光学学报

从感知到创造:图像视频生成式方法前沿探讨

林倞,杨斌斌*

中山大学计算机学院,广东 广州 510006

摘要 随着计算机软硬件的迅速发展,人工智能(AI)模型在感知型任务中实现了接近或者超越人类能力的性能水平。然而,为了开发能够全面理解世界的成熟AI系统,模型必须能够生成视觉概念,而非仅仅是识别它们。首先全面概述现有的生成框架,其中包括对抗生成网络、变分自动编码器、流模型和扩散模型;然后,回顾最近在图像和视频生成方面的最新进展,并讨论它们的局限性;最后,提出改进现有视觉生成模型的可行策略,并概述有前途的未来研究方向。这些讨论和探究对推动视觉生成建模领域的发展和全面挖掘AI系统在视觉概念生成领域的潜能具有重要的意义。

关键词 人工智能模型;视觉生成建模;扩散模型;图像和视频生成

中图分类号 TP391 **文献标志码** A **DOI**: 10.3788/AOS230758

1 引 言

人工智能(AI)[1-4]作为计算机科学技术领域的一个分支,研究的是如何让计算机通过与环境交互[5-8]进行经验性学习后,获得类似人类的智能并能够在特定的任务中做出合理的推断与决策。由于飞速进步的计算硬件[9-10]和软件[11-12]的加持,在人工智能领域很多的子领域、子任务中,机器的推断水平已经可以接近甚至超越人类能力的平均水平。然而,现有人工智能中的重大突破大多是感知型的,例如对物体的识别[13-16]、检测[17-21]、分割[22-24]等。这些感知型任务大多都可以转换为机器学习[25]中的分类与回归问题[26-29],进而利用深度神经网络强大的拟合能力来拟合它们的复杂决策边界。因此可以认为,随着深度学习的发展与进步,AI已经具有了感知世界的能力。

但是,对于最终的AI形态,感知世界是远远不够的,它需要真正地认识、理解整个世界的一切,并学会像人类一样,能够创造性地生产符合自然规则、人类认知的一些产物。在人类的发展中亦是如此:要让一个人学会生产,则需要让他首先对当下的事物有足够深度的认识与独到的理解,否则生产出来的产品会让大家产生不适并感到违和。例如,一个厨师只有在对食物的原材料、味道的搭配、厨具的使用原理、摆盘的讲究都非常熟练、研究到位之后才能烹饪出一道符合大家口味、观感上惊艳的美食。但凡以上几点的其中之一未能做到,都会导致烹饪出来的食物在味道或者视觉效果上不佳。因此,当下人工智能的挑战在于,如何让AI将已经获得的感知能力进一步演化为更深层次的生成能力。

从技术角度上来看,感知模型只需要利用神经网络拟合现有的数据分布的决策边界,将一个分布划分为多个子分布(对应于分类任务),而不需要考量数据分布是如何产生的[30-31]。而对于生成式AI[32-37],则需要站在一个更高的角度,全面地剖析真实数据分布的各种数学特征,进而通过参数优化的方式来得到一个逼近真实数据分布的分布。通过在这个逼近分布上进行数据点的采样,便可以让AI生成不同模态、不同结构的数据,例如文本、图像、音频等。近期,随着Masked Autoencoder掩码自编码器(MAE)[38]等自监督生成式模型的成功,进一步验证了,在无人工标注类别的情况下(不需要人工指明数据的子分布规律),以MAE为代表的自监督生成式模型[39-46]能够对单一自然图像的内在性质进行挖掘与理解,并且可以进一步将其学习到的知识拿来作分类与判别[47-51]。

从应用层面来说,当AI模型具备视觉理解与生成的能力后,大大推动了业界各方各面的进步与发展。例如:对古老黑白照片与电影进行彩色化[52-57]、高清化修复与重制作[58-64];通过多模态理解合成用于实时进行手语翻译的虚拟主播[65-69]和AI数字人[70-76];在短视频平台中引入创意特效合成,以方便个人视频的定制化拍摄;人物肖像、图片的风格化;电影特效的合成与场景渲染[77-85]等。因此,视觉生成模型的原理与方法的研究具有非常重要的理论意义与工业应用价值。

2 生成模型的原理与框架

所谓的视觉生成,即让AI在已有的图像数据集上进行训练学习,拟合真实图像数据中的潜在分布特征

收稿日期:2023-03-30;**修回日期**:2023-04-11;**录用日期**:2023-07-22;**网络首发日期**:2023-08-02

通信作者:*yangbb3@mail2.sysu.edu.cn

后，能够进行从无到有的采样、合成、创造、编辑。这个生成的过程，也可以称作AI绘画或AI创作。通常来说，研究人员会假定图像/视频从无到有的生成过程为一个从标准高斯分布$N(0, \boldsymbol{I})$到真实图像/视频分布之间的映射G'。当从标准高斯分布上随机采样一个噪声$\boldsymbol{z}$后，经映射变换得到的结果$\boldsymbol{x}=G'(\boldsymbol{z})$即对应所生成的图像或视频。而对于生成模型，其考量的是如何得到一个更好的映射G'，使得合成数据的分布P_{data}更加逼近真实数据的分布P_{real}。下面，根据这个生成映射G'的学习策略和表达形式的不同，分别介绍现有的几种生成模型的原理与框架[86]，分别包括对抗生成网络(generative adversarial network)、变分自动编码器(variational auto-encoder)、流模型(flow-based generative model)与扩散模型(diffusion model)。

如图1所示，对抗生成网络[87-88]作为一类大家最为熟悉的生成式模型，基本原理是利用协同对抗的方式，同时训练一个生成式网络G与判别式网络D。其中希望判别器能够一直分辨出输入的样本是来自真实分布还是模型生成的分布，其目标函数为$\max\left\{E_{x\sim P_{\text{real}}}\left[\log D(\boldsymbol{x})\right]+E_{z\sim P(z)}\left\{\log\left\{1-D\left[G(\boldsymbol{z})\right]\right\}\right\}\right\}$。前一项可以看作是最大化判别器识别出真实样本的可能性，后一项表达的是最大化判别器识别出生成器合成的假样本$G(\boldsymbol{z})$的概率。而生成器能够不断生成近似满足真实分布的样本来误导判别器作出错误的判断，其目标函数可以表示为$\min E_{z\sim P(z)}\left\{\log\left\{1-D\left[G(\boldsymbol{z})\right]\right\}\right\}$。因此，对抗生成网络的优化目标可以看成是在生成器与判别器之间进行最大最小化的博弈学习，总的目标优化函数为$\min\limits_{G}\max\limits_{D}\left\{E_{x\sim P_{\text{real}}}\left[\log D(\boldsymbol{x})\right]+E_{z\sim P(z)}\left\{\log\left\{1-D\left[G(\boldsymbol{z})\right]\right\}\right\}\right\}$。在对抗训练中，优化上述这一损失函数，实际上等价于在保持判别器最优的前提下，最小化真实样本分布与合成数据分布之间的Jensen-Shannon(JS)散度，即$D_{\text{JS}}(P_{\text{real}}, P_{\text{data}})=\frac{1}{2}\left[\int P_{\text{real}}\log\frac{P_{\text{real}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}+\int P_{\text{data}}\log\frac{P_{\text{data}}(\boldsymbol{x})}{P_{\text{real}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right]$。利用预定义好的分布距离的度量准则，生成器所生成的分布则会在这个度量下不断地逼近真实数据分布。但是通常会因为度量准则选择得不合适，或者两个网络的协同训练参数调节得不匹配，这种对抗的方式会导致训练不稳定，最终难以达到一个最优的平衡状态甚至是崩塌。例如JS散度虽然满足分布度量的基本定义与对称性准则(Kullback-Leiber(KL)散度则不满足两个分布的对称性准则)，但Arjovsky等[89]指出其会导致GAN的训练不稳定。这是因为两个分布P_{real}和P_{data}各自处在高维表示空间的两个低维流形上，所以可以很容易地找到一个分割超平面将这两个分布划分开来。这也意味着通过简单的优化，便可以得到一个完美的判别器，使得$D(\boldsymbol{x})=1, \forall\boldsymbol{x}\in P_{\text{real}}, D(\boldsymbol{x})=0, \forall\boldsymbol{x}\in P_{\text{data}}$，在这种情况下GAN的优化函数将一直近似为0，出现梯度消散的现象，导致对抗训练缓慢甚至停滞不前。而当判别器不够好时，给予生成器的反馈则不够准确，导致对抗博弈的过程中生成器无法学习到足够的“欺骗”能力。既然传统GAN所使用的JS散度或者KL散度在两个分布P_{real}和P_{data}不相交时无法提供有意义的度量，那么WGAN[90]提出了新的度量准则，即Wasserstein距离，进行更加平滑的度量。所谓的Wasserstein距离也称为运土距离(EM距离)，定义是将分布1变换为分布2所需要的成本。在Wasserstein距离的定义下，分布P_{real}和P_{data}之间的距离计算公式为$W(P_{\text{real}}, P_{\text{data}})=\inf\limits_{\gamma\sim\prod(P_{\text{real}}, P_{\text{data}})}E_{(x,y)\sim\gamma}\left(\|\boldsymbol{x}-\boldsymbol{y}\|\right)$，其中$\prod(P_{\text{real}}, P_{\text{data}})$是$P_{\text{real}}$

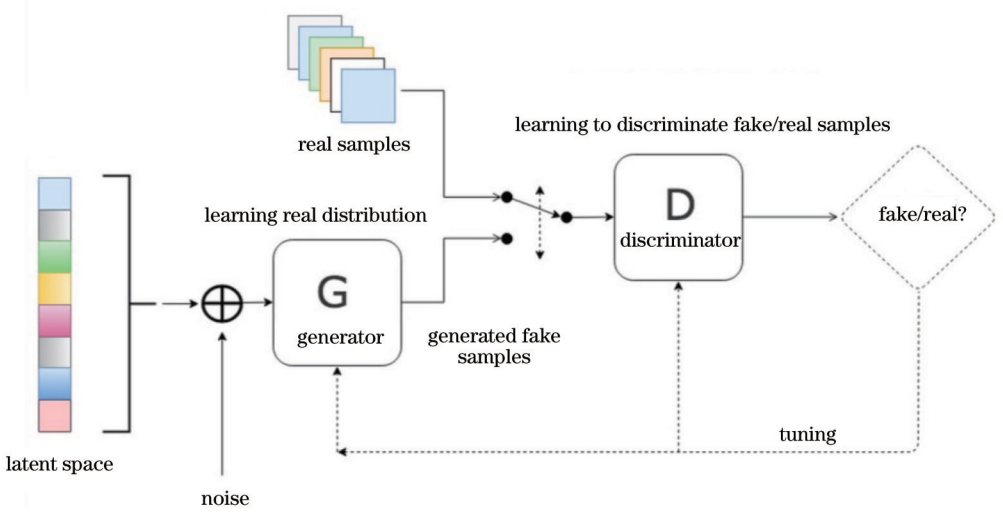

图1 对抗生成网络的原理示意图

Fig. 1 Overview of generative adversarial network (GAN) principle

和P_{data}的联合概率分布，$\gamma\sim\prod(P_{\text{real}},P_{\text{data}})$指其中一个传输策略。为了计算的方便性，通过Kantorovich-Rubinstein对偶性可以得到在Wasserstein距离下WGAN的目标优化函数，即$L_{\text{WGAN}}(P_{\text{real}},P_{\text{data}})=\max\limits_{\omega\in W}\left\{E_{x\sim P_{\text{real}}}\left[f_\omega(x)\right]-E_{z\sim P(z)}\left\{f_\omega\left[G(z)\right]\right\}\right\}$。在这里，WGAN的判别器不再是一个直接分辨真实样本和合成样本的分类器，而是通过学习一个K-Lipschitz连续的函数来帮助Wasserstein距离计算的分类器。

