计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2024/18(06)-1404-17
doi: 10.3778/j.issn.1673-9418.2401075

AI绘画研究综述

张泽宇[1],王铁君[2+],郭晓然[2],龙智磊[1],徐　魁[1]
1. 西北民族大学 中国民族语言文字信息技术教育部重点实验室,兰州 730030
2. 西北民族大学 数学与计算机科学学院,兰州 730030
+ 通信作者 E-mail: wtj@mail.lzjut.cn

摘　要:AI绘画,作为计算机视觉领域的热门研究方向,正通过自然语言处理技术、图文预训练大模型,以及新兴的扩散模型,不断拓展其在艺术创作、影视媒体、工业设计、艺术教育等领域的应用边界。将以图生图和以文生图两类AI绘画任务作为主线,深入分析了代表性模型及其关键技术和方法。对于以图生图方式,从基于自编码器和基于生成式对抗网络两类模型分别探讨了各自的发展脉络、生成原理以及优缺点,并总结了它们在公共数据集上的效果;对于以文生图方式,归纳了基于扩散模型等三类模型的结构区别,以及在三个数据集上各类模型的生成效果,同时指出利用扩散模型的以文生图方式已成为当下的热点,并预示着未来图像生成方式的多样化发展。对目前主流的AI绘画平台从使用方式、生成速度等角度进行了对比总结。最后在总结AI绘画在技术层面和社会层面所面临的问题与争议的基础上,展望了AI绘画与人类艺术家的互补发展、绘画过程互动性增强以及新职业和产业的出现等未来趋势。

关键词:AI绘画;以图生图;以文生图;图像生成;人工智能生成内容(AIGC)
文献标志码:A　**中图分类号:**TP301

Survey of AI Painting

ZHANG Zeyu[1], WANG Tiejun[2+], GUO Xiaoran[2], LONG Zhilei[1], XU Kui[1]
1. Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730030, China
2. School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, China

Abstract: AI painting, as a popular research direction in the field of computer vision, is expanding its application boundaries in the fields of art creation, film and media, industrial design, and art education through natural language processing, graphic pre-training models, and diffusion models. Two types of AI painting, namely, image-to-image and text-to-image, are taken as the main lines, and the representative models and their key technologies and methods are analyzed in depth. For the image-to-image, the development lineage, generation principle, and advantages and disadvantages of each model are explored from two types of models based on AE and GAN, and their effects on the public dataset are summarized. For the text-to-image, the structural differences of the three types of models based on diffusion model and other models, as well as the generation effects of various types of models on three datasets are summarized. It is pointed out that the text-to-image utilizing the diffusion model has become a hot topic nowadays, which predicts the diversified development of image generation in the future. And the current mainstream AI painting

基金项目:国家自然科学基金(62166035);甘肃省自然科学基金(21JR7RA163);中央高校项目-重大需求培育项目(31920230175)。
This work was supported by the National Natural Science Foundation of China (62166035), the Natural Science Foundation of Gansu Province (21JR7RA163), and the Central Universities Project - Major Needs Cultivation Project (31920230175).
收稿日期:2024-01-29　**修回日期:**2024-03-26

platforms are compared and summarized from the perspectives of usage and generation speed. Finally, on the basis of summarizing the problems and controversies faced by AI painting at the technical and social levels, future trends such as the complementary development of AI painting and human artists, the increased interactivity of the painting process, and the emergence of new professions and industries are envisioned.

Key words: AI painting; image-to-image; text-to-image; image generation; artificial intelligence generated content (AIGC)

绘画是人类最古老的艺术形式之一，随着人工智能的发展，出现了新的绘画方式——AI绘画。它是一种利用人工智能算法来生成或转换图片的过程，可以根据文字描述、风格示例或其他条件来创造出各种风格和主题的图片，或者对已有的图片进行风格转换、上色、修复等操作[1]。相比依赖于艺术家的情感、手法和技巧所绘制的传统绘画，AI绘画依赖于算法和深度学习技术，其创作过程更加理性，作品风格呈现出统一性、规范性和多样性等特点，作品价值更多体现在技术创新、数据质量、算法水平和算力消耗等方面。目前来看，AI绘画作为人工智能生成内容（artificial intelligence generated content，AIGC）的应用场景之一，具有广阔的发展前景和应用价值[2-3]。

AI绘画的历史可以追溯到20世纪70年代，当时主要采用传统的基于计算机软件或计算机控制程序与机械装置结合的方式进行绘画。例如Harold Cohen教授在1972年开发设计的一款名为“AARON”的计算机绘图程序[4]，这是最早的也是最为复杂的艺术作品生成程序之一，该程序通过控制一个机械臂在固定好的纸上进行绘画。2006年，由Colton开发的绘画软件The Painting Fool可以提取数码照内区域块的颜色，然后模拟现实世界中的绘画材料进行绘图[5-6]。

随着深度学习[7]、硬件算力和大规模数据集的发展，出现了各类生成式模型，并被应用于AI绘画任务中。其中最具代表性的是Goodfellow等[8]在2014年提出的生成对抗网络（generative adversarial network，GAN），通过使用一个生成器与一个判别器组合的方式来生成图像。一种基于马尔可夫链[9]的扩散模型[10]也逐渐被用于AI绘画任务，该类模型能够通过逐步去除和恢复图像的噪声来生成新图像，具有训练稳定、生成质量高、易于控制等优点。

目前，研究者提出了各种AI绘画模型，根据输入数据类型的不同可分为两大类：（1）以图生图模型（image-to-image）。通过输入已有的图像自动生成新图像。（2）以文生图模型（text-to-image）。依靠输入文本的描述生成具有该文本特征的新图像。目前还出现了以“文+图”生图的AI绘画方式。

本文主要对图像生成模型进行了分类总结，详细介绍并讨论了各代表模型的原理、结构、特点与应用优缺点，并对AI绘画面临的问题与挑战进行了全面总结与深入讨论。本文的工作有助于引导更多人了解AI绘画的技术原理与挑战，为之后开展更有价值的研究工作提供启发。

1 以图生图模型

以基于自编码器和基于生成式对抗网络这两类模型为主线，归纳和整理了代表模型的主要原理和特点，并进行对比分析。图1展示了以图生图模型的发展脉络。

1.1 基于自编码器的以图生图模型

隐空间是一种用于表示压缩数据的空间，它可以捕捉数据的重要特征和模式，同时去除冗余和噪声。其可以分为连续隐空间和离散隐空间。在以图生图模型中，隐空间可以反映图像的内容和风格。基于自编码器的以图生图模型是以隐空间表示为核心，通过重建误差来度量生成图像与输入图像的相似度。

1.1.1 基于连续隐空间模型

1986年，Rumelhart等[11]最早提出自编码器（auto-encoder，AE）的概念。Bourlard等[12]将AE描述为由编码器和解码器组成的一种基于无监督学习的生成模型，结构如图2所示。该模型具有重建过程简单、自动提取特征、可堆叠多层等优点，但是这种使用编码器和解码器的方式耗时长，且生成图清晰度不够。图2中，$\boldsymbol{X}$、$\boldsymbol{Z}$与$\boldsymbol{X}'$分别表示输入图像、编码器的输出，即图像的隐层特征、重建图像。编码器和解码器可以是多种形式，如全连接层、卷积层、循环层等。

