方法, 增强对人像生成的能力. LayerDiffuse[74]向 LDM 引入透明图层, 能够根据文本创建多个独立图层, 支持用户对生成结果更加灵活的控制.

除上述研究工作外, Midjourney②和 InvokeAI③等应用产品为用户提供专业工具, 进一步推动了文生图研究的发展与 AIGC 社区的建设.

3.4　图像转换问题

图像转换问题以给定的图像作为限制条件, 将输入图像转换为输出图像. 在计算机视觉、图像处理领域中, 大量问题可以被建模为图像转换问题, 如黑白照片上色修复问题、根据语义分割图生成相应真实图像、根据低光照图生成正常光照图像等问题. 图像转换问题分为监督和无监督 2 类: 监督任务指训练集中输入图像和输出图像之间具有逐像素对齐的性质, 可以训练深度神经网络在大规模数据下直接拟合出较为精确的映射关系; 无监督任务的训练集中, 输入图像域和输出图像域之间没有像素级别的对应关系. 随着深度学习的发展, 图像转换方法由最初的端到端网络演化至使用 GAN, 到当前利用条件扩散模型进行可控生成, 生成质量逐步提高.

3.4.1　面向监督任务的 pix2pix

2017 年, pix2pix[75]中提出创新性的网络结构

① https://blackforestlabs.ai/announcing-black-forest-labs

② https://www.midjourney.com/home

③ https://invoke-ai.github.io/InvokeAI

和损失函数处理监督的图像转换问题.

在生成器结构方面, 为了应对之前生成模型基于编码器-解码器结构导致的模糊和细节缺失问题, pix2pix 采用 U-Net 结构[76], 即在 n 层的编码器-解码器结构中, 在第 i 层和第 $n+1-i$ 层之间加入跨层连接, 直接将编码器的浅层特征加到解码器的深层特征上, 便于保持原始输入图像的边缘、纹理等底层信息. 在判别器结构方面, pix2pix 采用全卷积网络作为判别器, 不是将判别器的输出限制为 1×1 大小的分数, 而是 $m\times n$ 大小的分数, 则每个分数对应原图中的一个图像块, 使得生成器更加关注图像局部细节的真实性, 有效地避免了判别全图真实性导致的模糊问题.

针对图像转换问题, pix2pix 中提出兼顾图像真实性和输入输出图像一致性的损失函数

$$G^{*}=\arg\min_{G}\max_{D}L_{\mathrm{GAN}}(G,D)+\lambda L_{L_1}(G).$$

其中, G 表示生成器, D 表示判别器; 第 1 项表示对抗损失, 第 2 项表示重建损失; λ 表示超参数. 除了 GAN 的对抗损失外, 为了使输出尽可能地接近真实目标图像, pix2pix 还使用 L_1 损失对输出结果进行约束.

虽然 pix2pix 在多个图像转换任务上都取得了不错的结果, 但在实际应用中, 由于其训练需要大量像素级别一一对应的数据, 而这些数据在现实场景中往往难以获得, 如在黑夜白天转换问题中, 很难采集到同一场景同一角度不同时间的图像. 因此, 更进一步的研究范式是无监督的图像转换方法, 即学习 2 类图像域而非 2 幅图像之间的映射关系.

3.4.2 面向无监督迁移任务的 CycleGAN

针对现实场景缺少像素级别一一对应的成对训练数据的问题, CycleGAN[77]中提出解决无监督的图像转换问题的关键思路——循环一致性, 即图像从一个领域转换到另一个领域之后, 应该还能再转换回初始的领域, 通过 2 次迁移, 就能得到像素级别一一对应的结果, 形成用于建立对应关系的损失函数.

对于 2 个不同领域的图像数据集 X 和 Y, CycleGAN 设计了 2 个对应的生成器 $G:X\to Y$ 和 $F:Y\to X$, 分别将 X 中的图像映射到 Y 中和将 Y 中图像映射到 X 中. 此外, 还包含 2 个对应的判别器 D_X 和 D_Y, 用以判别生成的图像是否属于 X 或 Y. 循环一致性即是衡量 $F(G(X))$ 与 X 的一致性和 $G(F(Y))$ 与 Y 的一致性.

CycleGAN 中提出典型的 3 项损失函数

$$G^{*},F^{*}=\arg\min L_{\mathrm{GAN}}(G,D_Y,X,Y)+L_{\mathrm{GAN}}(F,D_X,X,Y)+\lambda L_{\mathrm{cyc}}(G,F).$$

其中, 前 2 项为标准的对抗损失, 即 D_Y 判断生成图像是否属于 Y, 促使 G 尽可能地生成真实逼真且属于 Y 的图像; D_X 判断生成图像是否属于 X, 促使 F 尽可能地生成真实逼真且属于 X 的图像. 最后一项为循环一致性损失, 可以保证图像被转换到另一个领域后仍然与原图像具有紧密的联系, 防止网络将任意输入图像转换到与输入无关但属于目标领域的图像. CycleGAN 分别对训练数据 x 和 y 进行采样, 不需要像素级别成对训练数据即可训练图像转换生成器及相应判别器, 这也是无监督图像迁移方法相对于监督方法的重要创新之处.

3.4.3 基于扩散模型的 ControlNet

自 2021 年基于扩散模型的 Stable Diffusion[43]被提出后, 许多研究者开始研究利用 Stable Diffusion 进行更可控的生成. ControlNet[78]能够根据用户提供的条件图像(如边缘图、语义分割图)进行可控生成, 完成多样化的图像转换任务. ControlNet 使用预训练好的 Stable Diffusion 的 U-Net 去噪网络作为特征提取网络进行条件图像的特征提取, 并通过零卷积层将提取出的条件图像特征注入 Stable Diffusion 原始的 U-Net 去噪网络中充当生成时的引导. 由于特征提取网络和去噪生成网络具有相同的网络结构, 二者的特征差异较小, 因此更容易起到引导和约束生成的作用.

3.4.4 图像引导的图像生成模型 IP-Adapter

Stable Diffusion 的强大之处除了在于其经过大规模数据集的训练而达到高质量生成能力外, 更在于能够根据文本提示引导的可控生成能力. 为此, IP-Adapter[79]设计了解耦的交叉注意力机制, 将条件图像按照原来文本提示的方式提供生成引导, 利用可以对齐视觉和语言的 CLIP[64]模型提取条件图像的特征, 并训练线性映射器完成交叉注意力的计算, 该方法的训练开销较低且兼容性强, 可以与其他控制模块如 ControlNet 一起使用.

然而, 由于 CLIP 模型不能提取丰富的局部细节特征, 当条件图像较复杂时, 单独使用 IP-Adapter 往往捉襟见肘, 如结构约束或采样时, 一般会搭配其他约束为条件联合进行条件生成.

3.4.5 基于扩散模型的零样本图像转换方法

2022 年, 基于 Stable Diffusion 强大的生成先验, SDEdit[80]中设计了高效的图像转换方法. 其核心思路是 2 个领域的图像在加噪到一定程度后分

布会逐渐重合. 因此, 可以在输入的条件图像上加入一定的噪声, 然后再次去噪, 带噪图像经历与加噪步数相同的去噪步数, 并在文本引导下逐步映射到目标图像领域, 完成图像转换, 整个过程无需任何训练, 称为零样本方法. SDEdit 是将 Stable Diffusion 高效应用于下游任务的代表工作之一.

Stable Diffusion 对文本的引导较为敏感, 难以通过修改文本实现对生成图像的内容进行相应的修改, 因为少量的文本修改导致输出图像整体布局和结构的巨大改变. 为了实现更精细的图像编辑, Prompt-to-prompt[81]观察到, 扩散网络中的交叉注意力部分决定了生成结果的布局结构, 进而通过替换交叉注意力图进行细粒度的图像编辑, 只修改与修改的文本对应的内容区域而保留其他区域; 此外, 通过给目标文本提示中的单词乘上系数可以调整生成强度.

Prompt-to-prompt 最初的设计是针对模型生成的图像进行编辑. 为了编辑真实图像, 一般采用基于 DDIM 反演的方法将真实图像映射为初始的高斯噪声采样. 然而, DDIM 反演的初始噪声求解方法或者无法得到准确的初始噪声结果以重建输入图像, 或者得到的初始噪声不够接近设定的高斯噪声分布, 原因是 Stable Diffusion 设定的无分类器引导技术会同时考虑目标文本去噪和空文本去噪 2 部分的融合. 因为 DDIM 反演没有考虑基于目标文本的去噪, 所以最终得到的加噪结果与原始噪声相差较大. 为了解决该问题, 2023 年被提出的 Null-Text Inversion 方法[82]可以找到输入图像在目标文本提示下的初始噪声, 进而完成定制化的生成, 达到图像转换的目的. 该方法设计了可微调的空文本机制, 即在反演过程中, 通过微调一个空文本提示来接近生成过程中的去噪中间结果. 利用此方法得到的初始噪声和微调的空文本包含原输入图像的更多信息, 在原始的文本提示引导下能更精确地重建输入图像, 并在不同文本提示的引导下将输入图像转换成不同的结果, 配合 Prompt-to-prompt 实现精准的图像编辑.