另外，变分自动编码器[91]和流模型[92]也是较为常见的两类生成式模型。如图2所示，变分自动编码器由编码器和解码器两部分组成，其中编码器负责将输入的图片压缩为一个低维的隐向量，而解码器负责将隐向量重构为与输入近似的图片。编码器与解码器通过利用变分推断进行参数优化得到训练，使得解码得到的图像分布逼近真实图像的分布。经过优化后，图像的生成过程可以通过随机采样隐向量并利用解码器来实现。具体地说，在假定由参数θ参数化的先验分布$p_\theta(z)$、似然分布$p_\theta(x|z)$和后验分布$p_\theta(z|x)$的情况下，变分自动编码器利用最大似然估计方式根据真实样本x_i对参数θ进行参数估计，表达式为$\theta^*=\arg\max\limits_\theta\sum\limits_{i=1}^{N}\log p_\theta(x_i)$。由于直接计算$p_\theta(x_i)$需要通过$p_\theta(x_i)=\int p_\theta(x_i|z_i)\,p_\theta(z_i)\mathrm{d}z$对所有的$z_i$进行遍历，这是十分困难的，所以需要引入一个编码器$q_\phi(z|x)$来逼近不易计算的后验分布$p_\theta(z|x)$。通过对$q_\phi(z|x)$和$p_\theta(z|x)$进行KL散度的度量并重新整理后可以得到$\log p_\theta(x)-D_{\text{KL}}\left[q_\phi(z|x),p_\theta(z|x)\right]=E_{z\sim q_\phi(z|x)}\left\{p_\theta(x|z)-D_{\text{KL}}\left[q_\phi(z|x),p_\theta(z)\right]\right\}$。当最大化等式左边的项后，即可以同时进行最大似然估计且最小化两个后验分布间的KL散度。因此，变分自动编码器的优化目标为$L_{\text{VAE}}(\phi,\theta)=-E_{z\sim q_\phi(z|x)}\left\{\log p_\theta(x|z)+D_{\text{KL}}\left[q_\phi(z|x),p_\theta(z)\right]\right\}$。

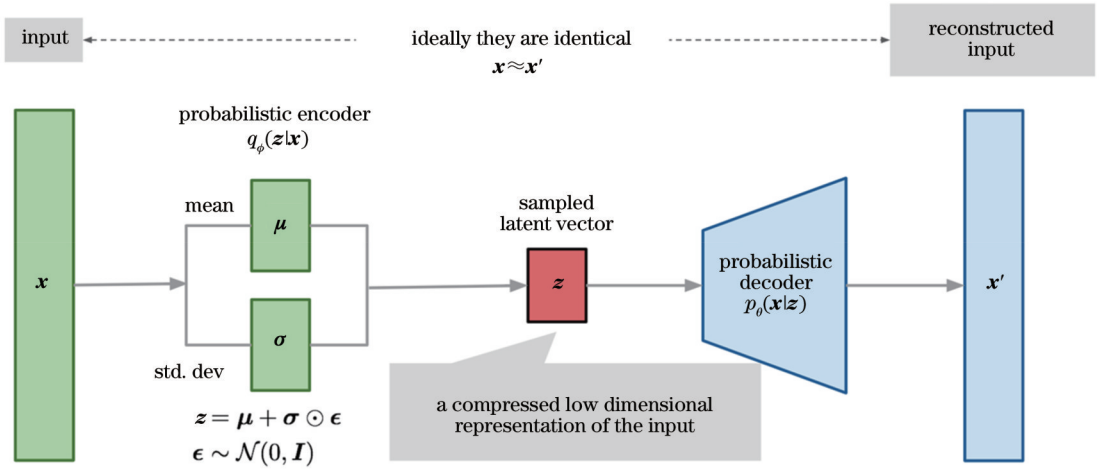

图2 变分自动编码器的原理示意图

Fig. 2 Overview of variational auto-encoder (VAE)

如图3所示，流模型则通过一系列的可逆映射$\{f_i\}$和较为简单的先验分布p_0来构建真实数据分布p_X与先验之间的可逆映射变换。但是往往因为不完美的逆变换，流模型生成的图像会产生形状的扭曲与畸变。与GAN和VAE不同的是，流模型通过显示的方式学习分布$p(x)$，损失函数即为样本的负对数似然，即$L(D)=-\dfrac{1}{|D|}\sum\limits_{x\in D}\log p(x)$。假设存在一个可逆的函

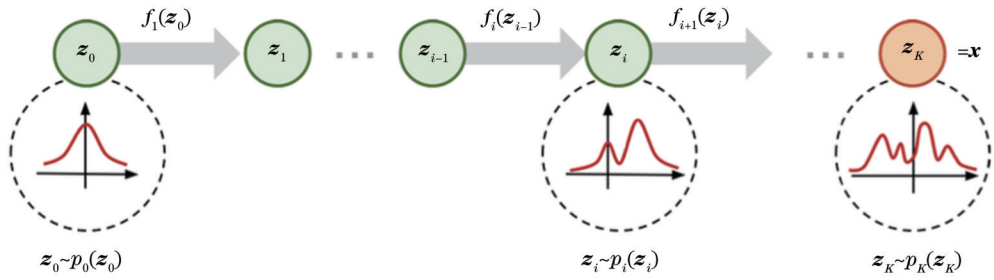

图3 流模型的原理示意图

Fig. 3 Overview of flow-based generative model

数或映射f,使得隐变量z可以通过f变化得到真实分布中的变量x,表达式为$f:\mathbb{R}^n\to\mathbb{R}^n$, s.t. $x=f(z)$; $z=f^{-1}(x)$,便可以得到$p_X(x)=p_Z[f^{-1}(x)]\left|\det\frac{\partial f^{-1}(x)}{\partial x}\right|$。其中$\frac{\partial f^{-1}(x)}{\partial x}$为$n\times n$的Jacobian矩阵。由$z=f^{-1}(x)$,可以得到$p_X(x)=p_Z(z)\left|\det\frac{\partial f(z)}{\partial z}\right|^{-1}$。在流模型中,通常这个可逆映射$f$都会用$\theta$进行参数化,所以变换的公式为$p_X(x;\theta)=p_Z[f_\theta^{-1}(x)]\left|\det\frac{\partial f_\theta^{-1}(x)}{\partial x}\right|$。流模型具有以下特点:映射变换的输入和输出的维度必须相同;映射变换f需要是可逆的;Jacobian矩阵及行列式的计算需要是可微且高效的。当f由深度神经网络表示时,每一个神经层即可以表达为一个映射过程。为了更清晰地解析流模型,这里给出一个经典的流模型real-valued non-volume preserving(RealNVP)[93]作为示例。RealNVP是通过堆叠一系列的可逆双射变换函数来实现归一化流的。对于每个双射变换$f:x\to y$,也称之为仿射耦合层,其输入数据被分为两个部分进行变换处理:输入的前d维特征保持不动;输入的第$d+1$到D维特征经过一个“缩放且移动”的仿射变换。其中缩放和移动的规模均由前d维特征的变换决定:$y_{1:d}=x_{1:d}$,$y_{d+1:D}=x_{d+1:D}\odot\exp[s(x_{1:d})]+t(x_{1:d})$,其中$s(\cdot)$和$t(\cdot)$是缩放和移动操作对应的变换函数,且它们均为$\mathbb{R}^d\to\mathbb{R}^{D-d}$空间上的函数。通过简单整理后可以得到$x_{1:d}=y_{1:d}$,$x_{d+1:D}=\left[y_{d+1:D}-t(y_{1:d})\right]\odot\exp\left[-s(y_{1:d})\right]$,对应的Jacobian矩阵为下三角矩阵,即$J=\begin{bmatrix}\mathbb{I}_d & 0_{d\times(D-d)}\\ \frac{\partial y_{d+1:D}}{\partial x_{1:d}} & \mathrm{diag}\{\exp[s(x_{1:d})]\}\end{bmatrix}$。其行列式可以简单地表达为对角线上的乘积,即$\det(J)=\prod_{j=1}^{D-d}\exp[s(x_{1:d})]_j=\exp\left[\sum_{j=1}^{D-d}s(x_{1:d})_j\right]$。在这基础上,RealNVP还可以使用多尺度架构,为高维度的样本输入构建一个更加高效的流模型。

近期,扩散模型[94]一类新型的基于似然估计的生成模型因为更为稳定的训练和生成图片惊人的高清程度进入了人们的视野。扩散模型是受自然界中的扩散现象启发而设计出来的生成模型,其正向加噪与反向去噪的训练原理如图4所示。与液体或气体的扩散类似,对于一幅图像,可以通过逐渐加入高斯噪声进行污染,其最终成为完全没有语义信息的随机噪声。通过在每个噪声扩散的步骤中进行变分推断,可以逼近这个高斯转移过程的逆过程。对所有扩散步骤都进行逆向估计后,便得到了扩散过程的生成流程:输入一张与原图同尺寸的完全噪声图,按照每个时间步所估计的逆向去噪过程对带噪图像进行去噪,迭代对应次数后,便可以实现图像的生成了。具体地说,假设x_0为真实分布$q(x)$中的一个数据样本$x_0\sim q(x)$,扩散模型的前向扩散过程由T步高斯转移定义,即$q(x_t|x_{t-1})=\mathcal{N}(x_t;\sqrt{1-\beta_t}\,x_{t-1},\beta_t\boldsymbol{I})$,其中$x_1,x_2,\cdots,x_T$为不断被噪声污染的图片样本,$\{\beta_t\in(0,1)\}_{t=1}^T$决定了每步添加的噪声强度。通过连续$T$步的噪声化干扰,原始图像中具有辨识度的特征逐渐消失,并且当$T\to\infty$时,x_T将趋于一个各向同性的标准高斯分布。假设上述的前向扩散过程可以逆转,那么便可以通过从标准高斯分布中随机采样一个高斯噪声$x_T\sim\mathcal{N}(0,\boldsymbol{I})$并输入这个反向过程来生成一个近似满足真实图像分布的样本。类似于VAE中的推导,尽管已知正向转移$q(x_t|x_{t-1})$,无法直接得到其逆转过程$q(x_{t-1}|x_t)$,但是可以通过类似的变分推断用另一个参数化的分布$p_\theta(x_{t-1}|x_t)$来逼近$q(x_{t-1}|x_t)$,表达式为$p_\theta(x_{t-1}|x_t)=\mathcal{N}[x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)]$,$p_\theta(x_{0:T})=p(x_T)\times\prod_{t=1}^T p_\theta(x_{t-1}|x_t)$。其中,尽管$q(x_{t-1}|x_t)$无法直接计算,但当$x_0$已知时,$q(x_{t-1}|x_t,x_0)$是可知的。由于扩散模型需要在每步之间逼近$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t)$,并且这个逼近过程也是通过变分推断所估计的,所以

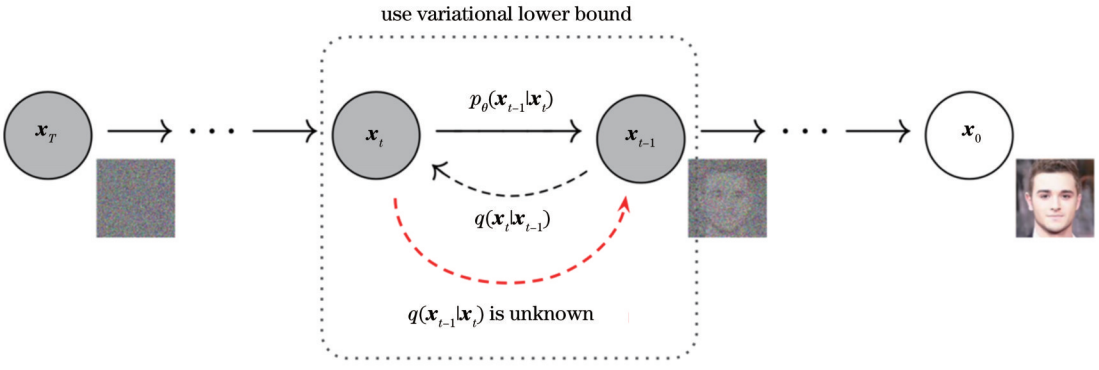

图4　扩散模型的原理示意图

Fig. 4　Overview of diffusion model

也可以认为扩散模型是一个渐进迭代的多步变分自动编码器。但是不同的是，通过重参数化，可以推导出实际上在优化扩散模型的变分下界时，等价于在每个时间 t 处估计一个噪声 $\boldsymbol{e}_t$，表达式为 $L_t=E_{t\sim[1,T],\,\boldsymbol{x}_0,\,\boldsymbol{e}_t}\left[\left\|\boldsymbol{e}_t-\boldsymbol{e}_\theta\left(\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\,\boldsymbol{e}_t,\,t\right)\right\|^2\right]$。由于生成图片的真实度、样本的多样性、灵活的可控性，这一类生成式模型进一步地革新了图像和视频生成领域的研究。作为总结，图 5 给出了 4 种生成模型的原理对比示意图，以便读者更好地理解与区分这几种模型的差异性与相似性。