Kingma等[13]提出了变分自编码器（variational auto-encoders，VAE），该编码器是基于变分贝叶斯推断的生成式网络结构，如图3所示。与AE不同的是，VAE不再去学习一个连续的表征，而是直接学习一个分布，对输入图像$\boldsymbol{X}$，通过采样得到服从高斯分布的隐变量$\boldsymbol{Z}$，最后解码$\boldsymbol{Z}$得到生成图像$\boldsymbol{X}'$。在图3中，μ与σ为$\boldsymbol{Z}$的均值与标准差，$P(\boldsymbol{X}|\boldsymbol{Z})$为真实图像的先

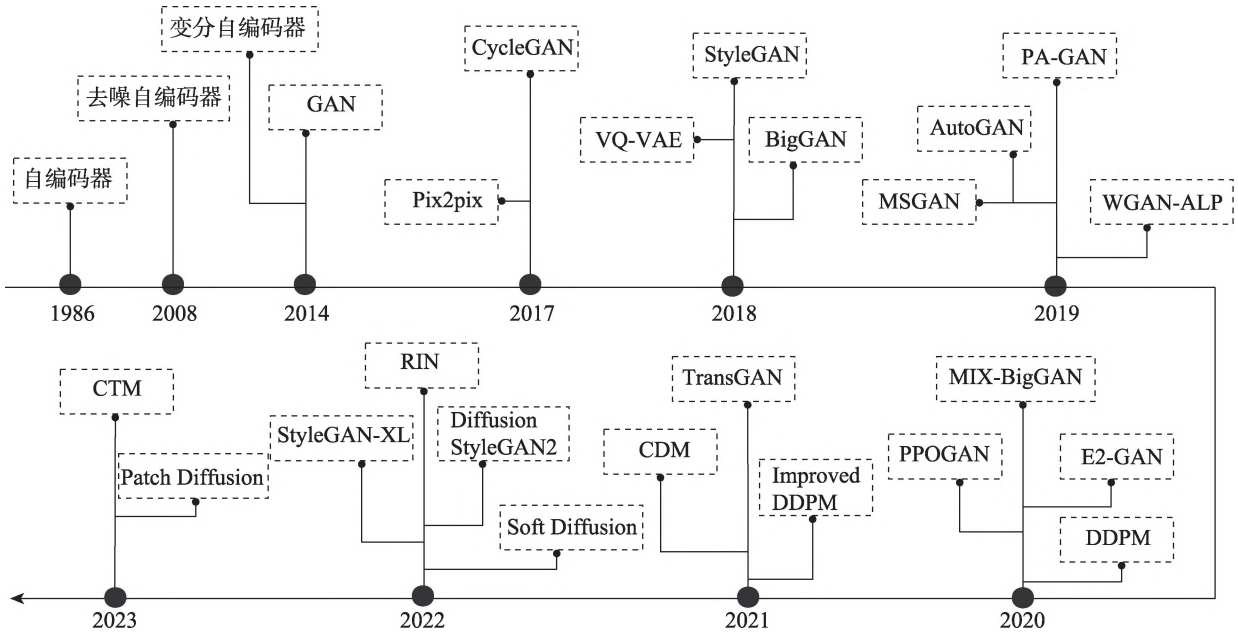


图1 以图生图模型发展脉络

Fig.1 Development line of image-to-image model

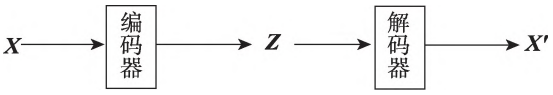


图2 AE结构图

Fig.2 AE structure diagram

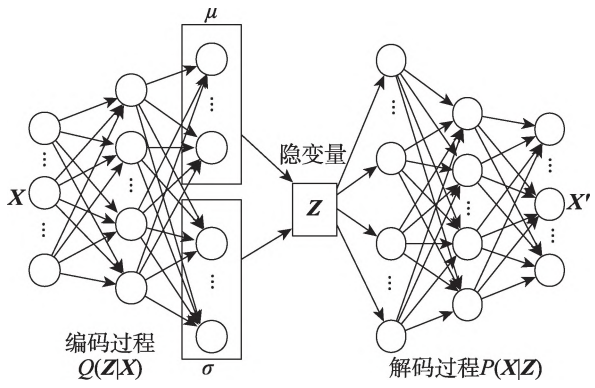


图3 VAE结构图

Fig.3 VAE structure diagram

验分布，$Q(Z|X)$ 为 Z 的近似后验分布，通过调节 μ 与 σ 来生成不同的输出数据，从而实现以图生图的功能。虽然VAE可以通过反向传播算法[14]进行快速训练，并且在图像生成方面有较好的效果，但是因为过强的先验假设限制了模型的拟合能力，导致其不能模拟复杂分布。

1.1.2 基于离散隐空间模型

相比于基于连续隐空间的模型，基于离散隐空间的模型在训练、生成和评估方面均有一些不同的难点和挑战。首先离散变量的梯度不可导，需要使用重参数化的技巧来进行优化，其次离散变量的变化会导致图像的剧烈变化，而不是连续变量的渐进变化，最后由于离散变量的分布和度量方式与连续变量不同，评估和比较离散隐空间模型的性能和效果也比连续隐空间模型更加困难。

Vincent等[15]在AE的基础上引入退化过程，提出去噪自编码器（denoising autoencoders，DAE），其设计的目的是通过改善噪声影响来提高所提取特征的稳健性，结构如图4所示。引入的退化过程使得由一系列离散的符号或整数构成的隐变量 Z 具有更好的可解释性和可操作性，模型提取的抽象特征也更能反映数据本质，但是增加了训练时间，并且退化过程中的退化率大小对模型的性能有一定影响。

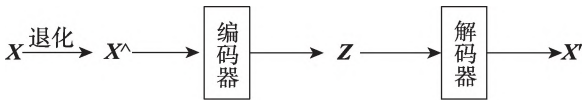


图4 DAE结构图

Fig.4 DAE structure diagram

Oord等[16]提出向量量化变分自编码器（vector quantized variational autoencoder，VQ-VAE），其结构如图5所示。Codebook是一个由 K 个 D 维向量组成的矩阵，其中存放的每个向量都是一个离散的隐变量，用于表示图像的某种特征或属性。输入图像通过编码

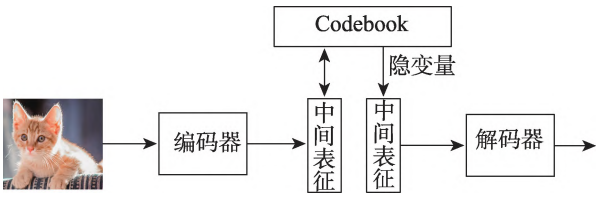


图5 VQ-VAE结构图

Fig.5 VQ-VAE structure diagram

器转换为一个低维的特征图，即中间表征。编码器的输出维度与Codebook的大小相同，以便于进行向量量化的操作。之后利用K近邻算法[17]在其中查询与中间表征相似的向量，解码器则经过向量量化的特征图还原为一个高维的图像，即重建的图像。其在编码阶段需要单独训练一个基于自回归的模型如PixelCNN[18]来学习先验。该模型通过高度一致性的Codebook查询操作代替混乱的中间表征，有效地提高了生成图像的可控性与丰富度，但是其离散的隐变量空间可能存在信息损失或者编码冗余的问题。

通过对上述模型的分析可知，基于自编码器结构的图像生成模型，其优点是不需要标注数据，只需要输入图像即可，而且可以提高重构质量、隐含表示的鲁棒性或多样性。缺点是它们不能保证隐含表示的连续性和可解释性。表1对上文所介绍的自编码器进行了总结。