上述代表性的基于扩散模型的零样本方法的优势是不需要任何训练和微调, 即可将预训练的文生图扩散模型适配图像转换任务, 该思路在其他图像和视频编辑研究中展现出广阔的应用场景.

4　视频生成

视频生成模型可以分为基于 GAN、基于自回归模型和基于扩散模型 3 类. 其中, 扩散模型由于其强大的图像生成能力, 又能进一步细分为依靠大量视频数据训练的数据驱动模型, 将预训练的图像生成模型适配到视频任务的单样本和零样本模型. 视频生成与编辑模型分类如图 5 所示.

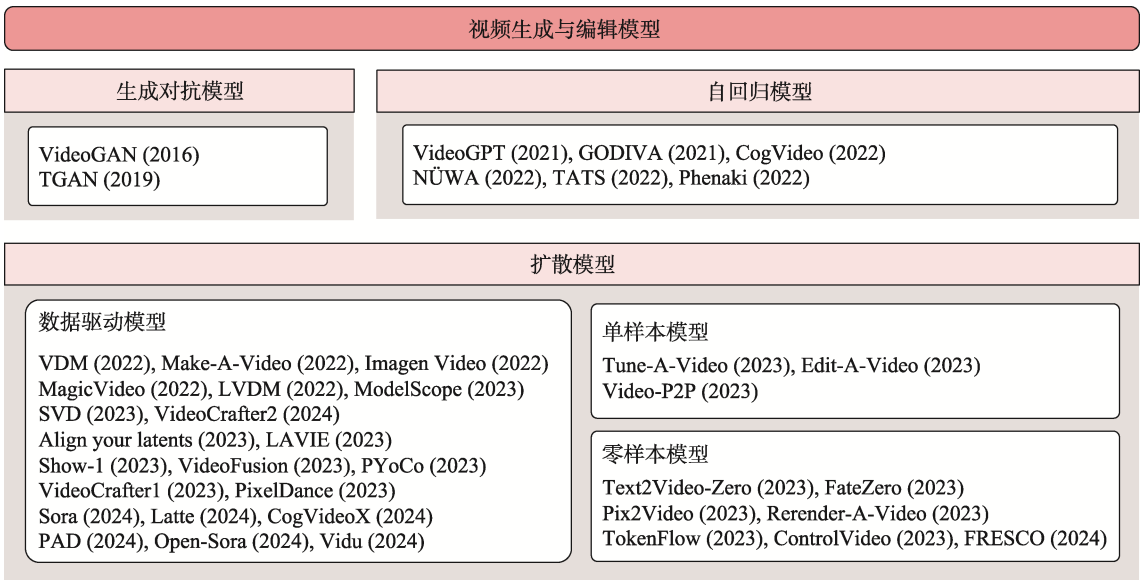

图 5　视频生成与编辑模型分类

4.1　数据驱动的视频生成

4.1.1　基于 GAN 的视频生成

GAN[2]是图像生成研究中经典的模型架构. 随着技术的发展, 研究者开始探索将 GAN 应用于更具挑战性的视频生成任务. VideoGAN[83]是首个将 GAN 应用于视频生成的工作, 其基于 DCGAN[46]

构建双流生成器架构，用于分别采样视频的前景和背景，这 2 部分信息通过掩模加权的方式合并，最终生成完整的视频输出；进一步，TGAN[84]中提出时间生成器和图像生成器的设计，其中，时间生成器负责生成一系列隐变量，而图像生成器则利用这些隐变量生成视频.

尽管上述方法取得了显著进展，但是当涉及高分辨率和长视频的生成任务时，基于 GAN 的架构仍然面临着诸多挑战，如 GAN 固有的训练不稳定性限制了其在视频生成任务上的性能.

4.1.2 基于 Transformer 的视频生成

ViT(vision Transformer)[85]面世之后，Transformer 架构[28]在图像生成任务得到了广泛的应用；VideoGPT[86]利用类 GPT(generative pre-trained transformer)结构[87]建模经由 VQ-VAE[13]自编码器压缩后的潜在特征，再解码生成视频；GODIVA[88]设计了三维稀疏注意力模型，在文本描述和经过压缩编码的稀疏视频特征上进行训练，使得模型具有根据文本提示生成视频的能力；CogVideo[89]将模型参数规模扩大至 90 亿，并采用多帧率分层训练策略以更好地对齐文本和视频内容；NÜWA[63]对不同模态的输入信息建立统一的三维表示，并通过编码器-解码器架构依据输入信息指导视频生成，该模型支持通过文本、草图和图像信息指导视频生成.

为了使模型具备生成长视频的能力，TATS[90]优化 VQ-GAN(vector quantized GAN)[91]编码器中的时序标记，并设计了分层时序 Transformer 架构，有效地提升了超过训练集视频长度的视频生成质量；Phenaki[92]通过双向 Transformer 架构增了长视频生成的可控程度，支持根据故事情节生成长视频：然而，这些模型难以生成具有复杂运动场景和丰富细节的视频.

4.1.3 基于扩散模型的视频生成

目前，扩散模型在图像和视频生成研究中均取得了较为突出的成就. 如图 6 所示，通过在大规模视频-文本数据上训练，扩散模型已经能生成符合运动规律且细节丰富的视频内容，即使是复杂的流体运动也能模拟出自然的结果. 在早期工作中，VDM(video diffusion models)[93]最早将扩散模型应用于视频生成任务，通过加入时序注意力层的方式将扩散模型中的二维 U-Net[76]扩展至三维，并在包含图像和视频的数据集上进行联合训练，实现基于扩散模型的视频生成；Make-A-Video[94]和 Imagen Video[95]在 VDM 的架构上进行优化，应

图 6 数据驱动的扩散模型的视频生成结果

用级联式模型结构进行多阶段视频生成，生成基础视频后，各阶段模块逐步对基础视频进行帧插值和超分辨率，最终输出流畅的高清晰度视频.

然而，随着模型复杂度的增加，带来的训练与推理的资源消耗是不容忽视的. 为了解决这个问题，MagicVideo[96]将 LDM[43]应用于视频生成任务，显著地降低了模型参数，提升了训练与推理效率，其在 LDM 中加入帧适配器和定向时序注意力模块，并设计了 Video VAE，在逐帧编解码视频时消除抖动；LVDM(latent video diffusion models)[97]设计隐空间上的分层扩散模型，并引入条件隐变量扰动和无条件指导机制，使模型能够生成超过 1 000 帧的高质量长视频；ModelScope[98]在预训练图像生成模型中加入时空卷积和注意力机制，支持生成多种帧数的视频；SVD(stable video diffusion)[99]测试并分析视频生成模型的训练方式，采用文本到图像预训练、视频预训练和高质量视频微调三阶段训练策略，并验证了这种策略的有效性；VideoCrafter2[100]通过在时序与空间模块上进行分布转移的方式，实现了低质量视频与高质量图像相结合的训练方法，有效地解决了高质量视频训练数据短缺的问题.

虽然资源消耗较大，但具有级联式架构的多阶段视频生成模型在早期工作中展现了较为理想的效果. 因此，部分工作将多阶段模型与 LDM 相结合. Align your latents[101]将图像生成模型依据视频时序微调以生成关键帧，之后依次通过多个 LDM 提高视频帧率和分辨率；LAVIE[102]引入时间注意力和旋转位置编码保持视频时域一致性，并构建级联 LDM 以提高生成视频的质量；Show-1[103]将级联 LDM 的视频生成、帧插值和部分超分辨率模块替换为像素级扩散模型，更好地对齐生成视频和提示文本，并降低资源消耗.

扩散模型生成图像或视频本质上是对初始噪声的去噪过程. 为了降低随机初始噪声对视频时域一致性造成的影响, VideoFusion[104]使用统一的基础噪声, 并在分配初始噪声时为每帧对应噪声加入残差; PYoCo(preserve your own correlation)[105]针对视频生成任务设计噪声先验模型, 并通过基于图像生成模型构建的多阶段视频生成模型, 验证了该先验在视频生成任务的适用性.