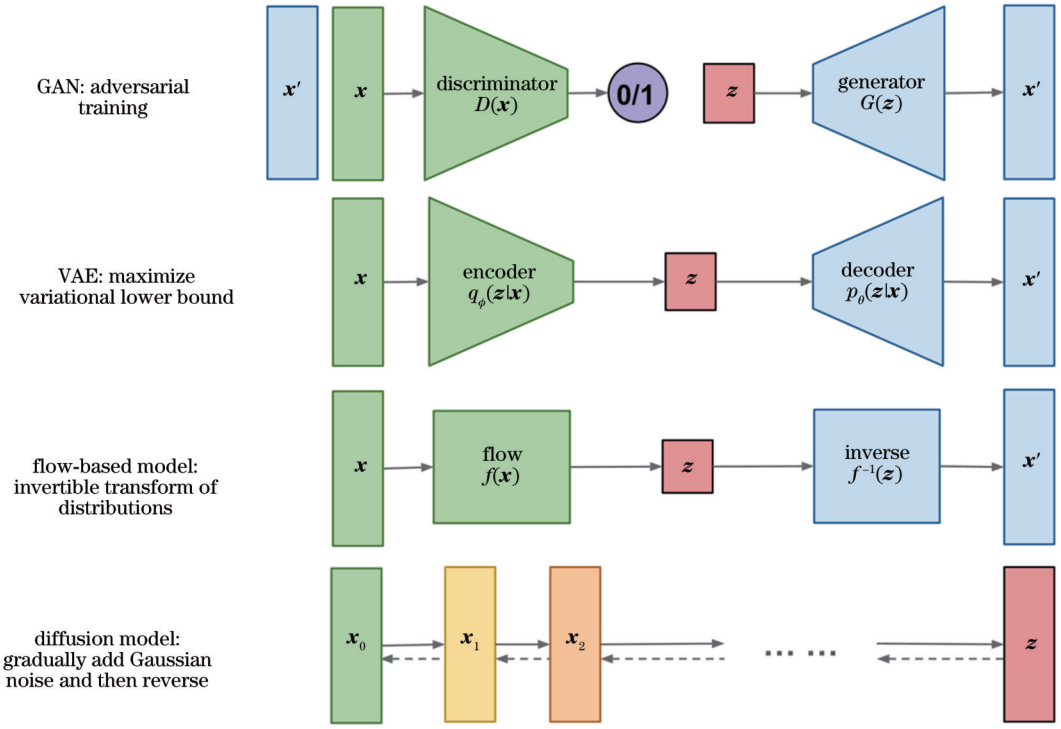

图 5 各种生成模型的原理对比示意图

Fig. 5 Comparison of different generative models

3 图像与视频生成的前沿方法

为了更全面、清晰地展示图像和视频生成领域的最新进展，对其进行了流派分类，如图 6 所示。在图像生成模型方面，分别介绍了几种文生图预训练模型和基于提示文本的图像编辑模型。对于基于提示文本的图像编辑模型，根据其利用的预训练的文生图模型将输入图片转换为目标图片的方式，将现有工作分类为基于微调权重的图片定制化、基于文本嵌入学习的图片定制化、基于掩码区域控制的图像编辑、基于改变提示文本的图像编辑、基于文本嵌入插值的图像编辑。在视频生成方面，由于动态视频生成相对静态图片所涉及

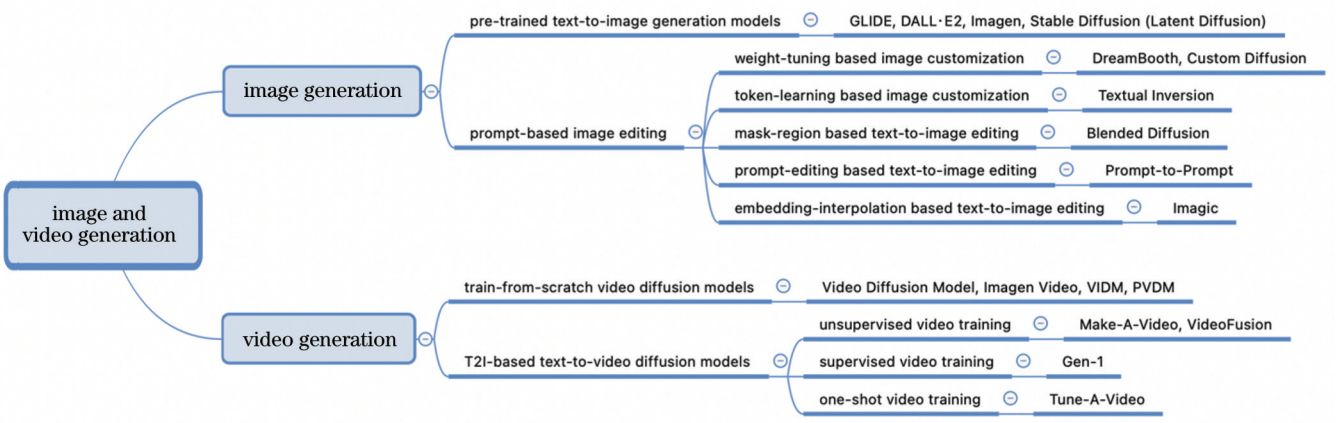

图 6 图像与视频生成方法流派概览

Fig. 6 Overview of image and video generation models

的问题更多且更复杂，目前的研究重点仍然集中在如何训练生成一个高清、稳定、连贯的视频。根据视频生成模型的训练方式，将现有工作划分为完全从头开始训练的视频生成模型和基于扩展预训练文生图模型的视频生成模型。其中，根据在训练过程中使用视频数据的方式，基于扩展预训练文生图模型的视频生成模型又可以分为基于无监督视频的训练、基于有监督视频（使用视频对应的文本描述作为监督信号）的训练、基于单一视频输入的生成训练（仅使用一个视频作为时序训练数据，而非使用大规模视频数据集）。接下来，将按照上述脉络框架，逐一介绍图像和视频生成领域的前沿进展，并对各个方法的优缺点进行探讨分析。

2021 年，OpenAI 公司推出了 Ablated Diffusion Model（ADM）[95]，揭示了扩散模型正式在图像生成（尤其是无条件、以类别为条件的图像生成）领域打破了 GAN 在图像生成领域的统治地位，进而成为当下最强有力的图像生成器。以类别为条件的图像生成如图 7 所示，指给定某一类别标签，指定生成对应类别的单一物体。ADM 证实现有 GAN 的成功大多依赖于其优越的网络训练技巧和网络结构设计，而当将 GAN 训练的网络结构设计中的一些技巧，例如增加网络的深度宽度比、增加注意力机制的头数、在多尺度而不是单一尺寸的特征上使用注意力机制等，迁移至扩散模型的噪声估计 U-Net 上后，扩散模型也能够在大规模 ImageNet 数据集上实现比 GAN 更好的无条件生成、以类别为条件的图像生成效果。除了具有更高的保真度，扩散模型生成的图片也具有更高的多样性和可控性，可以通过额外训练一个分类器，对扩散模型的去噪阶段进行偏移量修正的引导，使其生成出满足某一类别约束的图像。

图 7　以类别为条件的图像生成[95]

Fig. 7　Class-conditioned image generation[95]

由于 ADM 的训练是针对无条件或者以类别为条件的图像生成的，对图像内容的可控性较差，因此 OpenAI 公司的 GLIDE[98] 和 DALL·E2[96] 及 Google 公司的 Imagen[99] 模型几乎同期提出了将大规模语言模型与图像扩散生成模型结合的方法，实现基于文本提示的可控图像生成（如图 8 所示，模型根据提示文本描述的内容，生成对应的图像）。它们的成功证实了：在强有力的大规模预训练语言模型的加持下，扩散模型也具有非常惊人的从文本生成图像的能力。通过将跨模态对比预训练模型 CLIP[100] 或者在大规模语料数据上预训练的语言模型输出的文本嵌入作为生成条件，扩散模型便具备了文生图的能力，能够根据用户输入的提示文本输出具有对应语义的图片。其中，GLIDE 与 Imagen 分别基于 CLIP 的文本编码器和语言模型 T5-XXL，使用 classifier-free guidance 方式来实现文本控制的引导训练与图文对齐。而 DALL·E2 采用一种层次化预测的方式：给定输入的文本提示 $\boldsymbol{y}$，在生成图像 $\boldsymbol{x}$ 时，将其对应的 CLIP 图像嵌入 $\boldsymbol{z}_i$ 和文本嵌入 $\boldsymbol{z}_t$ 作为中间介质，把生成过程 $p(\boldsymbol{x}|\boldsymbol{y})$ 分解为一个先验模型 $p(\boldsymbol{z}_i|\boldsymbol{y})$ 和解码模型 $p(\boldsymbol{x}|\boldsymbol{z}_i, \boldsymbol{y})$，即 $p(\boldsymbol{x}|\boldsymbol{y})=p(\boldsymbol{x}, \boldsymbol{z}_i|\boldsymbol{y})=p(\boldsymbol{x}|\boldsymbol{z}_i, \boldsymbol{y})\times p(\boldsymbol{z}_i|\boldsymbol{y})$。其中，先验模型 $p(\boldsymbol{z}_i|\boldsymbol{y})$ 旨在根据输入的文本提示 $\boldsymbol{y}$ 预测图像对应的 CLIP 图像嵌入 $\boldsymbol{z}_i$，解码模型 $p(\boldsymbol{x}|\boldsymbol{z}_i, \boldsymbol{y})$ 旨在根据 CLIP 的图像嵌入 $\boldsymbol{z}_i$ 和文本提示 $\boldsymbol{y}$ 来生成最终的图像 $\boldsymbol{x}$。相比直接使用文本编码器的 GLIDE 和 Imagen，这种层次化编码的方式可以在 CLIP 空间得到图像和文本的紧致化表达，从而实现更加精准、灵活的控制，如对两张图像的 CLIP 嵌入进行插值。从图 8 可以看出，在给定一句文本提示的情况下，DALL·E2 利用跨模态的 CLIP 模型对文本语义与图像语义进行对齐，结合生成扩散模型和超分辨率扩散模型，能够生成十分高清的、符合文本描述的高分辨率图像。

尽管 DALL·E2 和 Imagen 等模型的成功证实了扩散模型能够在大规模训练下获得强大的文生图能力，

图 8　以文本为条件的图像生成[96]

Fig. 8　Text-conditioned image generation[96]

但是由于它们均是直接在像素空间上进行扩散训练的，并且隐变量多、维度高，无法高效地训练高分辨率、支持多种模态控制的扩散模型。随着隐式扩散模型[97]的横空出世，它通过将扩散模型的训练空间从高维像素空间转变为在 VQ-VAE 映射[101]下的低维隐空间，大大提高了扩散模型的计算效率，使高效训练多样化、支持多种控制、可以生成高分辨率图片的扩散模型（如 class-condition、text-to-image、layout-to-image）成为了可能。其中，跨模态交互与图文语义对齐是通过交叉注意力机制（cross-attention）的方式实现的。基于隐式扩散模型，StabilityAI 公司进一步推出了开源的稳定扩散模型（Stable Diffusion）[97]，图 9 展示了 Stable Diffusion 模型的文生图效果，将从文本生成图像的效果提高到一个更高的高度，也迎来了 AI Generated Content（AIGC）的大爆发。

由于 Stable Diffusion 的推出，图像生成的可控性、可定制性、可编辑性得到了大大提升。例如，Google 提出的 DreamBooth[102]、Nvidia 提出的 Textual Inversion[103]、Adobe 提出的 Custom Diffusion[104]方法，各自以微调 Stable Diffusion 网络权重与文本嵌入学习（embedding learning）的角度，使得预训练好的 Stable Diffusion 模型可以学习到用户输入图片中的概念，结合用户输入的文本提示（prompt）进行图片定制化创造。

图 10 展示了 DreamBooth 的生成结果，其根据用户输入的少量几张图片以及图片所述的类别，以微调网络权重的方式来使预训练好的 Stable Diffusion 网络生成出用户期望得到的定制化图片。其中，在标定用户输入图片中的主体时，需要通过其所属的类别结合一个未使用过的标识符来指代用户期望指代的物体。