1.2 基于生成式对抗网络的以图生图模型

虽然自编码器是以图生图的一类重要模型，但是它们缺乏对图像内容、风格的分离与控制，并不能很好地处理图像风格迁移这样的任务。为了解决这类问题，引入了生成式对抗网络。目前图像风格迁移算法主要有两类：基于图像迭代与基于模型迭代。基于图像迭代的算法是直接在白噪声图像上通过不断优化白噪声图像，最终实现风格迁移；基于模型迭代则是通过不断优化神经网络模型来实现风格迁移。陈淑環等[19]对基于图像迭代与部分基于模型迭代的图像风格迁移方法进行了总结；陈淮源等[20]对比分析了基于卷积神经网络和基于生成对抗网络的风格迁移方法。本文在上述文献基础上，进一步对两类算法的基本方法、代表性工作与优缺点进行了归纳总结，如表2所示。

相比于基于图像迭代方法，以GAN为代表的生成式对抗网络及其改进版本具有更加灵活和高效的特点。本节对通用性好、迁移效果逼真、影响范围广的生成式对抗网络模型进行介绍和比较。

表1 部分自编码器总结

Table 1 Summary of partial self-encoders

名称	在VE上的改进点	目的	优点	缺点
DAE[15]	将带有噪声的信息作为输入	提高重构数据的稳定性，增强其对噪声的鲁棒性	能够抑制过拟合，提取出数据中的本质特征	需要合理地选择噪声类型和程度，否则可能影响重构效果
VAE[13]	在AE的隐层表达上增加了一个对隐变量的分布约束，通过对分布采样重构数据	使编码器产生的隐层表达满足正态分布，提高模型的表达能力	能够生成连续和平滑的潜在空间，便于插值和生成新样本	由于使用了重参数化技巧，可能导致梯度消失和模糊重构
VQ-VAE[16]	在VAE的基础上将连续的高维向量离散化	解决VAE生成图模糊问题，减少模型复杂度，提高重构数据质量	能够有效地学习离散潜在表示，避免了后验坍塌问题	需要预先设定离散码本的大小和维度，可能影响模型的灵活性和泛化性

表2 两类风格迁移算法对比

Table 2 Comparison of two types of style transfer algorithms

类型	基本方法	代表性工作	优点	缺点
基于图像迭代	最大均值差异	Gatys等[21]提出一种基于卷积神经网络利用预训练的VGG[22]网络模型作为图像特征提取器来完成风格迁移任务的方法	可以更多地保留输入图像的细节和纹理，可控性好，无需训练数据	计算时间长，对预训练模型依赖性大，而不同的模型可能导致不同的结果
	马尔可夫随机场	Li等[23]将马尔可夫随机场与深度卷积神经网络结合用于图像风格迁移任务		
	深度图像类比	Liao等[24]通过区域块匹配迭代优化的深度图像类比方法提高图像风格迁移的有效性		
基于模型迭代	生成模型	Johnson等[25]提出快速风格迁移模型；Zhu等[26]提出CycleGAN模型	计算时间短，适合实时应用	合成图质量低，会出现失真、模糊、伪影等问题，可控性差
	图像重构解码器	Li等[27]提出多层次风格化策略，无须针对某一特定风格进行训练		

1.2.1 基于成对/非成对数据的图像转换

Isola等[28]提出的Pix2pix模型是一种基于条件生成对抗网络(conditional GAN,cGAN)[29]的图像到图像的转换模型,利用成对的图像进行图像转换,模型结构如图6所示。生成器使用基于U-Net[30]的网络架构代替传统的编码器和解码器结构。虽然该模型最终得到的生成图像接近真实图像,但是训练该模型时是以大量成对的图像数据进行训练,而收集成对的图像数据集是比较困难的。例如当用模型生成以《带珍珠耳环的少女》为代表作的荷兰画家约翰内斯·维米尔风格的画时,由于他留存于世的作品非常少,在此情景下不适合使用Pix2pix模型。

图6 Pix2pix结构图

Fig.6 Pix2pix structure diagram

CycleGAN则不需要成对的数据用于模型训练,该模型包含两个生成器 G 和 F 、两个判别器 D_x 和 D_y 。模型结构与循环一致性损失如图7所示。它的出现首次实现了不成对图像之间的变换,在风格迁移领域得到了广泛的应用,但也有如下缺点:(1)更倾向于改变输入图像的色彩风格而非几何结构。(2)当输入图像中包含训练过程中输入所不包含的图像时,生成器的映射会产生多样性变化。(3)模型并不能生成多样的结果且对数据的泛化能力比较弱。

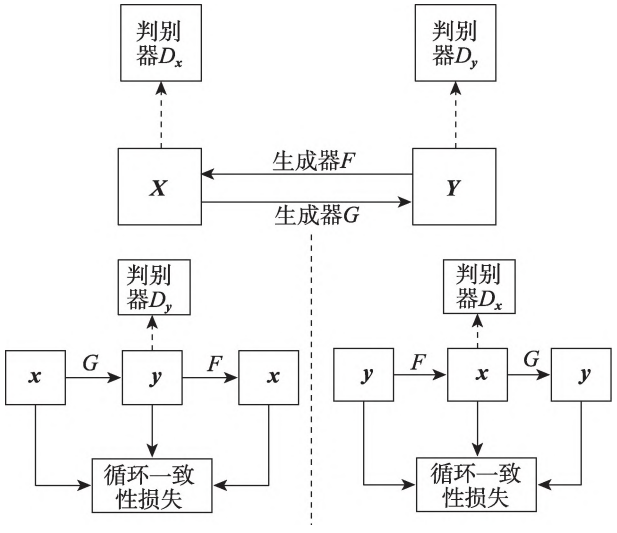


图7 CycleGAN结构图

Fig.7 CycleGAN structure diagram

Sanakoyeu等[31]针对CycleGAN泛化能力弱的问题,引入一个风格感知损失函数,可使模型训练出某一大类的艺术风格而不是特定类型的风格;Zhu等[32]针对CycleGAN生成结果单一的问题,提出的Bicycle-GAN模型实现了图像多类风格的生成。

1.2.2 基于条件/风格信息的图像生成

Brock等[33]提出的BigGAN是将多个CycleGAN组合成的更大的模型,相较于CycleGAN,其增加了模型规模,提高了训练效率和避免了局部最优解的问题,同时增加领域多样性,提高模型的性能和鲁棒性。模型结构如图8所示。首先将噪声向量通过分割等分成多块,然后和类别标签合并后一起送入网络中的不同层。并引入正则化思想,采用“截断技巧”通过对输入先验分布的适时截断使得模型对样本多样性和保真度可以进行精细控制。BigGAN的出现为之后大规模训练基于GAN原始结构的模型提供了一种方法,并且发现了大规模GAN特有的不稳定性以及提出可用于减少这种不稳定性的技术方法。虽然其具有生成图高度逼真的特点,但同时它也具有模型体量庞大、参数多、训练成本高等缺点。