为了进一步提高视频生成的可控程度, VideoCrafter1[106]支持用户以文本和图像相结合的方式对生成视频进行引导, 并严格根据给定图像内容的空间结构生成视频; PixelDance[107]在文本引导的基础上增加了第 1 帧与最后一帧的图像引导, 与纯文本引导相比, 这种引导方式更加直观地描述了视频中物体运动的起止状态, 便于模型生成具有复杂动态场景的视频.

经典的基于 U-Net 去噪网络的扩散模型受限于U-Net模型本身有限的可伸缩性, 难以扩大训练规模和参数规模. 在上文图像生成任务的介绍中, Stable Diffusion 3 已经证明了基于 Transformer 的 DiT 框架的有效性. Transformer 具有优秀的可伸缩性, 随着训练数据的增多和参数规模的增大, 生成质量能稳定提升. 最近, 基于 Transformer 的扩散模型在图像生成的成功应用已逐步拓展到视频生成任务, 其中最具代表性的是 2024 年初发布的 Sora[108], 其颠覆性的视频生成效果在各界引发了大量关注, Sora 将视频在空间和时间维度同时进行压缩和 Token 化, 并基于 Dit 架构进行长时间依赖建模、处理更复杂的时序信息, 提升了视频生成效果; 在视频生成的基础上, 还支持包括视频扩展、编辑、连接等一系列视频相关任务. 但是, Sora 尚未开源, 其具体模型架构细节尚不可知. Latte[109]通过大量的实验, 探索了如何有效地对视频中的时间和空间信息进行建模, 以及如何将视频信息有效地处理成 Tokens, 并提出一套基于 Dit 架构的文生视频模型开源方案; CogVideoX[110]采用 Stable Diffusion 3[66]中的多模态 DiT 架构, 文本和视频特征具有独立的 Transformer 层, 仅在注意力层互相感知; 为了缓解 DiT 架构带来的计算开销, PAD (pyramid attention broadcast)[111]中提出金字塔注意力广播, 利用注意力机制中的冗余计算加速视频生成.

2024 年, Sora[108]在各界引发大量关注, 也激发了工业界对视频生成模型的探索. Colossal-AI 团队提出 Sora 模型的复现方案, 并发布了开源视频生成模型 Open-Sora[112], 其架构包括预训练 VAE 与 STDiT 扩散 Transformer 模型, 并依据 SVD 提出的三阶段策略进行训练. 生数科技发布是 Vidu 视频模型[113]基于 U-ViT[114]架构, 是国内首个在生成效果上接近 Sora 的视频生成模型. 快手科技推出了视频生成大模型 Kling[115], 具备与 Sora 同样强大的视频生成能力, 并比 Sora 更早进行了开放测试. 此外, 工业界还有大量开源[116-118]与非开源[119-123]方案, 以满足用户视频生成任务的相关需求.

4.2　单样本视频生成

数据驱动的视频生成模型通常在大规模文本-视频数据集上进行训练扩散模型, 需要在大型硬件加速器上进行大规模训练, 既昂贵又耗时. 为了避免数据搜集和训练资源的要求, 单样本和零样本方法被提出.

基于扩散模型的单样本视频生成任务中, 使用单个视频微调预训练的图像生成扩散模型, 使之生成相关包含运动的视频内容. Tune-A-Video[124]是这类方法的代表性工作之一, 其只使用单个文本-视频对微调图像生成扩散模型(下文中默认为 Stable Diffusion), 用户能够根据一段视频了解物体该如何运动. 受此启发, 在单样本视频生成任务中, 扩散模型能够从输入视频中捕捉到关键的运动信息, 并合成满足文本提示的新视频. Tune-A-Video 在微调阶段, 首先修改原有的扩散模型, 在 U-Net 每层加入额外的时序注意力模块, 并将空间注意力扩展为时空注意力, 然后扩展自注意力层为跨帧注意力层, 使其支持多帧输入, 并依据当前帧与第 1 帧和前一帧的关系生成当前帧的内容; 微调时, 更新原有注意力模块的 Query 部分和新增的时域注意力模块, 其余部分固定; 推理时, 根据输入视频, 使用 DDIM 反演后得到的噪声和指定的文本提示生成新视频.

这种方法虽然可以保留动作信息, 但难以保留原视频的具体细节信息. 为了解决这个问题, 在 Tune-A-Video 的基础上, Edit-A-Video[125]引入 Null-Text Inversion 进一步拟合输入视频, 并且在采样过程中采用 Prompt-to-prompt 的混合遮罩保证非编辑区域的不变性; 由于原有的遮罩计算方法没有考虑到时域一致性导致抖动, 因此提出时域一致性的混合遮罩, 进一步提升了整体的稳定性.

为了进一步改善视频生成质量, Video-P2P[126]将 Prompt-to-prompt 应用到视频生成中, 提出一种解耦指导策略改善注意力控制.在 Tune-A-Video 和 Null-Text Inversion 微调控文本的基础上, Video-P2P

认为微调后空文本更适合重建原视频，而微调前的空文本更适合编辑，考虑注意力更倾向于在早期形成，因此在生成过程早期使用原本的空文本和原始文本提示形成交叉注意力，后期使用微调后的空文本和目标文本提示形成的交叉注意力，以实现编辑和重建的平衡.

4.3 零样本视频生成

单样本视频生成可能存在对长视频训练效率低和容易对单个视频过拟合的问题，与单样本视频生成需要用一个文字-视频对来微调模型不同，基于扩散模型的零样本视频生成无需对模型的微调，直接利用预训练的图像生成扩散模型(下文中默认 Stable Diffusion)进行视频生成编辑. 如图 7 所示，零样本模型通过在图像模型的推理过程中施加帧间一致性约束，能很好地提升视频编辑后的时域一致性.

图 7 零样本视频生成模型显著提升时域一致性

Text2Video-Zero[127]是零样本视频生成的代表性方法之一，其既可以在已有的扩散模型上直接进行文本到视频的生成，又可以结合 ControlNet 实现有条件的视频生成编辑. 为了保持视频生成的一致性, Text2Video-Zero 在扩散模型的推理过程中提出初始噪声的一致采样. 首先进行一次采样，对其进行少量步数的 DDIM 去噪；然后在去噪结果上模拟运动得到一系列帧；再进行相应步数的 DDPM 加噪，得到一致采样的初始噪声. 此外，为了保证细节一致，将自注意力替换为跨帧注意力，因为完全没有修改模型参数，该方法与基于预训练扩散模型的其他支撑模型如 ControlNet 兼容，所以可以实现基于人体姿势的视频生成.

在视频编辑方面, FateZero[128]将基于 DDIM 反演和 Prompt-to-prompt 的图像编辑方法扩展到视频领域，对自注意力层提出注意力混合机制，依据交叉注意力特征计算带编辑区域的混合遮罩，将源注意力和编辑后的注意力进行加权混合；此外，将自注意力层修改为跨帧注意力层，对每帧按照自身和视频中心帧，对其到该帧的信息一起进行注意力计算.

为了保持结构的连续性，Pix2Video[129]引入输入视频的深度图作为额外引导；同时，在 DDIM 去噪过程中引入隐变量优化机制，对于去噪的每一步都预测一个去噪结果，根据每帧和前一帧相应步数预测的去噪结果的梯度优化更新，要求每帧和前一帧在前一定步数中预测的去噪结果一致.

为了进一步提升视频时域一致性，Rerender A Video[130]中引入输入视频的光流场，使用前一帧的渲染结果通过原视频光流对齐到当前帧作为参考，并将第 1 帧作为锚点防止逐步累积的外观改变，将自注意力层扩展为跨帧注意力层提升整理风格一致性，在隐空间和像素空间同时利用光流程实施对齐融合，以保证结构和纹理的一致性.

TokenFlow[131]利用扩散模型特征空间的相似性保持最终视频输出的时域一致性. 首先对输入视频进行 DDIM 反演提取其自注意力层的特征，然后使用最近邻搜索的方法获得帧间特征对应关系. 在去噪过程中，根据预先建立的对应关系融合编辑过的特征，提升最终输出的一致性.

ControlVideo[132]在 ControlNet 的基础上实现视频生成，其中引入交替帧平滑机制，通过在选定的时序步长上进行交替插帧提升平滑性.

FRESCO[133]中提出视频帧时空域对应性，在 Rerender-A-Video 提出的基于光流场的时域一致性基础上引入空域一致性约束，要求视频编辑前后相似的区域仍然保持相似，约束编辑过程的稳定性，能有效地处理视频中快速运动或遮挡导致光流场难以估计的区域.