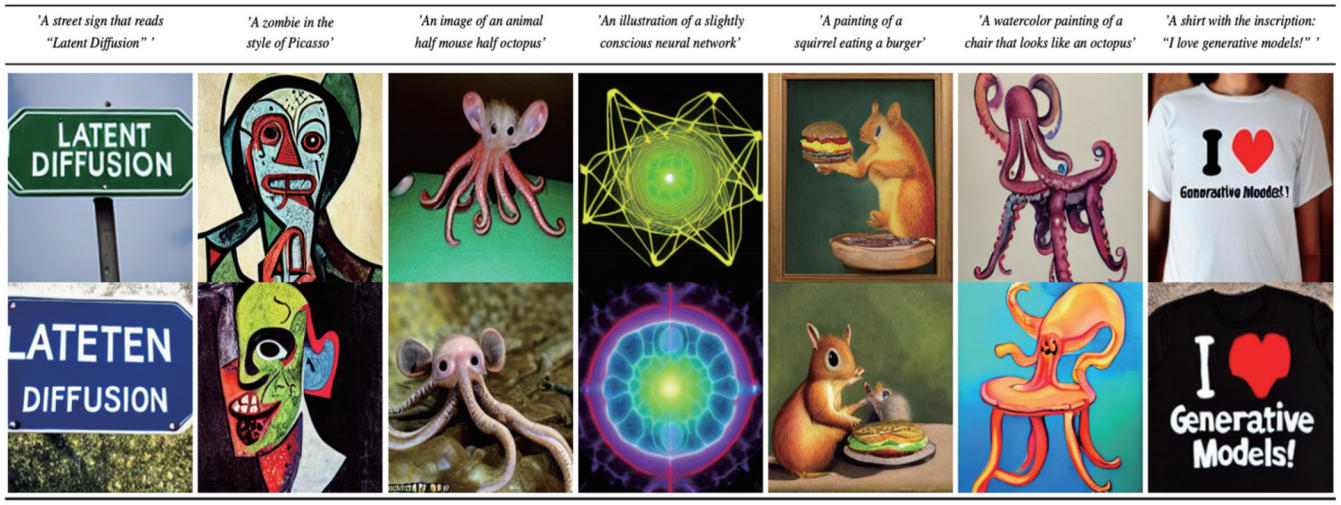

图 9 Stable Diffusion 的文本-图像生成效果[97]

Fig. 9 Text-to-image generation results of Stable Diffusion[97]

图 10 基于微调权重的定制化创造[102]

Fig. 10 Weight-tuning-based image customization[102]

另外，为了防止微调后的网络在当前用户输入的图片上过拟合，语言模型失去解析同类其他物体的能力，DreamBooth 中具有 class-specific prior preservation loss，即在微调时，利用原始的 Stable Diffusion 模型将同类别物体的生成结果作为正则化，防止模型遗忘同类其他物体。由于 DreamBooth 需要在用户输入的少量样本中微调 U-Net 中的网络权重，模型容易在这些输入中过拟合，图 11 所示的 Textual Inversion 则将网络权重固定住，通过文本嵌入学习的方式，将用户输入的图片转换为一个伪单词符号，利用标准噪声回归的重建损失函数，令伪单词符号的语义与用户输入的图片对应起来。通过这种嵌入学习的方式，用户可以将模型学习到的这个嵌入符号任意地插入到其提示文本中，从而生成用户所需要概念的定制化图片。从图 11 可以看到，学习到的伪单词符号包含了输入图片中物体的姿态、风格、内容等信息，所以定制化的图片具有高度的可控性和可编辑性。由于 DreamBooth 和 Textual Inversion 只能支持单个物体概念的提取与定制，使得用户在生成图片时无法对多种概念同时进行定制并将它们整合到同一张新图片里。因此 Custom Diffusion 在它们的基础上进一步引入了多概念组合学习的图片定制化生成框架。其中，Custom Diffusion 指

出在DreamBooth定制化概念的学习中，网络更多地是在更新交叉注意力机制（cross-attention）中的key和query特征变换矩阵，因此为了实现高效、快速的多概念学习，其在学习输入图片中的定制化概念时只微调这两组特征变换矩阵。同时，在这种高效微调的基础上，它通过引入约束优化的方式，进而支持同时学习多个概念的合成，并能对这些概念同时在一张定制化图片上进行组合展示，如图12所示。

图11 基于文本嵌入学习的定制化创造[103]

Fig. 11 Token-learning-based image customization[103]

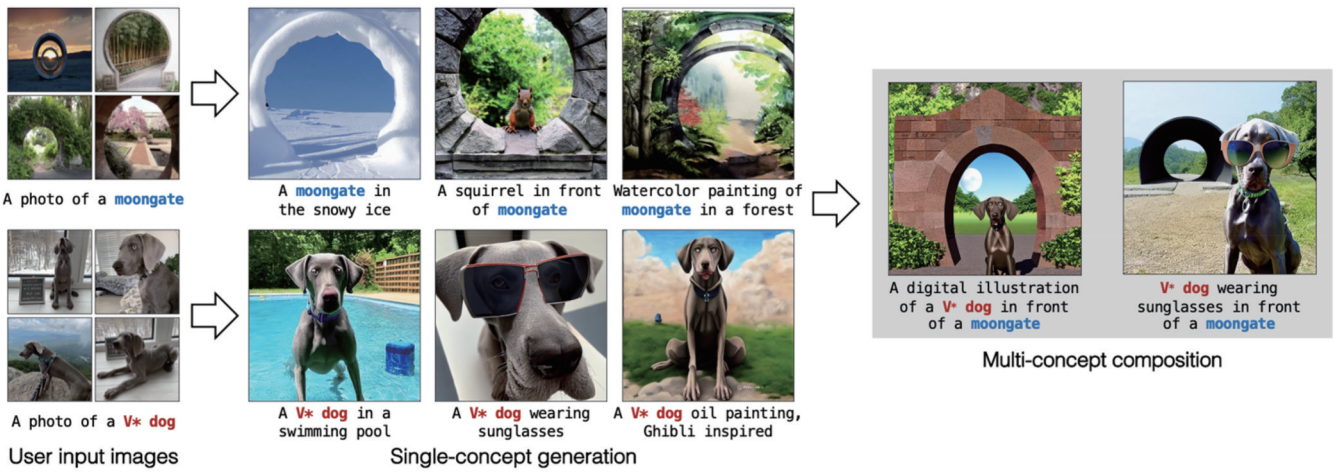

图12 支持多概念合成的定制化创造[104]

Fig. 12 Image customization with multi-concept composition[104]

除了定制化学习用户输入图像中的概念与风格之外，Blended Diffusion[105]、Prompt-to-Prompt[106]、Imagic[107]还支持各种形式的基于预训练Diffusion模型的图片编辑。例如Blended Diffusion允许用户输入一张图像、一个掩码区域、一句文本提示，通过利用混合新文本引导的隐变量的方式实现对输入图像的自然编辑。如图13所示，Blended Diffusion基于用户输入的提示文本，对掩码区域中的内容进行引导编辑，使得掩码区域中生成用户期望的内容，达到图像编辑的效果。具体来说，输入一张自然图像、待编辑的区域对应的掩码、待编辑区域中的修改内容，Blended Diffusion可以通过classifier-guidance的方式，计算掩码区域图片与提示文本的CLIP相似度，并基于这个CLIP相似度引导生成一张掩码区域内满足提示文本的图片，并在去噪生成的每个时间步对当前步引导生成的带噪图片与同样噪声级别下的原始图片利用掩码进行融合，融合的结果作为下一步去噪的输入。通过这样迭代融合与去噪后，便能得到掩码区域内满足提示输入要求、掩码区域外保持与原图一致的编辑后的图片。Prompt-to-Prompt利用对注意力注入的方式，可以只改变输入的

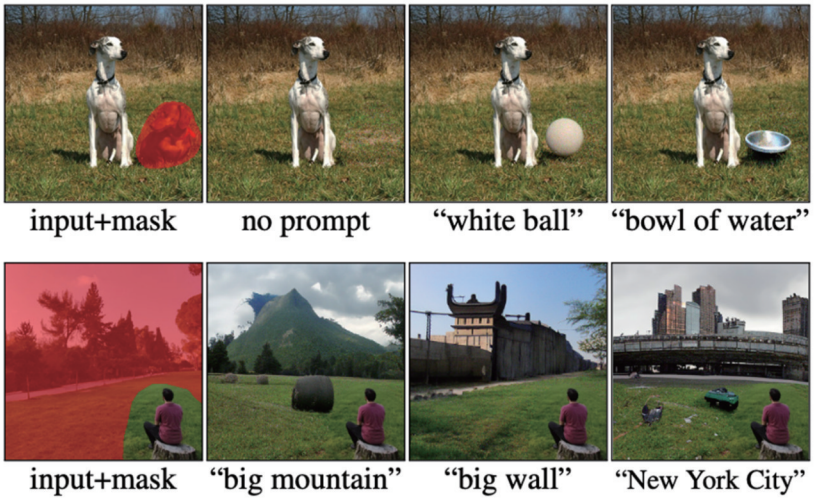

图 13 基于掩码区域的文本-图像编辑[105]

Fig. 13 Mask-region-based text-to-image editing[105]

提示文本来实现对生成图像的编辑。如图 14 所示，Prompt-to-Prompt 对初始提示进行形容词强度调整、单词替换、图片风格指定、物体描述精炼化的修改，并进行相应的语言符号所对应的注意力注入，便能够实现无需掩码的灵活图像编辑。Imagic 通过对输入的图像进行文本嵌入优化，寻找其在重建语义上最接近的文本嵌入，然后对其与目标编辑的文本嵌入进行插值，来实现对真实图像的编辑。在进行文本嵌入优化时，Imagic 使用类似 Textual Inversion 中的方式，固定住预训练的扩散模型的参数，然后利用标准的噪声重建损失函数来优化输入的文本嵌入。在得到文本嵌入后，通过同样的优化目标，固定学到的文本嵌入，并进一步优化扩散模型的网络参数，使得模型具备重建源图像的能力。最终，在源图像学到的文本嵌入和待修改的目标文本嵌入之间进行线性插值，便可以得到修改后的图像。如图 15 所示，输入一张自然图像时，Imagic 在优化得到图像所对应的文本嵌入后，对其与输入的目标编辑的不同图像文本嵌入进行插值，便可以得到对输入图像的不同编辑效果，同时不会失去图像的全局和谐度与保真度。

图 14 基于改变提示文本的文本-图像编辑[106]

Fig. 14 Prompt-editing-based text-to-image editing[106]

在视频生成方面，Google 提出的 Video Diffusion Model[109]首次利用 3D U-Net[110]对时空噪声进行建模，自然地将扩散模型应用到了视频上。如图 16 所示，Imagen Video[108]进一步地通过级联扩散模型的方式，实现了从文本到高清视频的生成。从这之后，便开启了扩散模型在视频生成方面的研究。由于 3D U-Net 的计算量大、计算效率低，视频扩散模型难以在大规模视频数据上进行训练。因此，如何更高效地在视频数

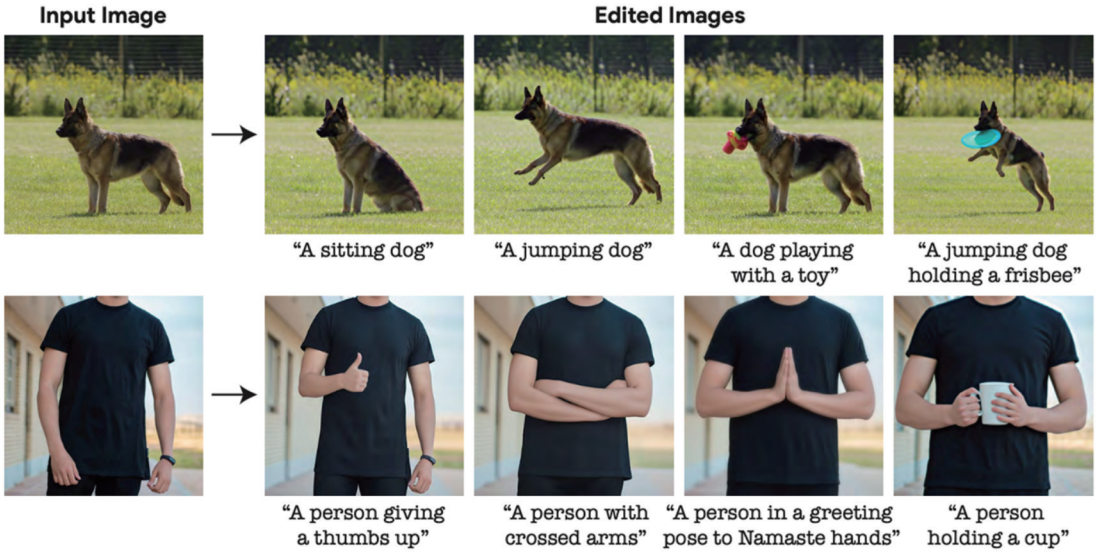

图 15 基于文本嵌入插值的文本-图像编辑[107]