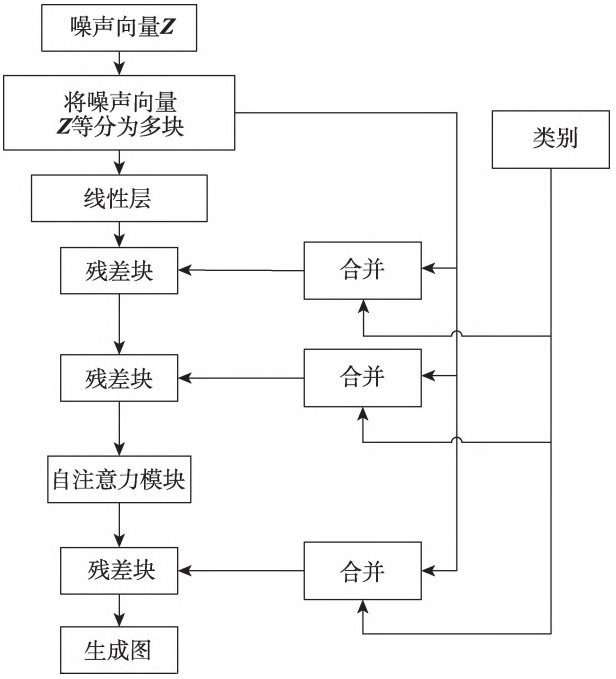


图8 BigGAN结构图

Fig.8 BigGAN structure diagram

相比于BigGAN通过扩大模型容量、使用分层的潜在空间等技术来提高生成图像的质量和多样性,

Karras等[34]提出的StyleGAN模型则引入了风格向量和自适应实例归一化来控制生成图像的风格和细节，结构如图9所示。该模型由映射网络 f 和生成网络 g 组成。f 将原始的潜在向量映射到一个解耦的风格空间，使得不同维度的风格向量可以对应不同层次的图像特征，实现更细粒度的图像编辑。g 则将风格向量通过自适应实例归一化操作注入到每个卷积层中，使得生成图像的风格和细节可以由风格向量控制。StyleGAN系列目前已经更新至StyleGAN3[35]，其解决了之前版本中图像在平移、旋转过程中细节沾黏问题，大幅度提高了图像的生成质量。

GAN在不同方面大大推动了AI绘画的发展，但是依然存在一些缺陷，比如当面对复杂和多样的数据时，模型的稳定性和收敛性较差，对输出结果的控制力较弱，容易产生随机图像等。更为重要的是，根据GAN的基本架构，生成的内容非常接近现有内容，这意味着输出的图像是对现有作品的模仿，而不是创新，这并不符合人们对于AI绘画的期待。表3对上述所介绍的GAN进行了归纳总结和对比分析。

1.3 以图生图模型评价方法

模型评价方法是一种用于衡量模型生成图像质量和多样性的方法，它可以帮助人们比较不同模型的优劣，指导模型的改进和应用。通常图像评价方法主要有客观和主观两种。客观的评价方法是指用一些数学或统计的指标来对生成图像进行量化或分类，根据它们与真实图像或目标图像之间的距离或相似度来评价图像的质量和多样性；主观的评价方法则根据人类的主观感受和偏好来评价图像的质量和多样性。表4总结了常用的客观图片评价方法，在表4中用Q表示图片质量评价方法，用D表示多样性评价方法，用QD表示同时考虑图片质量和多样性的评价方法。

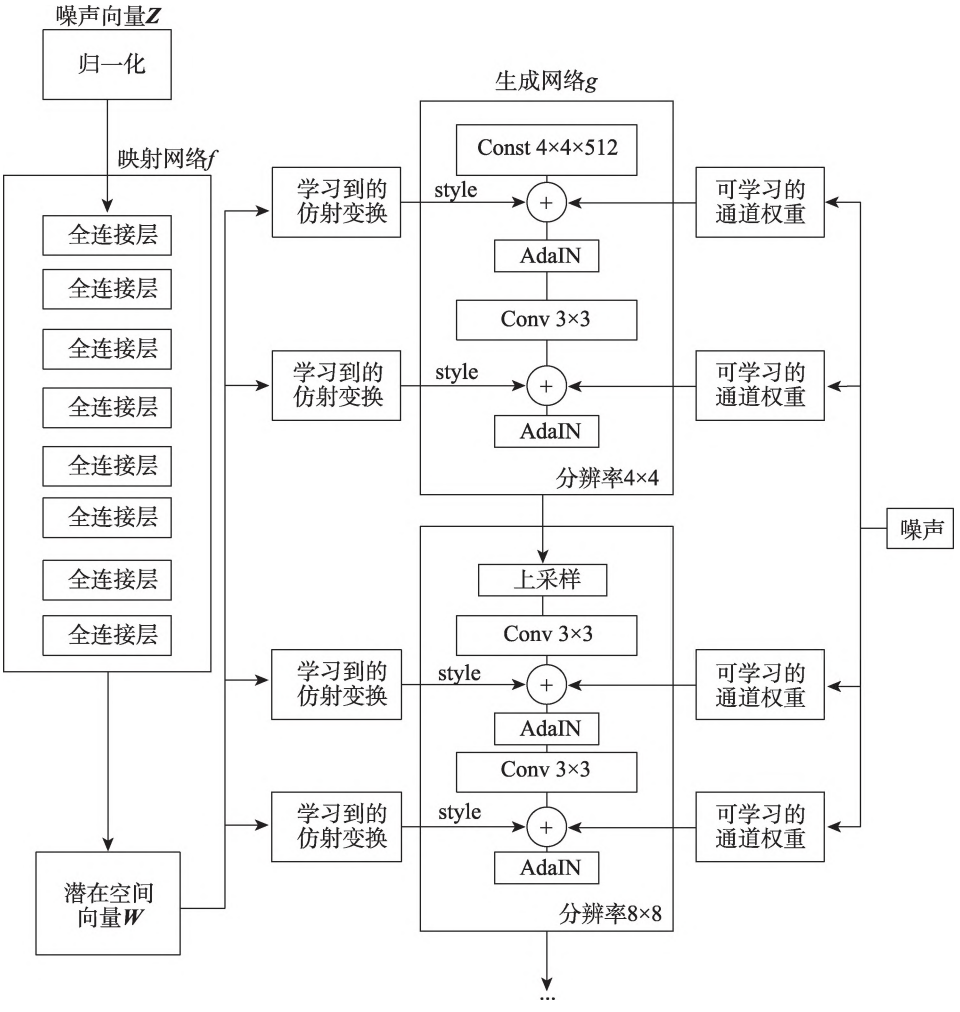


图9 StyleGAN结构图

Fig.9 StyleGAN structure diagram

表3 部分生成式对抗网络模型总结

Table 3 Summary of partial generative adversarial networks

模型	特点	优点	缺点	适用场景
Pix2pix[28]	基于cGAN[29]利用成对的数据进行图像翻译	可以实现高分辨率的图像翻译,同时保留内容和风格特征	需要成对的数据训练,只能学习一对一的映射,泛化能力弱	图像修复、图像着色、图像分割等
CycleGAN[26]	由两对生成器和判别器组成,通过循环一致性损失实现不同域之间的图像转换	不需要成对的数据训练;可实现多种风格转换	需要训练四个网络,计算量大,转换效果不自然或失真	图像风格迁移、图像着色、图像超分辨率等
BicycleGAN[32]	基于CycleGAN,通过双向映射实现多模态的图像生成,使用随机噪声作为潜在编码控制输出的多样性	不需要成对的数据训练;可实现一对多的映射,增加输出的多样性和鲁棒性	需要训练六个网络,计算量大,转换效果会出现不稳定或模糊	图像风格迁移、图像着色、图像超分辨率等
BigGAN[33]	通过扩大模型容量、使用分层的潜在空间、加入正交正则化和谱归一化等技术提高生成图质量和多样性	可以利用大规模的数据集,生成更多类别的高质量图像	需要大量的计算资源和数据集,训练时间长	图像生成、图像编辑、图像增强等
StyleGAN[34]	通过引入风格向量和自适应实例归一化来控制生成图的风格和细节,使得潜在空间更加解耦和可控	可以实现更细粒度的图像编辑和风格混合,生成效果逼真	需要针对不同的数据集进行调整和优化,难以泛化到其他领域	图像生成、图像编辑、图像风格化等