5 数据集与评价标准

5.1 图像生成数据集

在图像生成任务中，图像生成数据集用于训练和评估图像生成模型. 本节从数据集包含的图像数量、质量、内容和特点等方面简要介绍一些常用的图像生成数据集. 表 1 所示为不同的开源图像数据集对比.

表 1　不同开源图像数据集对比

数据集	出版年	图像数量	分辨率/像素	图像内容	标注形式
MNIST[134]	1998	70 000	28×28	手写数字	类标签
CIFAR-10[135]	2009	60 000	32×32	开放	类标签
CIFAR-100[135]	2009	60 000	32×32	开放	类标签
ImageNet[5]	2009	14 197 122		开放	类标签
LSUN[136]	2015	~69 000 000		大规模场景	场景类别、图像坐标、对象边界框等
CelebA[4]	2018	202 599	178×218	人脸	人脸特征
CelebA-HQ[4]	2018	30 000	1 024×1 024	人脸	人脸特征
FFHQ[48]	2019	70 000	1 024×1 024	人脸	
AFHQ[137]	2020	15 000	512×512	动物面部	类标签
MetFaces[138]	2020	1 336	1 024×1 024	艺术作品人像	
LAION-5B[6]	2022	5 850 000 000		开放	文本标签、NSFW 检测分数

(1) MNIST 数据集[134]. 一个手写数字图像数据集, 包含 60 000 个示例的训练集和 10 000 个示例的测试集. 该数据集是美国人口普查局员工书写数字的数据库和高中学生书写数字的数据库的子集, 为手写数字的单色图像, 这些数字的大小已标准化, 并置于 28×28 像素大小的图像中心.

(2) CIFAR-10 数据集[135]. 经典的图像分类数据集, 包含飞机、猫、轮船等 10 个类别共 60 000 幅带有分类标签的彩色图像, 其中, 每个类别由 5 000 幅图像的训练集和 1 000 幅图像的测试集组成, 图像分辨率为 32×32 像素.

(3) CIFAR-100 数据集[135]. 一组图像分类数据集, 与 CIFAR-10 的图像质量相同. 与 CIFAR-10 数据集的不同之处在于, CIFAR-100 数据集将图像分为 100 个类别, 这些类别又被分为 20 个大类; 每个类别有 600 幅图像, 包含 500 幅训练图像和 100 幅测试图像, 分辨率为 32×32 像素; 每幅图像都有一个精细标签(所属类别)和一个粗糙标签(所属大类).

(4) ImageNet 数据集[5]. 一个根据 WordNet 层次结构组织的大型分类图像数据集. 到目前为止, 该数据集共包含 14 197 122 幅标定好类别的图像; 依照训练任务的不同需要, 通常将图像大小限制在 64×64 像素、128×128 像素、256×256 像素等分辨率, 并采用整个数据集的一个包含 1 000 个类别的子集作为训练集.

(5) LSUN 数据集[136]. 属于大规模场景理解数据集, 提供了餐厅、卧室、户外教堂等多个场景类别的高分辨率图像, 每个图像都有相关的标注信息, 如场景类别、图像坐标、对象边界框等; 每个类别包含的图像数量大约为 12 000~3 000 000; 训练时, 通常将图像剪裁成 256×256 像素分辨率.

(6) CelebA/CelebA-HQ 数据集[4]. 名人人像数据集, 每幅图像都有特征标记, 包含人像 bbox 标注框、5 个人脸特征点坐标和 40 个属性标记(发色、年龄等). 其中, CelebA 数据集包含 10 177 个名人身份的 202 599 幅人像图像, 分辨率为 178×218 像素; CelebA-HQ 数据集是 CelebA 数据集的高分辨率版本, 包含 30 000 幅图像, 分辨率为 1 024×1 024 像素.

(7) FFHQ 数据集[48]. 一组高质量人像图像数据集, 其中的图像从 Flickr 上抓取, 进行自动对齐和裁剪, 在年龄、种族和图像背景方面具有多样性. 该数据集包含 70 000 幅分辨率为 1 024×1 024 像素的高质量人像图像, 这些图像在人像属性上也有丰富的变化, 如眼镜、帽子、发式等.

(8) AFHQ 数据集[137]. 一个动物面部图像数据集, 包含 15 000 幅 512×512 像素分辨率的高质量图像. 该数据集包含猫、狗和野生动物 3 个大类, 每大类有 5 000 幅图像, 同时包含多个细分品种; 所有图像都经过了以眼睛为中心、在垂直和水平方向上对齐的处理.

(9) MetFaces 数据集[138]. 一组艺术作品人像图像数据集, 由 1 336 幅艺术作品中的人像图像组成, 分辨率均为 1 024×1 024 像素. 这些艺术作品全部从大都会艺术博物馆收藏 API 下载, 经剪裁对齐而成.

(10) LAION-5B 数据集[6]. 一个大规模图文对数据集, 由 58.5 亿个经筛选的图像-文本对组成, 其中, 23 亿个为英文文本, 每幅图像还提供一个 NSFW(not suitable for work)检测分数. 训练测试时, 常将图像分辨率限定为固定值, 并使用整个数据集的一个子集, 如 LAION-400M.

5.2　视频生成数据集

与图像生成数据集类似, 视频生成数据集在视频的训练和评估中有十分重要的作用. 本节简要介绍一些常用视频生成数据集的特点, 并对其

关键特性进行比较. 表 2 所示为不同的开源数据集与文本-视频对对比, 其中, 标题是视频的优质文本标签. 可以看出, 相比之下, 类标签往往过于简单, 字幕不会与视频的视觉内容同步; 在开源数据集中, HD-VG-130M 数据集的视频数量最突出, 其标签满足了视频生成的要求.

表 2 不同开源数据集与文本-视频对对比

数据集	出版年	视频数量	分辨率/像素	视频内容	标注形式	视觉过滤
UCF101[139]	2012	13 320	240	人类动作	类标签	
ActivityNet 200[140]	2015	28 000		人类动作	类标签	
ACAV100M[141]	2021	100 000 000	360	开放	字幕	
HD-VILA-100M[142]	2022	103 000 000	720	开放	字幕	
HowTo100M[143]	2019	136 000 000	240	教学视频	字幕	
YT-Temporal-180M[144]	2021	6 000 000		开放	字幕	运动
MSVD[145]	2011	2 000		开放	摘要说明	视觉文本
YouCook2[146]	2018	2 000		烹饪视频	摘要说明	
MSR-VTT[147]	2016	10 000	240	开放	摘要说明	
VaTeX[148]	2019	41 250		开放	摘要说明	
LSMDC[149]	2015	118 081	1080	电影片段	摘要说明	
WebVid-10M[150]	2021	10 000 000	360	开放	摘要说明	
Panda-70M[151]	2024	70 000 000	720	开放	摘要说明	
HD-VG-130M[152]	2023	130 000 000	720	开放	摘要说明	
HD-VG-40M[152]	2023	40 000 000	720	开放	摘要说明	视觉文本、运动和美学

(1) UCF101 数据集[139]. 收集自 YouTube 的人体动作视频的动作识别数据集, 提供了来自 101 个动作类别的 13 320 个视频, 分辨率为 240 像素; 视频包括人与物体交互、单纯的肢体动作、人与人交互、演奏乐器和体育运动 5 大类动作, 每个视频内容包含标签标注.

(2) ActivityNet 数据集[140]. 属于人体动作识别数据集, 包含 200 个类别共 28 000 个视频, 涵盖了更多(200 多种)人体动作, 每个视频内容有标签标注.

(3) ACAV100M 数据集[141]. 一组开放内容视频语言数据集, 是从 1.4 亿个完整视频中剪辑得出 1 亿个具有高度视听关系的视频数据集, 其中, 文本信息以字幕的形式标注, 视频分辨率为 360 像素.

(4) HD-VILA-100M 数据集[142]. 一个大规模多样化的视频语言数据集, 共包含 330 万个 720 像素的视频, 被切分为 1.03 亿个视频片段, 均衡分布在 15 个类别中. 该数据集不限定视频内容, 文本信息以字幕的形式标注.

(5) HowTo100M 数据集[143]. 属于视频语言数据集, 其中的原始视频收集自 YouTube 的教学视频, 依靠原有字幕对视频内容进行字幕式标注. 该数据集共包含 1.36 亿个 240 像素分辨率的视频片段.