Fig. 15 Embedding-interpolation-based text-to-image editing[107]

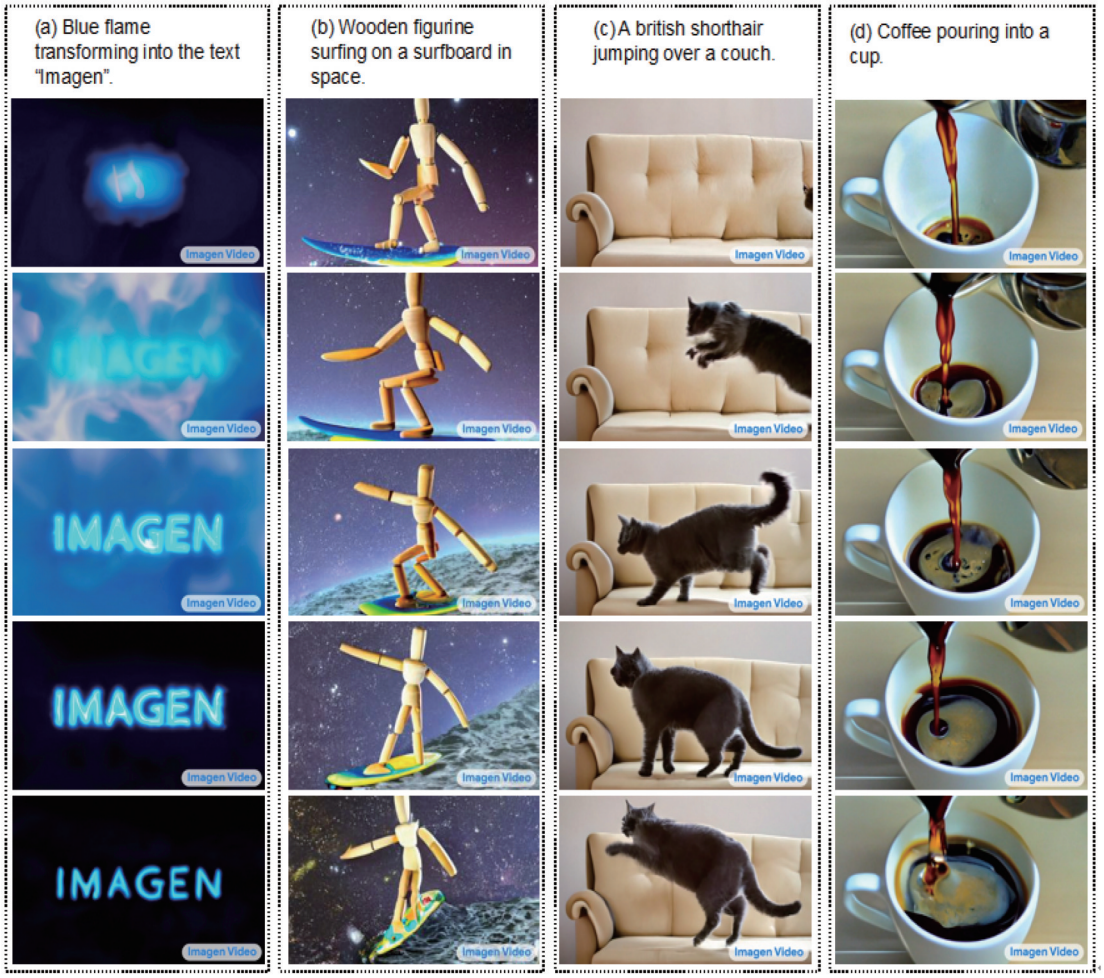

图 16 Imagen Video 的视频生成[108]

Fig. 16 Generated videos of Imagen Video[108]

据上直接训练视频扩散模型也成为了视频生成需要攻破的问题之一。如图 17 所示，VIDM[111]认为视频是由内容和运动两种信息组成的，因此利用了两个扩散模型来分别生成视频的第一帧，根据生成的第一帧来估计后续帧。如图 18 所示，PVDM[112]沿用了隐式扩散模型[97]的隐空间建模框架，用变分自动编码器将视频

的三维立体表示分解为3个二维的隐变量，从而可以避免直接在三维隐空间上进行高复杂度的扩散计算，提高模型训练的高效性。这一类视频生成模型只在视频数据上同时学习每一帧内容信息与动态信息，但由于帧间内容具有巨大冗余性，这种训练范式会影响视频内容学习的效率。

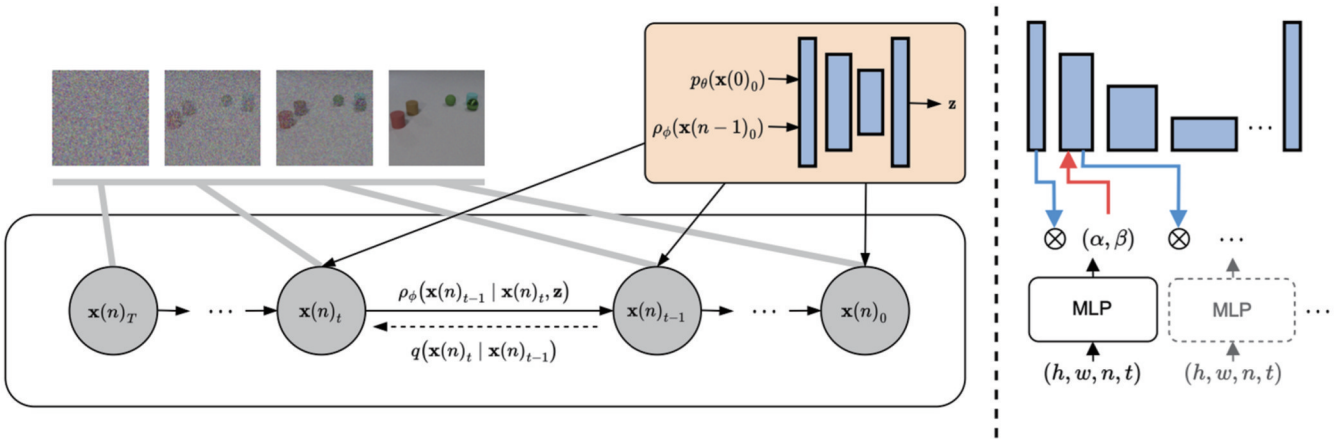

图 17 利用两个扩散模型分别来生成视频的内容和动作信息的VIDM框架[111]

Fig. 17 VIDM framework using two diffusion model to generate video content and action information respectively[111]

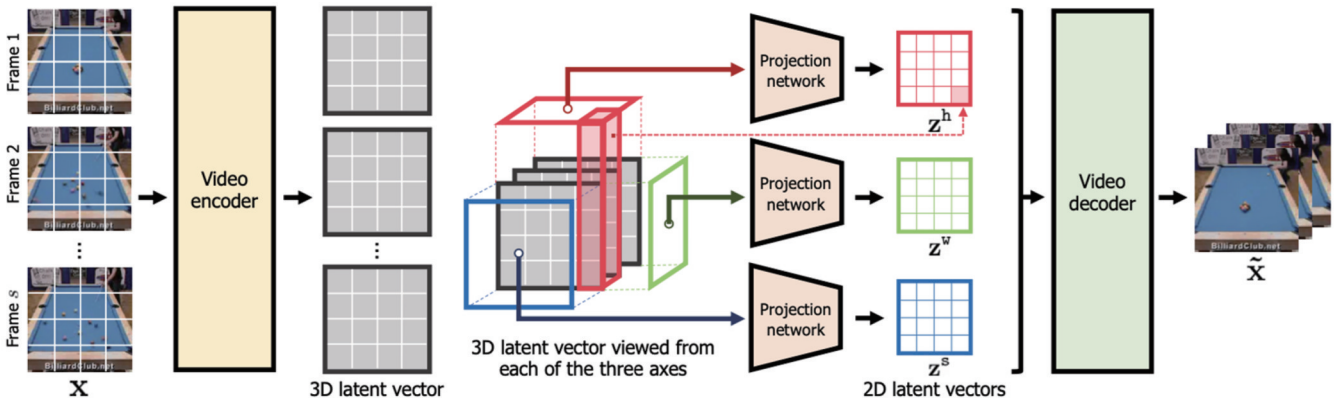

图 18 将视频表示为3个二维隐变量，从而利用二维扩散模型进行训练的PVDM框架[112]

Fig. 18 PVDM framework that represents the video as three two-dimensional hidden variables, and thus uses the two-dimensional diffusion model for training[112]

直接在视频数据上从头开始训练的效率会受限于视频在内容信息上的高度冗余性，因为模型需要同时学习生成空间内容与捕获帧间的时序信息。一种更为先进的视频生成框架是利用预训练好的文本-图像生成模型（如Stable Diffusion）来提供内容先验，生成视频的基础帧，然后在视频数据上只学习动态信息，从而不需要在视频数据上直接学习提示文本-视频的内容对齐信息，同时又能继承现有图像生成模型的高清、多样、跨模态对齐的优势。如图19所示，Meta AI提出的Make-A-Video[113]利用文本-图像扩散模型来生成视频的基础帧，然后为文本-图像扩散模型中的卷积层与注意力层扩展了时间维度，并在海量无监督的视频数据上训练扩展的时序层，大大提升了视频扩散模型的训练速度。

视频扩散模型训练的另一个问题是，现有模型均假设所有帧生成时使用的噪声是独立采样的，这忽略了建模帧间的关联性，也使得生成的视频帧间不够连贯。为了解决这个问题，阿里巴巴达摩院提出的VideoFusion[114]则在现有框架的基础上，进一步分析了基础帧与视频其余帧的噪声关联性，将每一帧的噪声分解为共享噪声（也即基础帧噪声）与独立的残差噪声，如图20所示，其中共享的噪声对应了视频的内容，残差噪声对应了当前帧的运动信息。基于这个分析，基础帧（内容）使用预训练的文本-图像生成模型来生成，而其余帧则利用专门的残差扩散模型来生成对应的残差噪声。但是这种在无监督的视频数据上的训练会导致提示文本与视频之间的对齐效果较为有限，只能生成结构简单、幅度不大的动作。

由于使用无监督的视频数据进行生成训练会导致可控性较差，Gen-1[115]在使用文本监督的视频进行训练之外，还引入了结构性控制，将提示文本和深度信息分别作为视频的内容引导和几何结构引导。如图21