表4 图片评价方法

Table 4 Image evaluation methods

评价方法	原理	优点	缺点	适用范围	类型
PSNR	基于均方误差(MSE)方法。值越大,则生成图像越接近真实图像	简单易用,可以直观地反映生成图像与真实图像之间的差异程度	忽略了人类对图像感知的非线性特性,不能很好地反映生成图像的视觉质量	衡量一些需要保持原始信息不变或者只有微小变化的生成任务,比如去噪、超分辨率等	Q
SSIM	基于结构相似性。值越接近1,则生成图像越接近真实图像	考虑了人类对图像感知的结构特性,可以更好地反映生成图像的视觉质量	忽略了人类对图像感知的高阶特征,如纹理、边缘等	衡量一些需要保持结构信息不变或者只有微小变化的生成任务,比如去模糊、去雾等	Q
bits/dimension	根据模型对每个像素的概率分布来计算。值越低,生成图越接近真实图	是一种无参考的评价方法,不需要额外的数据集或标准来进行比较	只能评价模型对图像的整体质量,不能反映出图像的局部细节或语义信息	适用于评价基于最大似然估计的图像生成模型	Q
IS	基于Inception网络。值越大,生成图像越清晰且多样	简单易用,可以同时反映生成图像的质量和多样性	依赖于Inception网络的结构和参数,不能很好地适应不同领域或任务的数据集	衡量一些需要产生清晰且多样的图像的生成任务,比如GAN等	QD
FID	基于Inception网络。值越小,生成图像越接近真实图像	不依赖于网络输出类别,可以更好地适应不同领域或任务数据集	计算复杂度较高,需要计算特征空间中的均值和协方差矩阵	衡量一些需要产生逼真且多样的图像的生成任务,比如GAN等	Q
LPIPS	基于深度学习网络。值越小,则生成图像越接近真实图像	考虑了人类对图像感知的高阶特征,可以更好地反映生成图像的视觉质量	需要训练权重参数,且对不同数据集或任务可能需要不同的权重参数	衡量一些需要产生高质量且多样的图像的生成任务,比如超分辨率、风格迁移等	QD

为了更好地评价以图生图模型的性能和效果,需要对不同模型进行横向的数据对比,考察它们在不同数据集和实验环境下的表现。表5展示了一些以图生图模型在CIFAR-10、ImageNet 64、CelebA 64这三个数据集上的性能指标。

从表5中可以看到,目前以图生图模型的发展趋势是朝着大模型和扩散模型的方向进行的。大模型可以利用更多的参数和计算资源来提高生成图像的质量和多样性,如StyleGAN-XL模型就是一个拥有1.2×10^{10}参数的超大模型。同时这两种方向可以结合,如Diffusion StyleGAN2模型就是将扩散模型和StyleGAN2模型结合起来的。一般来说,大模型的训练对硬件也有着较高的要求,需要更多的显存和内存,以及更快的处理器和操作系统。但是可以通过调整训练的Batch Size大小或使用分布式训练的方式来降低硬件的压力。

表5 部分以图生图模型对比

Table 5 Comparison of some image-to-image models

数据集	数据集介绍	模型	模型参数量/10^6	FID得分	IS得分
CIFAR-10	由60 000张32×32彩色图像组成。这些图像被标记为10个互斥类别之一：飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。每个类别有6 000张图片，其中有5 000张训练图片和1 000张测试图片	StyleGAN-XL[36]	≈12 000	1.85	—
		TransGAN[37]	≈22	9.26	9.02
		MIX-BigGAN[38]	≈50	8.17	9.67
		E2-GAN[39]	≈8.6	11.26	8.51
		PPOGAN[40]	≈8.7	10.70	8.69
		MSGAN[41]	≈8.9	11.40	—
		AutoGAN[42]	≈8.5	12.42	8.55
		WGAN-ALP[43]	≈8.4	12.96	8.34
		PA-GAN[44]	≈8.6	16.10	—
		BigGAN[33]	≈143	14.73	9.22
ImageNet 64	将原始ImageNet数据集下采样到64×64，共包含1 281 167张训练图像和50 000张测试图像，分为1 000个类别	RIN[45]	≈120	1.23	66.50
		CDM[46]	≈150	1.48	—
		StyleGAN-XL	≈12 000	1.51	—
		CTM[47]	≈140	1.73	64.29
		Improved DDPM[48]	≈8.6	2.92	—
CelebA 64	包含了202 599张名人人脸图片，每张图片都有40个属性标签，如性别、年龄、发型等	Diffusion StyleGAN2[49]	≈140	1.69	—
		Patch Diffusion[50]	≈150	1.77	—
		Soft Diffusion[51]	≈8.6	1.85	—

2 以文生图模型

这类模型可以实现从自然语言描述到图像生成，而不受现有图像数据的限制，具有很高的应用价值。

本章归纳和整理了以文生图的相关研究工作，将其发展历程按照生成机制分为三个阶段：基于GAN、基于VAE和预训练语言模型（pre-trained language model，PLM）、基于扩散模型。需要说明的是，这三个阶段并没有一个明显的划分，它们之间存在着相互影响和借鉴的关系。事实上，随着深度学习技术的不断发展和创新，文生图模型也在不断地探索和尝试新的生成机制和方法。因此，本章所总结的三个阶段只是为了方便理解和分析文生图模型的发展历程，并不代表文生图模型的发展是线性和单一的。图10给出了以文生图模型的发展脉络。

2.1 基于GAN的以文生图模型

该类模型主要利用GAN的生成器和判别器来学习文本和图像之间的映射关系，从而生成与文本语义相关的图像。赖丽娜等[52]对基于GAN的图文生成方法进行了综述，本文在此基础上细分为基于条件信息、基于多阶段结构与基于注意力机制的图像生成类别。

2.1.1 基于条件信息的图像生成

cGAN在GAN的基础上增加了一个标签信息，这样生成器就可以根据标签信息来生成与之匹配的图像，判别器可以根据标签信息来判断图像是否真实和一致。

Reed等[53]在cGAN的基础上提出的GAN-INT-CLS模型[54]成为首个基于GAN的以文生图模型，该网络包含两个判别器：GAN-CLS与GAN-INT。其中GAN-CLS不仅需要判断真伪，还需要判别图片是否按照文本的要求生成，而GAN-INT用于做数据插值。虽然模型最终只能生成分辨率为64×64的图像，但是为文本生成图像任务提供了新方法。Reed等[55]在GAN-INT-CLS模型的基础上，引入了注意力机制和写入网络，提出的基于渐进式增长生成对抗网络（progressive growing of GAN，PGGAN）[56]的GAWWN（generative adversarial what-where network）模型从条件编码入手，支持输入具体位置描述来绘制对应区域内容，将生成图像分辨率提升至128×128。

2.1.2 基于多阶段结构的图像生成

Han等提出的StackGAN[57]与StackGAN++[58]模型从优化生成器出发，提出堆叠思想，将多个GAN像栈一样串行，串中第一个模型生成包含对象的粗糙形状和颜色的低分辨率图像，后续的GAN在此基础上分别完成高分辨率生成任务。