(6) YT-Temporal-180M 数据集[144]. 一组开放内容视频语言数据集, 包含 600 万个视频共 1.8 亿帧. 该数据集视频的文本信息以字幕的形式标注, 收集不同集合的视频, 内容多样.

(7) MSVD 数据集[145]. 一组视频语言数据集, 由 2 000 多个视频和 12 万个句子组成. 该数据集的文本是人工观看视频后进行的描述、解释性的总结, 因此数据中捕获了释义和双语交替, 文本信息是概括性的说明, 其中的视频内容不限定.

(8) YouCook2 数据集[146]. 任务导向的教学视频数据集, 由涉及 89 个食谱的 2 000 个 YouTube 视频组成, 均为未经剪辑的长视频. 每个视频的程序步骤都带有时间界限的标注, 并带有说明性英文句子描述.

(9) MSR-VTT 数据集[147]. 一个包含视频和字幕的大规模数据集, 包括来自 20 个类别的 10 000 个 240 像素分辨率的视频片段, 被分为训练、验证和测试集 3 部分; 每个视频片段都被标注了大约 20 条说明性英文句子.

(10) VaTeX 数据集[148]. 一个大型多语言视频描述数据集, 包括 41 250 个视频和 825 000 组中英文字幕; 字幕文本中, 有超过 206 000 组英汉对应翻译. 该数据集不限定视频内容.

(11) LSMDC 数据集[149]. 一组高质量电影视频文字数据集, 包含从 202 部电影中提取的 118 081 个 1 080 像素分辨率的短视频片段, 每个视

频都附有字幕, 并带有说明性英文句子描述.

(12) WebVid-10M 数据集[150]. 一个大型短视频数据集, 包含 1 000 万个从网络上获取的、内容不限的带字幕视频片段, 分辨率为 360 像素, 但没有经过去除水印的处理.

(13) Panda-70M 数据集[151]. 一个高质量视频字幕对数据集, 包含 7000 万个 720 像素分辨率视频及其高质量的文本字幕, 描述文本为简短的摘要性描述.

(14) HD-VG-130M 数据集[152]. 一个高质量大型文本视频数据集, 包含来自开放域的 1.3 亿对文本视频, 分辨率为 720 像素, 具有高清、宽屏、无水印的优势. 其高质量子集 HD-VG-40M 数据集中, 增加了视觉文本、运动和美学特质, 适合更高要求的专业训练.

5.3 评价标准

质量评价是视觉生成与重建任务的重要子课题之一, 近年来出现了一些针对深度生成模型的质量评价指标, 下面介绍最具代表性的 IS(inception score), FID(Fréchet inception distance), KID(kernel inception distance), FVD(Fréchet video distance)和 CLIP score.

(1) IS[46]. 评估生成图像的质量和多样性, 首先使用预训练的 Inception 网络提取图像特征, 然后计算这些特征的标签的条件概率分布, 得分越高, 表明图像质量越好, 且类别间差异越大. 但是, IS 只考虑了边缘分布, 没有考虑真实图像和生成图像之间的关系.

(2) FID[153]. 衡量生成图像与真实图像在统计特征上的接近程度, 通过计算真实图像和生成图像的特征的 Fréchet 距离评估它们的相似性. 因为 FID 直接考虑了生成图像与真实图像之间的距离, 所以被广泛认为比 IS 更可靠.

(3) KID[154]. 另一种评估生成图像与真实图像相似度的方法, 使用核方法计算基于 Inception 网络提取的特征的距离, 类似于 FID 但通常更稳定, 尤其是在样本量较少时. KID 提供了一个无偏的估计, 计算速度通常比 FID 快.

(4) FVD[155]. 类似于 FID, 但专门针对视频数据, 用于评估视频生成模型的性能, 通过计算生成视频和真实视频在隐空间中的 Fréchet 距离衡量它们的相似性. FVD 考虑了视频的时域连续性和视觉质量, 是评估生成视频质量的重要指标.

(5) CLIP score 基于多模态预训练模型 CLIP[64], 可用于多种视觉任务. CLIP 通过联合训练图像和文本数据来理解图像内容与自然语言描述之间的关系, 其特征提取能力使得它可以用于生成图像的语义评估, 判断生成图像与输入的文本提示之间的一致性.

6 结语与展望

AIGC 是计算机视觉领域的热门研究课题, 具有广泛的应用前景. 针对面向视觉生成的 AIGC 方法与应用, 本文首先介绍了主流的基于深度学习的四类生成模型; 然后重点归纳现有的图像生成模型, 按照输入条件的类型进行分类和综述; 再梳理现有的视频生成模型, 依据训练数据的要求进行综述, 并总结了各类方法的优缺点; 最后介绍目前图像生成和视频生成数据集以及评价指标.

尽管当前扩散模型对生成高质量图像和视频具有优异的性能, 但仍然在以下方面存在一些问题和挑战.

(1) 高质量数据集的获取. 尽管业界出现了 LAION-5B[6]和 HD-VG-130M[152]等大规模图像视频数据集, 但现有数据集的获取方法以及高质量的文本数据标注本身仍然流程烦琐、消耗巨大. 而在深度方面, 面向专业性任务的高质量数据集仍然匮乏, 难以完成更专业化、更定制化的模型训练. 未来, 可以考虑结合用户反馈和自然语言大模型驱动的数据搜集和标注方法, 提升在指定任务上的数据集搜集效率.

(2) 大模型的推理效率. 虽然扩散模型在生成质量上具有显著优势, 但推理速度较慢是当前扩散模型的主要瓶颈之一. 为此, 大模型的推理加速是研究热点之一, 主要围绕以下方面展开. 首先是优化去噪机制, 研究能够跨步生成的采样方法, 如 DDIM[33], DPM-Solver[34]. 其次研究模型蒸馏, 将大模型的知识浓缩到更小的模型, 或者将需要较多采样步数的模型压缩为只需要较少采样步数的模型, 其中具有代表性的工作为 SDXL turbo[65]. 在近期的热点工作 Rectified Flow[156]中, 同时考虑扩散模型的优化策略和模型蒸馏, 采用走直线的方式实现图像和视频的快速生成. 最后, 通过优化模型架构进行模型结构修剪[157]和注意力稀疏化[158], 达到降低模型整体复杂度的目的. 尽管目前已有一些基于加速推理和模型蒸馏的方法, 但是这类方法或降低了生成质量或加速效率不足, 难以在端侧部署, 已经成为制约扩散模型大规模

应用的关键因素之一. 如何有效地提升扩散模型的推理效率, 或者结合传统快速推理的卷积神经网络和 GAN 的框架, 或者提出全新的理论框架破解扩散模型本身的瓶颈, 是未来需要研究者关注的方向.

(3) 长视频生成. 现阶段, 视频扩散模型在 10 s 内的超短视频生成方面取得了显著的进展, 但是面向超过 1 min 甚至更久的常规视频长度, 扩散模型在保持长时视觉一致性和生成内容的连贯性方面仍然有所欠缺: a. 长时间高质量的视频数据集更难以搜集; b. 算力和显存本身也限制了模型一次能够观察到的视频帧数量; c. 目前的模型架构也难以建模超长时间跨度下的帧间逻辑关系. 未来可考虑从以下方面展开研究: 首先引入具有强大关系、逻辑建模能力的视频理解神经网络, 如采用图神经网络建模时序之间的逻辑关系; 然后引入记忆机制, 在模型生成过程中记录下已生成的场景内容信息, 如动态更新的三维场景信息[159], 避免长时生成出现的遗忘问题; 再采用由粗到精的脚本生成、关键帧生成和非关键帧生成的多尺度生成模式, 逐步提升长视频生成质量; 最后从推理的角度出发, 研究基于自回归或先进先出[160]的视频生成模型, 实现无限时长的生成.

(4) 生成视频的可控性与用户定制. 在图像生成任务中, 已有很多可控和定制化生成方面的研究, 包括额外的结构约束[78], 图像信息的注入[79], 微调文本编码或去噪模型实现指定物体或风格的定制化生成[161-163]. 在视频生成中, 也出现了一些相关的可控性生成研究, 如基于人体动作信息驱动的视频生成[164-165]和基于镜头或者轨迹的视频生成[166-168], 但面对复杂场景和复杂动态, 如何精确地控制生成视频中人物动作、场景变化等具体元素, 仍是亟待解决的挑战性问题. 未来的研究可以集中在更精细化地控制信号注入, 基于微调或推断时优化的局部动态控制机制设计, 结合强化学习引入用户反馈作为生成过程的控制信号, 引入外部的场景和动作专家库提供物理视觉先验知识等方向.