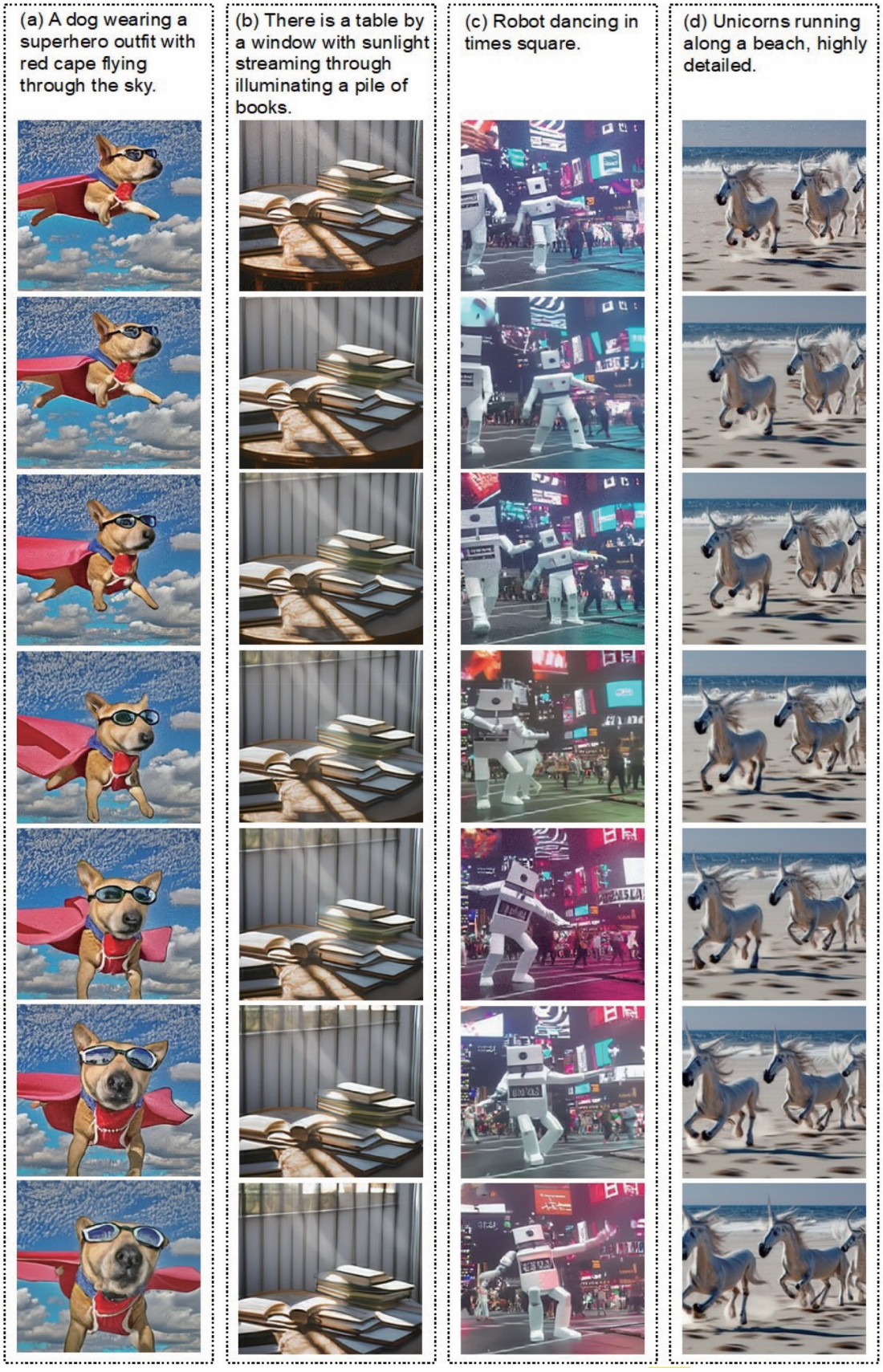

图 19 Make-A-Video的文本-视频生成效果[113]

Fig. 19 Text-to-video generation results of Make-A-Video[113]

所示，Gen-1将提示文本和深度信息作为用户的提示输入，并根据这些信息进行视频的生成与编辑。Gen-1通过在Stable Diffusion中插入时间序列层，并在图片和视频上进行联合训练，可以使得生成的视频具有更

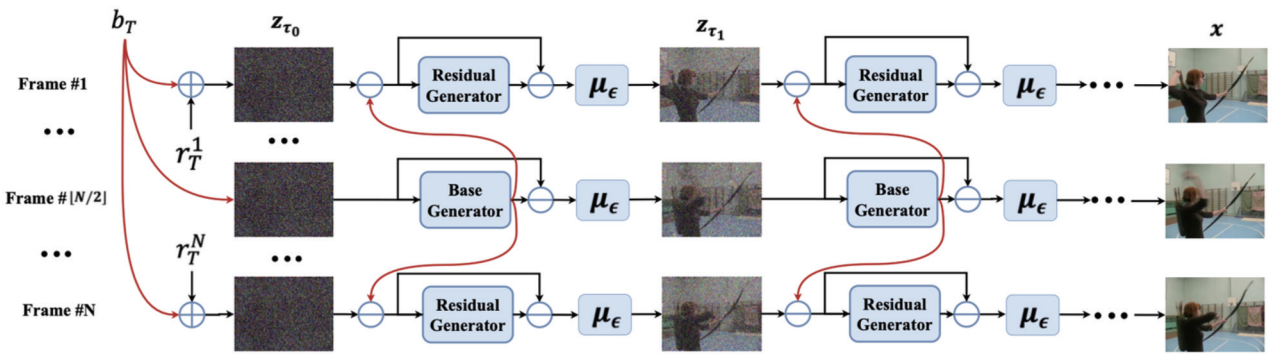

图 20 使用预训练的文本-图像扩散模型生成基础帧，并在视频数据上训练残差噪声生成器的 VideoFusion 框架[114]

Fig. 20 VideoFusion framework that uses pre-trained text-to-image diffusion model to generate base frame and uses video data to train a residual noise generator[114]

高的结构一致性和时间内容一致性。针对文本信息，Gen-1 使用交叉注意力机制进行融合，而针对结构信息，Gen-1 采用了对深度图与图像进行通道堆叠的方式进行对齐融合。从图 21 可以看到，Gen-1 不仅可以根据用户的提示输入生成对应的视频，还可以基于用户输入的文字或者图片对视频的内容、动作主体进行引导、编辑，实现对动作与内容的高度可控。

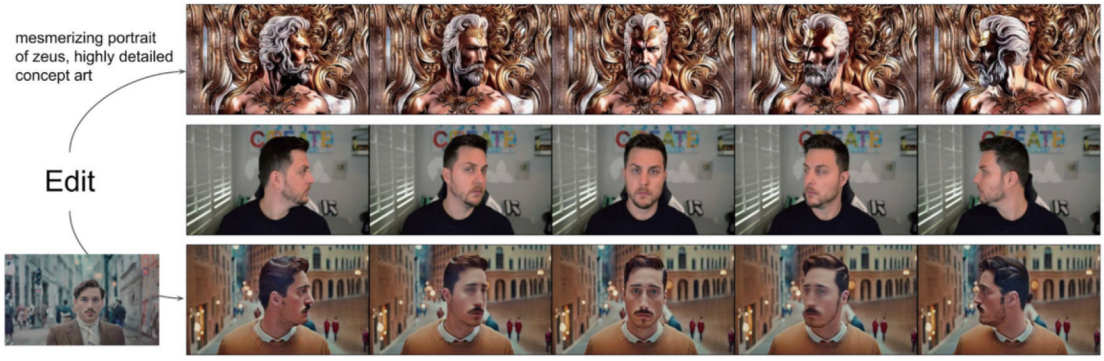

图 21 基于图像或者文本的视频编辑[115]

Fig. 21 Video editing based on input image or text prompt[115]

前面提到的视频生成模型的训练均依赖大量的视频数据（无论有无标签监督），训练的计算开销非常大。因此，近期腾讯提出了另一种视频生成的范式，即基于单个视频训练的视频生成，并推出了相应的模型 Tune-A-Video[116]。Tune-A-Video 证实了，通过扩展已经训练好的从文本到图像的扩散模型（Stable Diffusion），在其交叉注意力机制上增加跨帧的时间维度注意力层，便可以从单一视频训练样本中学会从提示文本中生成在时间上连贯的视频。图 22 展示了 Tune-A-Video 的单一视频微调效果，可以看到尽管只有一个用于微调网络权重的训练视频，模型也能很好地学习到动作姿态，并利用预训练好的文本-图像扩散生成模型学到的类别概念，对视频中的动作主体进行替换，同时保持动作在时间维度的连续一致性。

4 效果评价

由于生成模型的效果评价较为主观并且现有的评价指标并不能反映人类肉眼的视觉效果，因此图像与视频编辑、定制化生成的工作大多以可视化结合用户测评的方式进行效果评价。故本文难以利用统一的定量指标来总结现有工作生成效果的好坏。但是，对于文生图预训练模型和基于直接训练的视频扩散模型，可以给出它们在某些数据集下的定量评测结果。

对于文生图预训练模型，给出在 MS-COCO 测试数据集[117]上按照提示文本生成的 256 × 256 的图片的 Fréchet inception distance（FID）指标[118]结果，如表 1 所示。

表 1 在 MS-COCO 数据集上不同文生图预训练模型的 FID 比较

Table 1 FID comparison of different text-to-image pre-trained models on MS-COCO dataset

Method	FID ↓
LAFITE[119]	26.94
DALL·E[120]	17.89
LDM[100]	12.63
GLIDE[96]	12.24
DALL·E2[97]	10.39
Imagen[98]	7.27

图 22 基于单视频训练的文本-视频生成[116]

Fig. 22 One shot text-to-video generation[116]

对于视频生成模型，由于不同的工作使用的训练视频数据差异较大，且它们的结果评测以逐帧可视化结合用户测评的方式为主，它们在UCF-101数据集[121]和Sky Time-lapse数据集[122]上对基于类别生成的视频的FVD（Fréchet video distance）[123]、FID[118]、IS（inception score）[124]指标如表2所示。

表2 在UCF-101数据集与Sky Time-lapse数据集上，基于类别输入的视频生成模型的性能对比

Table 2 Performance comparison of different class-to-video generation methods on UCF-101 dataset and Sky Time-lapse dataset

Method	FVD ↓ (Sky Time-lapse 256×256)	FVD ↓ (UCF-101 256×256)	FID ↓ (UCF-101 128×128)	IS ↑ (UCF-101 128×128)
MoCoGAN[125]	206.6	1821.4		12.42
VideoGPT[126]	222.7	2880.6		24.69
MoCoGAN-HD[127]	164.1	1729.6	838	32.36
DIGAN[128]	83.1	471.9	655	29.71
VIDM[111]	57.4	294.7	306	53.34
PVDM[112]	55.4	343.6		74.40

5 现有方法的局限性

尽管现有的图像和视频生成已经取得了一定的进展，但它们依旧有一些局限性：1）尽管基于文本嵌入与图像特征的交叉注意力机制可以实现跨模态的控制，但是当输入的提示文本较为复杂时，现有方法依然没办法生成精准的、满足人们期望的图像与视频，会产生部分物体或者属性的丢失、物体与描述性属性的混淆等问题；2）在进行定制化概念的学习与图像编辑时，没办法实现连续的编辑，并且会出现忘记前面所学概念的灾难性遗忘现象；3）当生成的结果不能使用户满意时，模型无法得到反馈修正固有的缺陷，只能通过不断地试探性修改提示输入来尝试得到更符合需求的结果；4）当前视频生成相对图像生成取得的成功来说，依然不够成熟稳定，主要体现在帧间抖动较大，无法生成时间长、物体与场景动态变化较大、具有多阶段运动信息的视频；5）与图像领域中的文生图预训练模型相比，仍缺乏视频生成的大规模预训练模型，现有的视频生成模型仍无法提取到解耦的内容表征与运动性信息，以至于在做下游的视频编辑等任务时，无法针对性地提取源视频的运动主体的相应信息，无法进行定制化的视频生成。

6 未来改进的方向和展望

针对上述提到的局限性，本文提出一些未来可能的研究方向：1）利用自监督生成式学习，使得模型能够更好地捕获图像中上下文的信息，结合更强大的大规模预训练语言模型和跨模态对齐策略，实现更为精准

的生成与编辑；2）引入可扩展式的生成式框架，通过混合专家模型（Mixture of Expert）机制扩展扩散模型，使模型可以不断地吸收新概念（针对新的概念增加新的子专家模型），并同时保有已有的知识（固定已有的子专家，并通过动态路由机制激活选择），实现可持续可扩展的生成式学习；3）引入多轮交互机制，针对训练好的视觉生成模型将 ChatGPT 等人机对话模型做提示微调（prompt-tuning），使得模型能在多轮人机交互中利用用户的反馈不断修正前几轮生成中不满意的地方，同时基于新的需求对生成的图像或视频做进一步的润色；4）进一步改进现有的视频生成模型中的时序模块，将自回归式的稀疏时序注意力机制修改为双向细粒度时序注意力机制，同时将提示文本的固定嵌入表示改进为随着视频帧变化的动态文本嵌入，以在不同时间戳上均可以进行动态对齐，突破当前视频生成的短时长、小变化的瓶颈；5）采集高质量的大规模视频数据，结合并扩展视频自监督与文生图预训练模型，训练大规模的文生视频生成训练模型。

毋庸置疑，在扩散模型飞速演进之下，人工智能正式地从感知走向了创造。无论是在图像还是在视频上，AI 已经能够生成在感知上真实的、和谐的、高清的视觉数据，甚至可以基于人类输入的某些视觉或者语言概念进行定制化的创造、编辑、控制。在此，针对生成式模型的发展与进步，本文也对 AI 的未来形态做出进一步的展望：在感知与认知的能力都具备时，AI 模型便能够创造出属于自己的一个开放式世界，并允许人们在这个开放式世界中进行自由发挥，实现“所想即所得”，不再受到现实生活中许多条条框框的条件式约束。在这个开放式环境中，AI 模型的训练也不再受限于数据的收集，因此改变了许多现有的机器学习范式。诸如迁移式的判别学习（领域自适应）、主动学习等可能不再重要，AI 或许能够在自己所创造的开放世界中实现自我交互、自我学习与自我进步，从而达到更高层次的智能，真正意义上地改变人类的生活方式。

参 考 文 献

[1] Müller V C, Bostrom N. Future progress in artificial intelligence: a survey of expert opinion[M]//Müller V C. Fundamental issues of artificial intelligence. Synthese library. Cham: Springer, 2016, 376: 555-572.