Yin等[59]提出的SDGAN（semantics disentangling

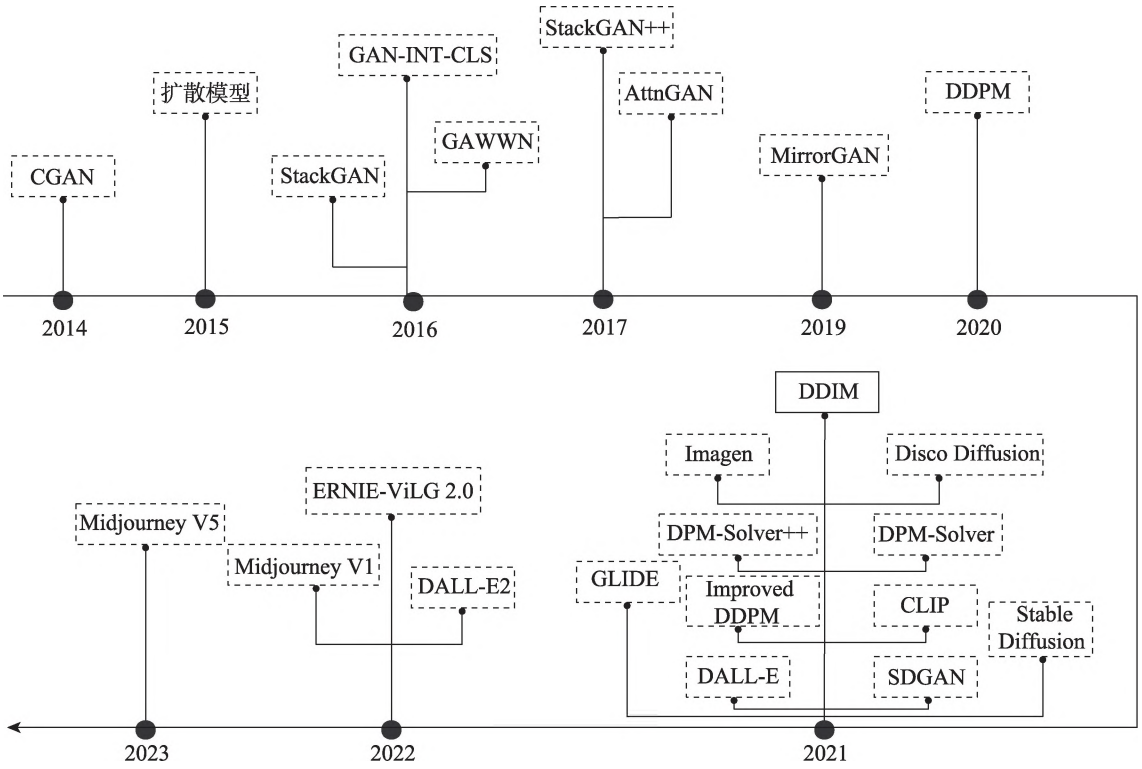


图10 以文生图模型发展脉络

Fig.10 Development line of text-to-image model

GAN)设计了一个基于孪生机制的判别器,使用对比损失来学习一致的高层语义和一个基于注意力机制的生成器,使用语义条件批归一化来融合多样的低层语义。

2.1.3 基于注意力机制的图像生成

Xu等[60]在GAN模型中引入注意力机制,提出AttnGAN模型,通过注意力机制和多阶段细化,克服图像细节丢失问题,实现了对文本描述的高质量细粒度图像生成。

Qiao等[61]在AttnGAN的基础上提出MirrorGAN模型,该模型使用了一个镜像组件,引入了一个全局与局部协同关注模块,利用局部词注意和全局句子注意,逐步增强生成图像的多样性和语义一致性。

2.2 基于VAE和PLM的以文生图模型

这类模型主要利用VAE来学习文本和图像之间的潜在空间,通过结合PLM来提高文本理解能力和生成图像的精度与多样性。

2.2.1 基于VAE和Transformer

DALL-E[62]是基于VQ-VAE-2[63]和Transformer[64]的文生图模型。它结合了一个高容量的Transformer模型,作为编码器和解码器的一部分,使用VQ-VAE-2将图像压缩为一组离散的编码,然后用Transformer模型来理解和生成这些编码。虽然在语义理解和创造性图像生成方面表现出色,能进行融合创作、场景理解和风格转换,但其在零样本和小众专业领域的图像生成质量可能不高,且分辨率有限。

2.2.2 基于VAE和对比学习

CLIP(contrastive language-image pre-training)[65]模型可以从自然语言的监督中有效地学习视觉概念,并在不直接优化任务的情况下根据文本预测最相关的图像。利用对比学习方法,通过让模型在大量的文本图像对上进行预训练,模型能够区分正负样本。在CLIP提出之后,人们思考可以依靠其图文匹配验证机制来引导图像特征向量匹配指定的文本条件编码向量,这样就可以得到符合文字描述的图片。基于将CLIP与其他图像生成模型结合的思路,Patashnik等[66]提出CLIP与StyleGAN的混合模型StyleCLIP;Crowson等[67]提出VQGAN-CLIP模型。

DALL-E2[68]将图像分辨率提升至1 024×1 024,并且图像更加贴合语义信息,该模型结构如图11所示。图11中虚线上方为CLIP模型,用于对齐图像文本特征,虚线下方由先验模型和解码器及扩散模型组成,其中先验模型用于接收文本信息,将其转换为CLIP图像表征,扩散模型用于接收图像表征来生成

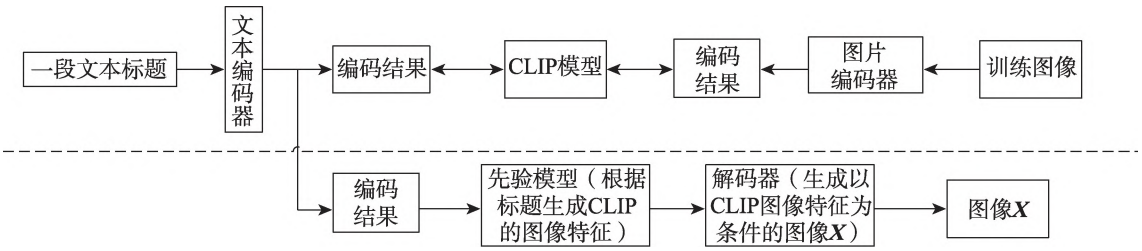


图11 DALL-E2结构图

Fig.11 DALL-E2 structure diagram

完整图像 X 。

Midjourney绘画平台则将VQGAN与CLIP进行了结合，从图像的潜在向量中生成图像。在其V5版本中使用了一种基于Transformer的被称为OpenCLIP的预训练语言模型。该模型可以同时处理文本和图像，并学习它们之间的语义对齐，同样还使用了一种自适应的生成策略，可以根据文本描述的复杂度和细节动态调整图像的分辨率和样式。

2.3 基于扩散模型的以文生图模型

扩散模型通过一个从数据分布到高斯分布的扩散过程和一个从高斯分布到数据分布的去噪过程来建模数据。扩散模型虽然具有如训练稳定性高、样本多样性高、不需要对数似然或分区函数、数学性质良好等优点，尤其是在与文本条件或其他引导技术结合时，可以生成逼真且符合语义的图像，但由于前向运算和逆向扩散过程需要较长的时间，导致模型训练和推理速度都比较慢，无法像前两个阶段模型那样高效地一步生成图像，因此扩散模型起初并没有受到人们广泛的关注。Yang等[69]对扩散模型的现有变体进行了全面回顾，并讨论了扩散模型与生成模型之间的联系。

2.3.1 无条件图像生成扩散模型

在提出扩散模型概念五年后，Ho等[10]提出了去噪扩散概率模型（denoising diffusion probabilistic models，DDPM）。该模型不仅能够生成高质量图像，且在图像超分辨率和图像修复方面也有较好的性能。模型原理如图12所示。前向扩散过程为图12中从右到左即 $\boldsymbol{X}_0 \rightarrow \boldsymbol{X}_T$ 的过程，是指对数据逐步添加高斯噪声直至数据变为随机噪声的过程。逆向扩散过程为从左到右 $\boldsymbol{X}_T \rightarrow \boldsymbol{X}_0$ 的过程，通过逐步预测噪声的分布，进而去除噪声生成图像。