(5) 隐私安全问题. 随着生成技术的快速发展, 大模型生成的图像和视频已经能够以假乱真, 随之而来的是虚假信息生成和版权保护等方面的挑战. 早在 GAN 的时代, 研究者就开始关注 AI 换脸等应用可能带来的安全问题. 为此, 2020 年, Facebook 联手 Microsoft 举办了全球 DeepFake 检测竞赛. 近年来, 人们发现扩散模型存在对训练数据集的记忆能力, 引发了公众对大模型版权保护的担忧. 为此, 一些研究者采用定制化生成相反的策略, 研究对模型进行指定概念的混淆训练[169], 避免生成相关概念的图像. 针对 AIGC 的安全问题, 在监管层面, 需要进一步探索如何建立有效的监管机制, 对大模型的使用进行约束和引导, 防止其被用于制造和传播虚假信息; 在模型生成之前, 对模型输入条件进行意图分析, 在生成初始阶段阻止存在恶意的用户; 在模型生成过程中及结束后, 研究数字水印等防伪技术, 对生成内容进行标识和追踪, 支撑版权保护.

致谢 白宇航、兰晰程、李一凡和高文硕同学在完成本文过程中协助对综述方法进行查漏补缺和整理数据集相关数据项, 在此表示感谢!

参考文献(References):

[1] Kingma D P, Welling M. Auto-encoding variational bayes[C] // Proceedings of the 2nd International Conference on Learning Representations. Banff: ICLR Press, 2014: 1-14

[2] 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. Cambridge: MIT Press, 2014: 2672-2680

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.574

[4] Liu Z W, Luo P, Wang X G, *et al*. Large-scale celebfaces attributes (CelebA) dataset[EB/OL]. [2024-06-06]. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html/CelebA/CelebA.html

[5] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 248-255

[6] Schuhmann C, Beaumont R, Vencu R, *et al*. LAION-5B: an open large-scale dataset for training next generation image-text models[C] //Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2022: Article No.1833

[7] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[C] //Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR Press, 2019: 1-35

[8] Larochelle H, Murray I. The neural autoregressive distribution estimator[C] //Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR Press, 2011: 29-37

[9] Burda Y, Grosse R B, Salakhutdinov R. Importance weighted autoencoders[C] //Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR Press, 2016: 1-14

[10] Sohn K, Yan X C, Lee H. Learning structured output representation using deep conditional generative models[C] //Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 3483-3491

[11] Abbasnejad M E, Dick A, van den Hengel A. Infinite variational autoencoder for semi-supervised learning[C] //Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 781-790

[12] Zhao S J, Song J M, Ermon S. InfoVAE: information maximizing variational autoencoders[OL]. [2024-06-06]. https://arxiv.org/abs/1706.02262

[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. Red Hook: Curran Associates Inc, 2017: 6309-6318

[14] Gregor K, Danihelka I, Graves A, *et al*. DRAW: a recurrent neural network for image generation[C] //Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR Press, 2015: 1462-1471

[15] Kulkarni T D, Whitney W F, Kohli P, *et al*. Deep convolutional inverse graphics network[C] //Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2539-2547

[16] Kusner M J, Paige B, Hernandez-Lobato J M. Grammar variational autoencoder[C] //Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR Press, 2017: 1945-1954

[17] Mao X D, Li Q, Xie H R, *et al*. Least squares generative adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2813-2821

[18] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[OL]. [2024-06-06]. https://arxiv.org/abs/1701.07875

[19] Gulrajani I, Ahmed F, Arjovsky M, *et al*. Improved training of Wasserstein GANs[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 5769-5779

[20] Miyato T, Kataoka T, Koyama M, *et al*. Spectral normalization for generative adversarial networks[C] //Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR Press, 2018: 1-26

[21] Mirza M, Osindero S. Conditional generative adversarial nets[OL]. [2024-06-06]. https://arxiv.org/abs/1411.1784

[22] Zhang H, Goodfellow I J, Metaxas D, *et al*. Self-attention generative adversarial networks[C] //Proceedings of the 36th International Conference on Machine Learning. New York: JMLR Press, 2019: 7354-7363

[23] Denton E, Chintala S, Szlam A, *et al*. Deep generative image models using a laplacian pyramid of adversarial networks[C] // Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 1486-1494

[24] Shaham T R, Dekel T, Michaeli T. SinGAN: learning a generative model from a single natural image[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4569-4579

[25] van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks[C] //Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR Press, 2016: 1747-1756

[26] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780

[27] van den Oord A, Kalchbrenner N, Vinyals O, *et al*. Conditional image generation with PixelCNN decoders[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2016: 4797-4805

[28] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6000-6010

[29] Ding M, Yang Z Y, Hong W Y, *et al*. CogView: mastering text-to-image generation via transformers[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.1516

[30] Tian K Y, Jiang Y, Yuan Z H, *et al*. Visual autoregressive modeling: Scalable image generation via next-scale prediction[OL]. [2024-06-06]. https://arxiv.org/abs/2404.02905

[31] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.672

[32] Ho J, Salimans T. Classifier-free diffusion guidance[OL]. [2024-06-06]. https://arxiv.org/abs/2207.12598

[33] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models[C] //Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR Press, 2021: 1-22

[34] Lu C, Zhou Y H, Bao F, *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. Red Hook: Curran Associates Inc, 2022: Article No.418

[35] Saharia C, Chan W, Chang H, *et al*. Palette: image-to-image diffusion models[C] //Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings. New York: ACM Press, 2022: Article No.15

[36] Saharia C, Ho J, Chan W, *et al*. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726

[37] Lugmayr A, Danelljan M, Romero A, *et al*. RePaint: Inpainting using denoising diffusion probabilistic models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 11451-11461

[38] Li B, Xue K, Liu B, *et al*. BBDM: image-to-image translation with Brownian bridge diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 1952-1961

[39] Zabari N, Azulay A, Gorkor A, *et al*. Diffusing colors: image

colorization with text guided diffusion[C] //Proceedings of the SIGGRAPH Asia 2023 Conference Papers. New York: ACM Press, 2023: Article No.61

[40] Ramesh A, Dhariwal P, Nichol A, *et al*. Hierarchical text-conditional image generation with clip latents[OL]. [2024-06-06]. https://arxiv.org/abs/2204.06125

[41] 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. ICML, 2022: 16784-16804

[42] 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. Red Hook: Curran Associates Inc, 2022: Article No.2643

[43] Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 10674-10685

[44] Podell D, English Z, Lacey K, *et al*. SDXL: improving latent diffusion models for high-resolution image synthesis[C] //Proceedings of the 12th International Conference on Learning Representations. Vienna: ICLR Press, 2024: 1-21

[45] Peebles W, Xie S N. Scalable diffusion models with transformers[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 4172-4182

[46] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[C] //Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR Press, 2016: 1-16

[47] Karras T, Aila T, Laine S, *et al*. Progressive growing of GANs for improved quality, stability, and variation[C] //Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR Press, 2018: 1-26

[48] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4396-4405

[49] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 1510-1519

[50] Karras T, Laine S, Aittala M, *et al*. Analyzing and improving the image quality of StyleGAN[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8107-8116

[51] Karras T, Aittala M, Laine S, *et al*. Alias-free generative adversarial networks[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.66

[52] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs[C] //Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR Press, 2017: 2642-2651

[53] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[OL]. [2024-06-06]. https://arxiv.org/abs/2312.00752

[54] Peng B, Alcaide E, Anthony Q, *et al*. RWKV: reinventing RNNs for the transformer era[C] //Proceedings of the Findings of the Association for Computational Linguistics. Singapore: ACL Press, 2023: 14048-14077

[55] Ma N Y, Goldstein M, Albergo M S, *et al*. SiT: exploring flow and diffusion-based generative models with scalable interpolant transformers[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2024: 23-40

[56] Chen R T Q, Rubanova Y, Bettencourt J, *et al*. Neural ordinary differential equations[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2018: 6572-6583

[57] Song Y, Sohl-Dickstein J, Kingma D P, *et al*. Score-based generative modeling through stochastic differential equations[C] //Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR Press, 2021: 1-36

[58] Reed S E, Akata Z, Yan X C, *et al*. Generative adversarial text to image synthesis[C] //Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR Press, 2016: 1060-1069

[59] Zhang H, Xu T, Li H S, *et al*. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 5908-5916

[60] Zhang H, Xu T, Li H S, *et al*. StackGAN++: realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947-1962

[61] Ramesh A, Pavlov M, Goh G, *et al*. Zero-shot text-to-image generation[C] //Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR Press, 2021: 8821-8831