[2] Došilović F K, Brčić M, Hlupić N. Explainable artificial intelligence: a survey[C]//2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 21-25, 2018, Opatija, Croatia. New York: IEEE Press, 2018: 210-215.

[3] Lu Y. Artificial intelligence: a survey on evolution, models, applications and future trends[J]. Journal of Management Analytics, 2019, 6(1): 1-29.

[4] Henry W P. Artificial intelligence[M]. Massachusetts: Addison-Wesley Longman Publishing Co., Inc., 1984.

[5] Huang C X, Wang G R, Zhou Z B, et al. Reward-adaptive reinforcement learning: dynamic policy gradient optimization for bipedal locomotion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7686-7695.

[6] Huang C X, Zhang R H, Ouyang M Z, et al. Deductive reinforcement learning for visual autonomous urban driving navigation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5379-5391.

[7] Wu J, Li G B, Han X G, et al. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos[C]//Proceedings of the 28th ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, USA. New York: ACM Press, 2020: 1283-1291.

[8] Xie S R, Huang J N, Lei L X, et al. NADPEx: an on-policy temporally consistent exploration method for deep reinforcement learning[EB/OL]. (2018-12-21)[2023-03-02]. https://arxiv.org/abs/1812.09028.

[9] Garland M, le Grand S, Nickolls J, et al. Parallel computing experiences with CUDA[J]. IEEE Micro, 2008, 28(4): 13-27.

[10] Kalaiselvi T, Sriramakrishnan P, Somasundaram K. Survey of using GPU CUDA programming model in medical image analysis[J]. Informatics in Medicine Unlocked, 2017, 9: 133-144.

[11] Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library[EB/OL]. (2019-12-03)[2023-03-02]. https://arxiv.org/abs/1912.01703.

[12] Abadi M. TensorFlow: learning functions at scale[C]//Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, September 18 - 24, 2016, Nara, Japan. New York: ACM Press, 2016.

[13] Wang G R, Lin L, Chen R C, et al. Joint learning of neural transfer and architecture adaptation for image recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(10): 5401-5415.

[14] Chen T S, Lin L, Chen R Q, et al. Knowledge-guided multi-label few-shot learning for general image recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1371-1384.

[15] Wang K Z, Zhang D Y, Li Y, et al. Cost-effective active learning for deep image classification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(12): 2591-2600.

[16] Wang X L, Lin L, Huang L C, et al. Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 3334-3341.

[17] Wang K Z, Lin L, Zuo W M, et al. Dictionary pair classifier driven convolutional neural networks for object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2138-2146.

[18] Wang K Z, Yan X P, Zhang D Y, et al. Towards human-machine cooperation: self-supervised sample mining for object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1605-1613.

[19] Jiang C H, Xu H, Liang X D, et al. Hybrid knowledge routed modules for large-scale object detection[EB/OL]. (2018-10-30)[2023-03-02]. https://arxiv.org/abs/1810.12681.

[20] Xu H, Jiang C H, Liang X D, et al. Reasoning-RCNN: unifying adaptive global reasoning into large-scale object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 6412-6421.

[21] Yang B B, Deng X C, Shi H, et al. Continual object detection via prototypical task correlation guided gating mechanism[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA,

USA. New York: IEEE Press, 2022: 9245-9254.
[22] Wu Y X, Zhang G W, Xu H, et al. Auto-panoptic: cooperative multi-component architecture search for panoptic segmentation [EB/OL]. (2020-10-30) [2023-02-03]. https://arxiv. org/abs/2010.16119.
[23] Wu Y X, Zhang G W, Gao Y M, et al. Bidirectional graph reasoning network for panoptic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 9077-9086.
[24] Yang J H, Xu R J, Li R Y, et al. An adversarial perturbation oriented domain adaptation approach for semantic segmentation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12613-12620.
[25] Jordan M I, Mitchell T M. Machine learning: trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.
[26] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
[27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2023-03-06]. https://arxiv.org/abs/1409.1556.
[28] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 1440-1448.
[29] Ren S Q, He K M, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39: 1137-1149.
[30] Strumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory[J]. Journal of Machine Learning Research, 2010, 11: 1-18.
[31] Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions[J]. Journal of Machine Learning Research, 2010, 11: 1803-1831.
[32] Oussidi A, Elhassouny A. Deep generative models: survey[C]//2018 International Conference on Intelligent Systems and Computer Vision (ISCV), April 2-4, 2018, Fez, Morocco. New York: IEEE Press, 2018.
[33] Pan Z Q, Yu W J, Yi X K, et al. Recent progress on generative adversarial networks (GANs): a survey[J]. IEEE Access, 2019, 7: 36322-36333.
[34] Harshvardhan G M, Gourisaria M K, Pandey M, et al. A comprehensive survey and analysis of generative models in machine learning[J]. Computer Science Review, 2020, 38: 100285.
[35] Cao H Q, Tan C, Gao Z Y, et al. A survey on generative diffusion model[EB/OL]. (2022-09-06) [2023-03-02]. https://arxiv.org/abs/2209.02646.
[36] Wiatrak M, Albrecht S V, Nystrom A. Stabilizing generative adversarial networks: a survey[EB/OL]. (2019-09-30)[2023-03-02]. https://arxiv.org/abs/1910.00927.
[37] Wang K F, Gou C, Duan Y J, et al. Generative adversarial networks: introduction and outlook[J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(4): 588-598.
[38] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 15979-15988.
[39] Wang G R, Tang Y S, Lin L, et al. Semantic-aware auto-encoders for self-supervised representation learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 9654-9665.
[40] Guo Z Y, Zhang R R, Qiu L T, et al. Joint-MAE: 2D-3D joint masked autoencoders for 3D point cloud pre-training[EB/OL]. (2023-02-27)[2023-03-02]. https://arxiv.org/abs/2302.14007.
[41] Tan Q Y, Liu N H, Huang X, et al. S2GAE: self-supervised graph autoencoders are generalizable learners with graph masking [C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, February 27, 2023, Singapore, Singapore. New York: ACM Press, 2023: 787-795.
[42] Bao H B, Dong L, Piao S H, et al. BEiT: BERT pre-training of image transformers[EB/OL]. (2021-06-15) [2023-03-02]. https://arxiv.org/abs/2106.08254.
[43] Xia L H, Huang C, Huang C Z, et al. Automated self-supervised learning for recommendation[EB/OL]. (2023-03-14) [2023-04-02]. https://arxiv.org/abs/2303.07797.
[44] Chen A, Zhang K, Zhang R R, et al. PiMAE: point cloud and image interactive masked autoencoders for 3D object detection [EB/OL]. (2023-03-14) [2023-04-05]. https://arxiv. org/abs/2303.08129.
[45] Ren S C, Wei F Y, Albanie S, et al. DeepMIM: deep supervision for masked image modeling[EB/OL]. (2023-03-15) [2023-04-05]. https://arxiv.org/abs/2303.08817.
[46] Wei C, Fan H Q, Xie S N, et al. Masked feature prediction for self-supervised visual pre-training[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 14648-14658.
[47] He K M, Fan H Q, Wu Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 9726-9735.
[48] Chen X L, Fan H Q, Girshick R, et al. Improved baselines with momentum contrastive learning[EB/OL]. (2020-03-09)[2023-06-09]. https://arxiv.org/abs/2003.04297.
[49] Chen X L, Xie S N, He K M. An empirical study of training self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9620-9629.
[50] Chen X L, He K M. Exploring simple Siamese representation learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 15745-15753.
[51] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning[EB/OL]. (2020-06-13) [2023-03-05]. https://arxiv.org/abs/2006.07733.
[52] Cheng Z Z, Yang Q X, Sheng B. Deep colorization[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 415-423.
[53] Xiao Y, Zhou P Y, Zheng Y, et al. Interactive deep colorization using simultaneous global and local inputs[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 12-17, 2019, Brighton, UK. New York: IEEE Press, 2019: 1887-1891.
[54] Richard Z, Phillip I, Efros A A. Colorful image colorization [M]//Leibe B, Matas J, Sebe N, et al. Computer vision – ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9907: 649-666.
[55] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization[M]//Leibe B, Matas J, Sebe N, et al. Computer vision – ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 577-593.
[56] Zhang R, Zhu J Y, Isola P, et al. Real-time user-guided image colorization with learned deep priors[EB/OL]. (2017-05-08)

[2023-02-05]. https://arxiv.org/abs/1705.02999.

[57] He M M, Chen D D, Liao J, et al. Deep exemplar-based colorization[J]. ACM Transactions on Graphics, 2018, 37(4): 1-16.

[58] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 105-114.

[59] Sharif S M A, Ali Naqvi R, Ali F, et al. DarkDeblur: learning single-shot image deblurring in low-light condition[J]. Expert Systems With Applications, 2023, 222: 119739.

[60] Li B C, Li X, Lu Y T, et al. Hst: hierarchical swin transformer for compressed image super-resolution[M]//Karlinsky L, Michaeli T, Nishino K. Computer vision - ECCV 2022 workshops. Lecture notes in computer science. Cham: Springer, 2022, 13802.

[61] Liang J Y, Cao J Z, Sun G L, et al. SwinIR: image restoration using swin transformer[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11-17, 2021, Montreal, BC, Canada. New York: IEEE Press, 2021: 1833-1844.

[62] Zamir S W, Arora A, Khan S, et al. Multi-stage progressive image restoration[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14816-14826.

[63] Yang F Z, Yang H, Fu J L, et al. Learning texture transformer network for image super-resolution[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 5790-5799.

[64] Dai T, Cai J R, Zhang Y B, et al. Second-order attention network for single image super-resolution[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 11057-11066.

[65] Liu X, Wu Q Y, Zhou H, et al. Audio-driven co-speech gesture video generation[EB/OL]. (2022-12-05) [2023-02-03]. https://arxiv.org/abs/2212.02350.

[66] Cui R P, Cao Z, Pan W S, et al. Deep gesture video generation with learning on regions of interest[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2551-2563.

[67] Saunders B, Camgoz N C, Bowden R. Anonysign: novel human appearance synthesis for sign language video anonymisation[C]// 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), December 15-18, 2021, Jodhpur, India. New York: IEEE Press, 2022.

[68] Natarajan B, Elakkiya R. Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks[J]. Soft Computing, 2022, 26(23): 13153-13175.

[69] Ferstl Y, Neff M, McDonnell R. Multi-objective adversarial gesture generation[C]//Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games, October 28-30, 2019, Newcastle upon Tyne, United Kingdom. New York: ACM Press, 2019.

[70] Zeng D, Liu H, Lin H, et al. Talking face generation with expression-tailored generative adversarial network[C]// Proceedings of the 28th ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, USA. New York: ACM Press, 2020: 1716-1724.

[71] Zhou H, Liu Y, Liu Z W, et al. Talking face generation by adversarially disentangled audio-visual representation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 9299-9306.

[72] Zhang B W, Qi C Y, Zhang P, et al. MetaPortrait: identity-preserving talking head generation with fast personalized adaptation[EB/OL]. (2022-12-15) [2023-03-05]. https://arxiv.org/abs/2212.08062.

[73] Zhou H, Sun Y S, Wu W, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 4174-4184.