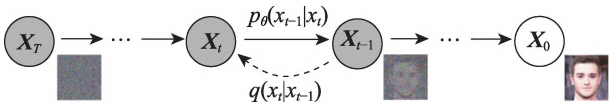


图12 DDPM结构图

Fig.12 DDPM structure diagram

DDPM的出现重新激发了人们对扩散模型的研究兴趣，在此之后有众多改进模型被提出。表6对部分改进模型进行了汇总整理。

Disco Diffusion绘画平台利用CLIP来引导扩散过程，用户可以根据需求选择不同的CLIP模型来影响生成图像的风格和质量。它还提供了一个次级扩散模型，该模型根据当前扩散步数和CLIP损失来动态调整模型的参数，以提高模型的表现。

Nichol等[73]提出的GLIDE（guided language to image diffusion for generation and editing）模型可以根据任意文本生成1 024×1 024分辨率的图像。该模型使用了一个潜在扩散模型（latent diffusion model，LDM）作为图像生成过程的基础，还使用了一个潜在变量

表6 基于DDPM模型的部分改进模型

Table 6 Partially improved models based on DDPM model

模型	改进点	效果
Improved DDPM[48]	针对DDPM在对数似然方面的不足进行了改进	在图像生成方面取得了更好的效果，并且可以达到更高的对数似然
DDIM[70]	使用一个基于损失函数的重要性采样方法	提高了采样的效率和多样性
DPM-Solver[71]	使用了一个高阶常微分方程求解器来进行采样	提高了采样精度和速度，减少了截断误差
DPM-Solver++[72]	在DPM-Solver的基础上，使用了一个神经网络来学习方差序列	提高了模型的灵活性和表达能力，还可以保留更多的图像细节和信息

编码器、U-Net网络和一个文本编码器，其中潜在变量编码器用于将图像编码为类别信息和样式信息的潜在表示，U-Net网络用于将文本编码器的输出和潜在变量编码器的输出融合，文本编码器则用于引导扩散过程符合文本描述。同时该模型还使用了一个全局-局部注意力机制来增强文本编码器和扩散解码器之间的对齐，但其在一些少见的文本描述上很难产生合乎逻辑的图像。

2.3.2 条件图像生成扩散模型

GLIDE模型是以文生图领域的一个重要里程碑，但它并不是终点。谷歌提出基于扩散模型思想的新的文本生成图像模型Imagen[74]。该模型引入了一种阈值扩散采样器，这种采样器可以使用较大的无分类器指导权重，同时还使用了一种新的U-Net网络架构，这种架构具有更高的计算效率、内存效率和更快的收敛速度。Imagen使用T5-XXL语言大模型直接编码文本信息，然后使用条件扩散模型直接用文本编码生成图像。由于其直接使用T5-XXL模型而无需学习先验模型，其语义知识相对于CLIP要丰富很多。

Rombach等[75]提出的Stable Diffusion模型是一种高分辨率、细节丰富、语义一致的以文生图模型。模型采用LDM图像生成范式，不同于以往在像素空间进行扩散过程，LDM则考虑在较低维度的潜在空间中进行扩散过程，这样就极大减轻了训练以及推理成本。

百度在2022年也发布了其文图生成大模型ERNIE-ViLG 2.0[76]，该模型在公开数据集MS-COCO验证集中的zero-shot FID-30K和人工盲评上均超越了Stable Diffusion、DALL-E2等模型。ERNIE-ViLG 2.0模型采用基于知识增强算法的混合降噪专家建模，让模型在不同的生成阶段选择不同的“降噪专家”网络，从而实现更加细致的降噪任务建模，进而提升生成图像的质量。

2.4 以文生图模型总结

本文对这三个阶段的模型从主要贡献和局限性方面进行详细的对比，分析它们的优势和局限性。对比结果如表7所示。

2.5 以文生图模型评价方法

对于以文生图模型的评价，主要使用的评价指标有：(1)语义一致性，指生成图像与输入文本之间的语义匹配程度。(2)创新性，评价生成图像的新颖性，判断生成图是否符合输入文本的描述，同时又有一些新颖的创意。(3)多模态性，评价模型是否能有效地融合文本和图像信息，生成与输入文本高度一致的图像。表8展示了一些以文生图模型在MS-COCO、CUB-200-2011和Multi-Modal-CelebA-HQ这三个数据集上的性能指标，包括模型参数量、FID得分和IS得分。

3 AI绘画平台对比

目前市面上主流的AI绘画平台有DALL-E2、Stable Diffusion、Midjourney等，本文对目前较为流行的平台从不同方面进行汇总对比，结果如表9所示。

目前，各大AI绘画平台都对用户的绘画任务进行收费，采用收费模式可能有以下几点原因：(1)算力成本，AI绘画平台需要使用高性能的显卡或云端服务器来运行复杂的AI模型。(2)模型开发，各大平

表7 第一、二、三阶段模型对比

Table 7 Comparison of the first, second and third stage models

阶段	主要贡献	局限性	代表模型
第一阶段	开创了文本到图像合成领域的研究方向，提出了多种有效而创新的方法来改进GAN的结构和损失函数，提高了图像生成的质量和多样性	依赖预训练的文本编码器和图像编码器来提取特征；使用固定长度的特征向量来表示文本描述；使用随机噪声作为生成器的输入之一，可能导致生成器无法控制图像的风格和内容或生成图像之间不连贯	StackGAN[57]、AttnGAN[60]、MirrorGAN[61]等
第二阶段	将对抗学习和其他正则化方法结合；并且可以与其他任务如图像检索、图像编辑、图像描述等结合，为各种领域提供了新的视觉表达方式	需要大量的数据和计算资源来训练；使用隐变量作为生成器的输入之一，这可能导致隐变量的后验分布与先验分布不匹配，或信息量不足，从而影响生成质量和多样性	DALL-E[62]、CLIP[65]、DALL-E2[68]等
第三阶段	开创性地提出扩散思想；可以与其他任务如图像修复、图像增强等结合，展示了强大的生成能力和灵活性	需要较长采样时间和更复杂的优化方法，可能导致效率低下或者不稳定；使用固定长度的特征向量来表示文本描述；难以评估生成图像与文本之间的相似度	DDPM[10]、Improved DDPM[48]、DDIM[70]、Stable Diffusion[75]等

表8 部分以文生图模型对比

Table 8 Comparison of some text-to-image models

数据集	数据集介绍	模型	模型参数量/10^6	FID得分	IS得分
MS-COCO	是由微软开发维护的大型图像数据集，具有非常高的行业地位且规模非常庞大。数据集共包含123 287幅图像，包含80 000张用于训练的图像和40 000张用于测试的图像。其中每个图像包含5个句子注释	Re-Imagen[77]	≈1 200	5.25	—
		TLDM[78]	—	6.29	—
		ERNIE-ViLG 2.0[76]	≈2 400	6.75	—
		Imagen[74]	≈1 500	7.27	—
		DALL-E[62]	≈1 500	27.50	17.90
		DALL-E2[68]	≈350	10.39	—
		GLIDE[73]	—	12.24	—
		Stable Diffusion[75]	≈600	12.63	—
CUB-200-2011	该数据集是精细视觉分类任务中使用最广泛的数据集。它包含属于鸟类的200个子类别的11 788张图像，其中5 994张用于训练，5 794张用于测试。每张图像都具有1个子类别标签、15个部分位置信息、312个二进制属性和1个边界框	TLDM	—	6.72	—
		Swinv2-Imagen[79]	≈3 000	9.78	8.44
		StackGAN++[58]	≈138.3	15.30	3.82
		GAWWN[55]	≈41.7	67.22	3.62
		MirrorGAN[61]	≈175.8	—	4.56
		AttnGAN[60]	≈86.7	—	4.36
Multi-Modal-CelebA-HQ	是一个大规模人脸图像数据集，具有30 000张高分辨率人脸图像，每个图像都对应10个描述性文本，还有高质量的分割蒙版、素描和透明背景图像	Swinv2-Imagen	≈3 000	10.31	—
		TediGAN-B[80]	≈41.7	101.42	—
		TediGAN-A[81]	≈175.8	106.37	—