[62] Ding M, Zheng W D, Hong W Y, *et al*. CogView2: faster and better text-to-image generation via hierarchical transformers[C] //Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2022: Article No.1229

[63] Wu C F, Liang J, Ji L, *et al*. NÜWA: visual synthesis pre-training for neural visual world creation[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 720-736

[64] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR Press, 2021: 8748-8763

[65] Stability.ai. Introducing SDXL turbo: a real-time text-to-image generation model[EB/OL]. [2024-06-06]. https://stability.ai/news/stability-ai-sdxl-turbo

[66] Stability.ai. Stable diffusion 3: research paper[EB/OL]. [2024-06-06]. https://stability.ai/news/stable-diffusion-3-research-paper

[67] Chen J S, Yu J C, Ge C J, *et al*. PixArt-α: fast training of diffu-

sion transformer for photorealistic text-to-image synthesis[OL]. [2024-06-06]. https://openreview.net/pdf?id=eAKmQPe3m1

[68] Chen J S, Ge C J, Xie E Z, *et al*. PixArt-Σ: weak-to-strong training of diffusion transformer for 4K text-to-image generation[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2024: 74-91

[69] Chen J S, Wu Y, Luo S M, *et al*. PixArt-δ: fast and controllable image generation with latent consistency models[OL]. [2024-06-06]. https://arxiv.org/abs/2401.05252

[70] Li Z M, Zhang J W, Lin Q, *et al*. Hunyuan-DiT: a powerful multi-resolution diffusion transformer with fine-grained Chinese understanding[OL]. [2024-06-06]. https://arxiv.org/abs/2405.08748

[71] Betker J, Goh G, Jing L, *et al*. Improving image generation with better captions[J]. Computer Science, 2018, 41(8): 1947-1962

[72] OpenAI. Introducing ChatGPT[EB/OL]. [2024-06-06]. https://openai.com/index/chatgpt/

[73] Li D Q, Kamko A, Akhgari E, *et al*. Introducing playground v2.5[EB/OL]. [2024-06-06]. https://playground.com/blog/playground-v2-5

[74] Zhang LM, Agrawala M. Transparent image layer diffusion using latent transparency[J]. ACM Transactions on Graphics, 2024, 43(4): Article No.100

[75] Isola P, Zhu J Y, Zhou T H, *et al*. Image-to-image translation with conditional adversarial networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5967-5976

[76] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[77] Zhu J Y, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2242-2251

[78] Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 3813-3824

[79] Ye H, Zhang J, Liu S B, *et al*. IP-Adapter: text compatible image prompt adapter for text-to-image diffusion models[OL]. [2024-06-06]. https://arxiv.org/abs/2308.06721

[80] Meng C L, He Y T, Song Y, *et al*. SDEdit: guided image synthesis and editing with stochastic differential equations[C] // Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR Press, 2022: 1-33

[81] Hertz A, Mokady R, Tenenbaum J, *et al*. Prompt-to-prompt image editing with cross-attention control[C] //Proceedings of the 11th International Conference on Learning Representations. Kigali Rwanda: ICLR Press, 2023: 1-19

[82] Mokady R, Hertz A, Aberman K, *et al*. Null-text inversion for editing real images using guided diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 6038-6047

[83] Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2016: 613-621

[84] Ding Z H, Liu X Y, Yin M, *et al*. TGAN: deep tensor generative adversarial nets for large image generation[OL]. [2024-06-06]. https://arxiv.org/abs/1901.09953

[85] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: transformers for image recognition at scale[C] //Proceedings of the 9th International Conference on Learning Representations. Ababa: ICLR Press, 2021: 1-21

[86] Yan W, Zhang Y Z, Abbeel P, *et al*. VideoGPT: video generation using vq-vae and transformers[OL]. [2024-06-06]. https://arxiv.org/abs/2104.10157

[87] Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training[EB/OL]. [2024-06-06]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[88] Wu C F, Huang L, Zhang QX, *et al*. GODIVA: generating open-domain videos from natural descriptions[OL]. [2024-06-06]. https://arxiv.org/abs/2104.14806

[89] Hong W Y, Ding M, Zheng W D, *et al*. CogVideo: large-scale pretraining for text-to-video generation via transformers[OL]. [2024-06-06]. https://openreview.net/pdf?id=rB6TpjAuSRy

[90] Ge S W, Hayes T, Yang H, *et al*. Long video generation with time-agnostic VQGAN and time-sensitive transformer[C] // Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 102-118

[91] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 12868-12878

[92] Villegas R, Babaeizadeh M, Kindermans P J, *et al*. Phenaki: variable length video generation from open domain textual descriptions[C] //Proceedings of the 11th International Conference on Learning Representations. Virtual Event: ICLR Press, 2022: 1-14

[93] Ho J, Salimans T, Gritsenko A, *et al*. Video diffusion models[C] //Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2022: Article No.628

[94] Singer U, Polyak A, Hayes T, *et al*. Make-A-Video: text-to-video generation without text-video data[C] //Proceedings of the 11th International Conference on Learning Representations. Kigali Rwanda: ICLR Press, 2023: 1-16

[95] Ho J, Chan W, Saharia C, *et al*. Imagen video: high definition video generation with diffusion models[OL]. [2024-06-06]. https://arxiv.org/abs/2210.02303

[96] Zhou D Q, Wang W M, Yan H S, *et al*. MagicVideo: efficient video generation with latent diffusion models[OL]. [2024-06-06]. https://arxiv.org/abs/2211.11018

[97] He Y Q, Yang T Y, Zhang Y, *et al*. Latent video diffusion models for high-fidelity long video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2211.13221

[98] Wang J N, Yuan H J, Chen D Y, *et al*. ModelScope text- to-video technical report[OL]. [2024-06-06]. https://arxiv.org/

abs/2308.06571

[99] Blattmann A, Dockhorn T, Kulal S, *et al*. Stable video diffusion: scaling latent video diffusion models to large datasets [OL]. [2024-06-06]. https://arxiv.org/abs/2311.15127

[100] Chen H X, Zhang Y, Cun X D, *et al*. VideoCrafter2: overcoming data limitations for high-quality video diffusion models[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 7310-7320

[101] Blattmann A, Rombach R, Ling H, *et al*. Align your latents: high-resolution video synthesis with latent diffusion models[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 22563-22575

[102] Wang Y H, Chen X Y, Ma X, *et al*. Lavie: high-quality video generation with cascaded latent diffusion models[OL]. [2024-06-06]. https://arxiv.org/abs/2309.15103

[103] Zhang D J, Wu J Z, Liu J W, *et al*. Show-1: marrying pixel and latent diffusion models for text-to-video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2309.15818

[104] Luo Z X, Chen D Y, Zhang Y Y, *et al*. Notice of removal: VideoFusion: decomposed diffusion models for high-quality video generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 10209-10218

[105] Ge S W, Nah S, Liu G L, *et al*. Preserve your own correlation: a noise prior for video diffusion models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 22873-22884

[106] Chen H X, Xia M H, He Y Q, *et al*. VideoCrafter1: open diffusion models for high-quality video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2310.19512

[107] Zeng Y, Wei G Q, Zheng J N, *et al*. Make pixels dance: high-dynamic video generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 8850-8860

[108] OpenAI. Video generation models as world simulators [EB/OL]. [2024-06-06]. https://openai.com/index/video-generation-models-as-world-simulators

[109] Ma X, Wang Y H, Jia G Y, *et al*. Latte: latent diffusion transformer for video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2401.03048

[110] Yang Z Y, Teng J Y, Zheng W D, *et al*. CogVideoX: text-to-video diffusion models with an expert Transformer[OL]. [2024-06-06]. https://arxiv.org/abs/2408.06072

[111] Zhao X L, Jin X L, Wang K, *et al*. Real-time video generation with pyramid attention broadcast[OL]. [2024-06-06]. https://arxiv.org/abs/2408.12588

[112] Gallery. Open-Sora: democratizing efficient video production for all[EB/OL]. [2024-06-06]. https://hpcaitech.github.io/Open-Sora

[113] Bao F, Xiang C D, Yue G, *et al*. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models[OL]. [2024-06-06]. https://arxiv.org/abs/2405.04233

[114] Bao F, Nie S, Xue K W, *et al*. All are worth words: a ViT backbone for diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 22669-22679

[115] Kling. Kuaishou unveils proprietary video generation model ‘kling;’ testing now available[EB/OL]. [2024-06-06]. https://ir.kuaishou.com/news-releases/news-release-details/kuaishou-unveils-proprietary-video-generation-model-kling