[74] Zeng D, Zhao S T, Zhang J J, et al. Expression-tailored talking face generation with adaptive cross-modal weighting[J]. Neurocomputing, 2022, 511: 117-130.

[75] Chen L L, Maddox R K, Duan Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 7824-7833.

[76] Zhang Z M, Li L C, Ding Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 3660-3669.

[77] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: representing scenes as neural radiance fields for view synthesis [M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision - ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 405-421.

[78] Park K, Sinha U, Barron J T, et al. Nerfies: deformable neural radiance fields[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 5845-5854.

[79] Niemeyer M, Geiger A. GIRAFFE: representing scenes as compositional generative neural feature fields[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 11448-11459.

[80] Pumarola A, Corona E, Pons-Moll G, et al. D-NeRF: neural radiance fields for dynamic scenes[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 10313-10322.

[81] Martin-Brualla R, Radwan N, Sajjadi M S M, et al. NeRF in the wild: neural radiance fields for unconstrained photo collections[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 7206-7215.

[82] Chan E R, Monteiro M, Kellnhofer P, et al. Pi-GAN: periodic implicit generative adversarial networks for 3D-aware image synthesis[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 5795-5805.

[83] Chan E R, Lin C Z, Chan M A, et al. Efficient geometry-aware 3D generative adversarial networks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 16102-16112.

[84] Li Z Q, Niklaus S, Snavely N, et al. Neural scene flow fields for space-time view synthesis of dynamic scenes[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 6494-6504.

[85] Oechsle M, Peng S Y, Geiger A. UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 5569-5579.

[86] Weng L. What are diffusion models [EB/OL]. [2023-03-06]. https://lilianweng. github. io/posts/2021-07-11-diffusion-models/.
[87] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[EB/OL]. (2014-06-10) [2023-03-02]. https://arxiv.org/abs/1406.2661.
[88] Chen Z L, Yang Z F, Wang X X, et al. Multivariate-information adversarial ensemble for scalable joint distribution matching[EB/OL]. (2019-07-08) [2023-03-02]. https://arxiv.org/abs/1907.03426.
[89] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks[EB/OL]. (2017-01-17)[2023-03-02]. https://arxiv.org/abs/1701.04862.
[90] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[EB/OL]. (2017-01-26) [2023-03-02]. https://arxiv. org/abs/1701.07875.
[91] Kingma D P, Welling M. Auto-encoding variational Bayes[EB/OL]. (2013-12-20) [2023-03-02]. https://arxiv. org/abs/1312.6114.
[92] Rezende D, Mohamed S. Variational inference with normalizing flows[EB/OL]. (2015-05-21) [2023-03-02]. https://arxiv. org/abs/1505.05770.
[93] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP[EB/OL]. (2016-05-21) [2023-03-02]. https://arxiv.org/abs/1605.08803.
[94] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [EB/OL]. (2020-06-19) [2023-03-02]. https://arxiv. org/abs/2006.11239.
[95] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[EB/OL]. (2021-05-11) [2023-03-02]. https://arxiv.org/abs/2105.05233.
[96] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022-04-13)[2023-02-03]. https://arxiv.org/abs/2204.06125.
[97] 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 (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 10674-10685.
[98] Nichol A Q, Dhariwal P, Ramesh A, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2021-12-20) [2023-03-02]. https://arxiv.org/abs/2112.10741.
[99] Saharia C, William C, Saurabh S, et al. Photorealistic text-to-image diffusion models with deep language understanding[EB/OL]. (2022-05-23) [2023-03-02]. https://arxiv. org/abs/2205.11487.
[100] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2023-03-02]. https://arxiv.org/abs/2103.00020.
[101] van den Oord A, Vinyals O. Neural discrete representation learning[EB/OL]. (2017-11-02)[2023-03-02]. https://arxiv.org/abs/1711.00937.
[102] Ruiz N, Li Y Z, Jampani V, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation[EB/OL]. (2022-08-25) [2023-03-05]. https://arxiv. org/abs/2208.12242.
[103] Gal R, Alaluf Y, Atzmon Y, et al. An image is worth one word: personalizing text-to-image generation using textual inversion[EB/OL]. (2022-08-02) [2023-03-05]. https://arxiv.org/abs/2208.01618.
[104] Kumari N, Zhang B L, Zhang R, et al. Multi-concept customization of text-to-image diffusion[EB/OL]. (2022-12-08) [2023-03-05]. https://arxiv.org/abs/2212.04488.
[105] Avrahami O, Lischinski D, Fried O. Blended diffusion for text-driven editing of natural images[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 18187-18197.
[106] Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with cross attention control[EB/OL]. (2022-08-02) [2023-03-05]. https://arxiv.org/abs/2208.01626.
[107] Kawar B, Zada S, Lang O, et al. Imagic: text-based real image editing with diffusion models[EB/OL]. (2022-10-17) [2023-03-02]. https://arxiv.org/abs/2210.09276.
[108] Ho J, Chan W, Saharia C, et al. Imagen video: high definition video generation with diffusion models[EB/OL]. (2022-10-05) [2023-02-03]. https://arxiv.org/abs/2210.02303.
[109] Ho J, Salimans T, Gritsenko A A, et al. Video diffusion models [EB/OL]. (2022-04-07) [2023-03-02]. https://arxiv. org/abs/2204.03458.
[110] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention – MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.
[111] Mei K F, Patel V M. VIDM: video implicit diffusion models [EB/OL]. (2022-12-01) [2023-02-03]. https://arxiv. org/abs/2212.00235.
[112] Yu S, Sohn K, Kim S, et al. Video probabilistic diffusion models in projected latent space[EB/OL]. (2022-04-07)[2023-03-02]. https://arxiv.org/abs/2302.07685.
[113] Singer U, Polyak A, Hayes T, et al. Make-a-video: text-to-video generation without text-video data[EB/OL]. (2022-09-29) [2023-03-05]. https://arxiv.org/abs/2209.14792.
[114] Luo Z X, Chen D Y, Zhang Y Y, et al. VideoFusion: decomposed diffusion models for high-quality video generation [EB/OL]. (2023-03-15) [2023-05-06]. https://arxiv. org/abs/2303.08320.
[115] Esser P, Chiu J, Atighehchian P, et al. Structure and content-guided video synthesis with diffusion models[EB/OL]. (2023-02-06)[2023-03-05]. https://arxiv.org/abs/2302.03011.
[116] 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 [EB/OL]. (2022-12-22) [2023-03-05]. https://arxiv. org/abs/2212.11565.
[117] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision – ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.
[118] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [EB/OL]. (2017-06-26) [2023-03-02]. https://arxiv. org/abs/1706.08500.
[119] Zhou Y F, Zhang R Y, Chen C Y, et al. LAFITE: towards language-free training for text-to-image generation[EB/OL]. (2021-11-27)[2023-03-05]. https://arxiv.org/abs/2111.13792.
[120] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[EB/OL]. (2021-02-24) [2023-03-02]. https://arxiv.org/abs/2102.12092.
[121] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. (2012-12-03)[2023-03-06]. https://arxiv.org/abs/1212.0402.
[122] Xiong W, Luo W H, Ma L, et al. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2364-2373.
[123] Unterthiner T, van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: a new metric & challenges [EB/OL]. (2018-12-03) [2023-03-05]. https://arxiv. org/abs/1812.01717.
[124] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[EB/OL]. (2016-06-10) [2023-03-

02]. https://arxiv.org/abs/1606.03498.
[125] Tulyakov S, Liu M Y, Yang X D, et al. MoCoGAN: decomposing motion and content for video generation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1526-1535.
[126] Yan W, Zhang Y Z, Abbeel P, et al. VideoGPT: video generation using VQ-VAE and transformers[EB/OL]. (2021-04-20)[2023-03-05]. https://arxiv.org/abs/2104.10157.
[127] Tian Y, Ren J, Chai M L, et al. A good image generator is what you need for high-resolution video synthesis[EB/OL]. (2021-04-30)[2023-03-05]. https://arxiv.org/abs/2104.15069.
[128] Yu S, Tack J, Mo S, et al. Generating videos with dynamics-aware implicit generative adversarial networks[EB/OL]. (2022-02-21)[2023-03-05]. https://arxiv.org/abs/2202.10571.

From Perception to Creation: Exploring Frontier of Image and Video Generation Methods

Lin Liang, Yang Binbin*

School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, *Guangdong, China*

Abstract

Significance In recent years, advancements in computing software and hardware have led to artificial intelligent (AI) models achieving performance levels approaching or surpassing human capabilities in perceptive tasks. However, in order to develop mature AI systems that can comprehensively understand the world, models must be capable of generating visual concepts, rather than simply recognizing them because creation and customization require a thorough understanding of high-level semantics and full details of each generated object.

From an applied perspective, when AI models obtain the capability of visual understanding and generation, they will significantly promote progress and development across diverse aspects of the industry. For example, visual generative models can be applied to the following aspects: colorizing and restoring old black and white photos and films; enhancing and remastering old videos in high definition; synthesizing real-time virtual anchors, talking faces, and AI avatars; incorporating special effects into personalized video shooting on short video platforms; stylizing users′ portraits and input images; compositing movie special effects and scene rendering, and so on. Therefore, research on the theories and methods of image and video generation models holds significant theoretical significance and industrial application value.

Progress In this paper, we first provide a comprehensive overview of existing generative frameworks, including generative adversarial networks (GAN), variational autoencoders (VAE), flow models, and diffusion models, which can be summarized in Fig. 5. GAN is trained in an adversarial manner to obtain an ideal generator, with the mutual competition of a generator and a discriminator. VAE is composed of an encoder and a decoder, and it is trained via variational inference to make the decoded distribution approximate the real distribution. The flow model uses a family of invertible mappings and simple priors to construct an invertible transformation between real data distribution and the prior distribution. Different from GANs and VAEs, flow models are trained by the estimation of maximum likelihood. Recently, diffusion models emerge as a class of powerful visual generative models with state-of-the-art synthesis results on visual data. The diffusion model decomposes the image generation process into a sequence of denoising processes from a Gaussian prior. Its training procedure is more stable by avoiding the use of an adversarial training strategy and can be successfully deployed in a large-scale pre-trained generation system.

We then review recent state-of-the-art advances in image and video generation and discuss their merits and limitations. Fig. 6 shows the overview of image and video generation models and their classifications. Works on pre-trained text-to-image generation models study how to pre-train a text-to-image foundation model on large-scale datasets. Among those T2I foundation models, stable diffusion becomes a widely-used backbone for the tasks of image/video customization and editing, due to its impressive performance and scalability. Prompt-based image editing methods aim to use the pre-trained text-to-image foundation model, e. g. , stable diffusion, to edit a generated/natural image according to input text prompts. Due to the difficulty of collecting large-scale and high-quality video datasets and the expensive computational cost, the research on video generation still lags behind image generation. To learn from the success of text-to-image diffusion models, some works, e. g. , video diffusion model, imagen video, VIDM, and PVDM, have tried to use enormous video data to train a video diffusion model from scratch and obtain a video generation foundation model similar to stable diffusion. Another line of work aims to resort to pre-trained image generators, e. g. , stable diffusion, to provide content prior to video generation and only learn the temporal dynamics from video, which significantly improves

the training efficiency.

Finally, we discuss the drawbacks of existing image and video generative modeling methods, such as misalignment between input prompts and generated images/videos, further propose feasible strategies to improve those visual generative models, and outline potential and promising future research directions. These contributions are crucial for advancing the field of visual generative modeling and realizing the full potential of AI systems in generating visual concepts.

Conclusions and Prospects Under the rapid evolution of diffusion models, artificial intelligence has undergone a significant transformation from perception to creation. AI can now generate perceptually realistic and harmonious data, even allowing visual customization and editing based on input conditions. In light of this progress in generative models, here we provide prospects for the potential future forms of AI: with both perception and cognitive abilities, AI models can establish their own open world, enabling people to realize the concept of "what they think is what they get" without being constrained by real-life conditions. For example, in this open environment, the training of AI models is no longer restricted by data collection, leading to a reformation of many existing paradigms in machine learning. Techniques like transfer learning (domain adaptation) and active learning may diminish in importance. AI might be able to achieve self-interaction, self-learning, and self-improvement within the open world it creates, ultimately attaining higher levels of intelligence and profoundly transforming humans′ lifestyles.

Key words artificial intelligent model; visual generative modeling; diffusion model; image and video generation