[116] Sastry A, Mullan J, Crawbuck D. Direct text-to-video synthesis with enhanced motion dynamics and large-scale text-video pair training[EB/OL]. [2024-06-06]. https://hotshot.co/act-one

[117] Guo Y W, Yang C Y, Rao A Y, *et al*. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning[C] //Proceedings of the 12th International Conference on Learning Representations. ICLR, 2024: 1-13

[118] Xia Z Q, Chen Z K, Wu B, *et al*. MuseV: infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising[EB/OL]. [2024-06-06]. https://tmelyralab.github.io/MuseV_Page/

[119] Esser P, Chiu J, Atighehchian P, *et al*. Structure and content-guided video synthesis with diffusion models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 7312-7322

[120] GenMo. Meet Replay, the next generation in AI video[EB/OL]. [2024-06-06]. https://blog.genmo.ai/log/replay-ai-video

[121] Plai. A new way to create[EB/OL]. [2024-06-06]. https:// plaiday.io/a-new-way-to-create/

[122] Assitive. Introducing assistive video[EB/OL]. [2024-06-06]. https://assistive.chat/blog/introducing-assistive-video

[123] Moran T. Complete guide to videoleap’s AI: streamline & enhance your editing[EB/OL]. [2024-06-06]. https://www.videoleapapp.com/blog/introducing-videoleap-ai

[124] Wu J Z, Ge Y X, Wang X T, *et al*. Tune-A-Video: one-shot tuning of image diffusion models for text-to-video generation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 7589-7599

[125] Shin C, Kim H, Lee C H, *et al*. Edit-A-Video: single video editing with object-aware consistency[C] //Proceedings of the Asian Conference on Machine Learning. Hanoi: PMLR Press, 2023: 1215-1230

[126] Liu S T, Zhang Y C, Li W B, *et al*. Video-P2P: video editing with cross-attention control[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 8599-8608

[127] Khachatryan L, Movsisyan A, Tadevosyan V, *et al*. Text2Video-Zero: text-to-image diffusion models are zero-shot video generators[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 15908-15918

[128] Qi C Y, Cun X D, Zhang Y, *et al*. FateZero: fusing attentions for zero-shot text-based video editing[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 15886-15896

[129] Ceylan D, Huang C H P, Mitra N J. Pix2Video: video editing using image diffusion[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 23149-23160

[130] Yang S, Zhou Y F, Liu Z W, *et al*. Rerender a video: zero-shot text-guided video-to-video translation[C] //Proceedings of the SIGGRAPH Asia 2023 Conference Papers. New York: ACM Press, 2023: Article No.95

[131] Geyer M, Bar-Tal O, Bagon S, *et al*. TokenFlow: consistent diffusion features for consistent video editing[C] //Proceedings of the 12th International Conference on Learning Representations. Kigali Rwanda: ICLR Press, 2024: 1-13

[132] Zhang Y B, Wei Y X, Jiang D S, *et al*. ControlVideo: training-free controllable text-to-video generation[C] //Proceedings of the 12th International Conference on Learning Representations. ICLR, 2024: 1-11

[133] Yang S, Zhou Y F, Liu Z W, *et al*. FRESCO: spatial-temporal correspondence for zero-shot video translation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 8703-8712

[134] LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[135] Krizhevsky A. Learning multiple layers of features from tiny images[EB/OL]. [2024-06-06]. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[136] Yu F, Seff A, Zhang Y D, *et al*. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[OL]. [2024-06-06]. https://arxiv.org/abs/1506.03365

[137] Choi Y, Uh Y, Yoo J, *et al*. StarGAN v2: diverse image synthesis for multiple domains[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8185-8194

[138] Karras T, Aittala M, Hellsten J, *et al*. Training generative adversarial networks with limited data[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.1015

[139] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[OL]. [2024-06-06]. https://arxiv.org/abs/1212.0402

[140] Caba Heilbron F, Escorcia V, Ghanem B, *et al*. ActivityNet: a large-scale video benchmark for human activity understanding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 961-970

[141] Lee S, Chung J, Yu Y, *et al*. ACAV100M: automatic curation of large-scale datasets for audio-visual video representation learning[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 10254-10264

[142] Xue H W, Hang T K, Zeng Y H, *et al*. Advancing high-resolution video-language representation with large-scale video transcriptions[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 5026-5035

[143] Miech A, Zhukov D, Alayrac J B, *et al*. HowTo100M: learning a text-video embedding by watching hundred million narrated video clips[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 2630-2640

[144] Zellers R, Lu X M, Hessel J, *et al*. MERIOT: multimodal neural script knowledge models[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.1810

[145] Chen D L, Dolan W B. Collecting highly parallel data for paraphrase evaluation[C] //Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Stroudsburg: ACL, 2011: 190-200

[146] Zhou L W, Xu C L, Corso J. Towards automatic learning of procedures from web instructional videos[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 1-9

[147] Xu J, Mei T, Yao T, *et al*. MSR-VTT: a large video description dataset for bridging video and language[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 5288-5296

[148] Wang X, Wu J W, Chen J K, *et al*. VaTeX: a large-scale, high-quality multilingual dataset for video-and-language research[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4580-4590

[149] Rohrbach A, Rohrbach M, Tandon N, *et al*. A dataset for movie description[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3202-3212

[150] Bain M, Nagrani A, Varol G, *et al*. Frozen in time: a joint video and image encoder for end-to-end retrieval[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 1708-1718

[151] Chen T S, Siarohin A, Menapace W, *et al*. Panda-70M: captioning 70M videos with multiple cross-modality teachers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 13320-13331

[152] Wang W J, Yang H, Tuo Z X, *et al*. VideoFactory: swap attention in spatiotemporal diffusions for text-to-video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2305.10874v1

[153] Salimans T, Goodfellow I, Zaremba W, *et al*. Improved techniques for training GANs[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2016: 2234-2242

[154] Bińkowski M, Sutherland D J, Arbel M, *et al*. Demystifying MMD GANs[C] //Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR Press, 2018: 1-36

[155] Unterthiner T, van Steenkiste S, Kurach K, *et al*. FVD: a new metric for video generation[C] //Proceedings of the Deep Generative Models for Highly Structured Data. New Orleans: ICLR Press, 2019: 1-9

[156] Liu X C, Gong C Y, Liu Q. Flow straight and fast: learning to generate and transfer data with rectified flow[C] //Proceedings of the 11th International Conference on Learning Representations. Kigali: ICLR Press, 2023: 1-15

[157] Fang G F, Ma X Y, Wang X C. Structural pruning for diffusion models[C] //Proceedings of the 37th International Conference

on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.731

[158] Liu S H, Ye J W, Ren S C, *et al*. DynaST: dynamic sparse transformer for exemplar-guided image generation[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 72-90

[159] Yu H X, Duan H Y, Hur J, *et al*. WonderJourney: going from anywhere to everywhere[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 6658-6667

[160] Kim J, Kang J, Choi J, *et al*. FIFO-diffusion: generating infinite videos from text without training[OL]. [2024-06-06]. https://arxiv.org/abs/2405.11473

[161] Gal R, Alaluf Y, Atzmon Y, *et al*. An image is worth one word: personalizing text-to-image generation using textual inversion[C] //Proceedings of the 11th International Conference on Learning Representations. Kigali: ICLR Press, 2023: 1-14

[162] Ruiz N, Li Y Z, Jampani V, *et al*. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 22500-22510

[163] Hu E J, Shen Y L, Wallis P, *et al*. LoRA: low-rank adaptation of large language models[C] //Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR Press, 2022: 1-13

[164] Li H. Animate anyone: consistent and controllable image-to-video synthesis for character animation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 8153-8163

[165] Xu Z C, Zhang J F, Liew J H, *et al*. MagicAnimate: temporally consistent human image animation using diffusion model[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 1481-1490

[166] Wang Z X, Yuan Z Y, Wang X T, *et al*. MotionCtrl: a unified and flexible motion controller for video generation[C] //Proceedings of the ACM SIGGRAPH 2024 Conference Papers. New York: ACM Press, 2024: Article No.114

[167] Xiao Z Q, Zhou Y F, Yang S, *et al*. Video diffusion models are training-free motion interpreter and controller[OL]. [2024-06-06]. https://arxiv.org/abs/2405.14864

[168] Wu J Z, Li X T, Zeng Y H, *et al*. MotionBooth: motion-aware customized text-to-video generation[OL]. [2024-06-06]. https://arxiv.org/abs/2406.17758

[169] Kumari N, Zhang B L, Wang S Y, *et al*. Ablating concepts in text-to-image diffusion models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 22634-22645