表9 AI绘画主流平台对比

Table 9 Comparison of AI painting mainstream platforms

AI绘画平台	基础模型	使用方式	生成速度/s	图像分辨率
DALL-E2	扩散模型+CLIP	网页输入文本描述	≈1	1 024×1 024
Midjourney	VAE+PLM+GAN	Discord输入文本描述	≈10	2 048×1 280
Imagen	扩散模型+CLIP+其他正则化方法	目前无法使用	—	1 024×1 024
Stable Diffusion	扩散模型+CLIP+其他正则化方法	网页输入文本描述或上传图像进行生成、编辑、转换等	≈10	取决于输入文本描述或上传图像的分辨率
Fotor	VAE+GAN+其他正则化方法	网页上传图像进行编辑、美化、滤镜等	≈5	取决于上传图像的分辨率
NightCafe	GAN+其他正则化方法	网页上传图像进行风格迁移、艺术效果等	≈10	取决于上传图像的分辨率
DeepAI	GAN+其他正则化方法	网页输入文本描述或上传图像进行生成、编辑、转换等	≈5	取决于输入文本描述或上传图像的分辨率
百度文心一格	ERNIE-ViLG+GAN+其他正则化方法	网页输入文本描述或上传图像进行生成、编辑、转换等	≈5	取决于输入文本描述或上传图像的分辨率

台需要不断地研发和优化自己的AI模型，以提高生成质量，适应不同的用户需求和场景。(3)版权保护，平台需要对版权问题进行合法合规的处理，以避免法律风险。(4)商业模式，AI绘画作为一种创新的技术应用，也需要探索可持续的商业模式，以实现盈利和发展。

4 AI绘画存在的问题

AI绘画在近年来取得了显著的进步和发展，各大AI绘画平台展示了人工智能在图像生成方面的强大能力和潜力，为人们提供了一个快速、便捷、有趣的方式来生成各种风格和主题的图像，但是它们并不是完美无缺的。事实上，AI绘画还有许多需要解决的问题和困难。

4.1 AI绘画在技术层面存在的问题

本节将从数据集、映射和控制、生成质量、训练成本、模型评价五个方面来分析AI绘画在技术层面存在的问题。

4.1.1 数据集的标注、选取与划分问题

在数据集的内容标注方面，由于以文生图的输

入为文本信息，其语义空间要比词组空间更大，需要标注大量的数据用于模型训练。在数据集的选取方面，数据集往往难以获取或成本高昂，如果数据集的质量不高、规模不够大，那么生成图可能会出现噪声、失真等不自然现象。在数据集划分方面，由于划分方式的不同，会导致实验结果的可对比性减弱，即无法确定当下实验结果的优劣是模型自身的原因，还是数据集划分方式不一致带来的影响。

4.1.2 文本到图像的映射和控制问题

该问题表现为生成图中所包含的语义信息与人们输入的语义信息不符。目前主要有两种方法来实现文本到图像的映射和控制，一种是基于cGAN的方法，通过引入文本编码器来约束图像生成过程，另一种是基于跨模态的方法，它通过使用如CLIP等预训练的语言-视觉模型来计算文本和图像之间的相似度，并用其作为优化目标。两种方法都有各自的优缺点，cGAN可以生成更精确和更细致的图像，但也容易受到文本编码器质量和数据集规模的影响；跨模态方法可以生成更多样性和创新性的图像，但也容易出现文不对题或画面结构混乱的问题。

4.1.3 AI绘画的生成质量问题

存在的问题有：(1)无法保证输出图像均符合质量要求，例如《太空歌剧院》作品是该作者从900多张生成图中挑选的效果最好的一张，并且在生成过程中进行了多次人为输入文本的修改。(2)模型对文本的理解能力有限，例如对物体之间的位置关系理解力偏弱。(3)生成图像不符合人类常识，如在绘制人体的眼睛、手、脚等部位细节时，往往生成的内容不符合大多数人的生理特征，存在手指或脚指的数量出错、手指关节的移位等问题。(4)在风格化和个性化方面表现不足。

4.1.4 训练成本问题

目前AI绘画模型的参数量由十亿至百亿不等，这是一个相当大的参数量，而拥有如此庞大参数量的模型其训练成本也是非常高昂的。如此耗时耗财的大模型反馈到用户身上就是使用这些大模型需要支付一定的费用，并且对于普通学者与众多小公司来说无法训练如此规模的模型。

4.1.5 模型评价问题

对于模型评价方法而言，目前还没有一个普适的评价方法来对AI绘画进行客观和量化的评估，因为AI绘画涉及到很多主观和复杂的因素，如美感、创意、语义、多样性等。不同的用户、不同的任务、不同的数据集、不同的模型都可能导致不同的评价标准和结果。因此，目前对AI绘画的评价主要是基于人的主观评价方法与客观评价方法相结合。主观评价方法虽然可以反映人类对图像的真实感知和喜好，但也存在一些问题，如评价标准不统一、评价成本高、评价结果不稳定等。

4.2 AI绘画在社会影响层面存在的问题

AI绘画不仅是一种技术，也是一种艺术，它涉及到人类的创造力、审美、情感和价值观等方面。因此，AI绘画也会对社会产生一定的影响，这些影响可能是积极的，也可能是消极的，甚至是有争议的。本节将从三个方面来分析AI绘画在社会影响层面存在的问题。

4.2.1 版权问题

由于AI绘画的基础算法模型所需要的训练数据大多是从互联网的素材库中获取的，AI绘画作品到底受不受版权保护，权利又归谁所有，目前还没有明确规定，处于争论状态。一方认为AI绘画的实质是应用计算机算法程序与规则来产出作品，是一种算法，创作体现了开发者的思想与脑力劳动，应当受到法律的保护且权利属于AI绘画的创作者。另一方则认为AI模型生成的图像是对原始数据的复制和重组，并没有真正的创新和表达，因此版权应该归属于原始数据的版权方或共享版权。

4.2.2 伦理和道德问题

不法分子可能会通过AI作图用于政治宣传、邪教宣传等扩散不良内容，这可能会带来伦理、道德和法律问题，甚至为色情、暴力、赌博、毒品等违法信息提供平台。而在Stable Diffusion 2.0版本中，其为了防止色情内容和名人肖像的滥用，保护人类艺术家的权益，模型中添加了LAION(large-scale artificial intelligence open network)的NSFW(not safe for work)过滤器，该过滤器可以过滤掉色情内容，使得AI绘画作品不能被用于色情宣传。

4.2.3 行业挑战问题

该问题表现为AI绘画是否会造成人类艺术家失业或降低收入、AI绘画作品是否属于艺术品等问题。本文认为，虽然AI绘画越来越逼真，内容也越来越精彩，但是究其根本是不会让艺术家失业的。AI仅仅是工具，绘画依然需要画家本人的创造力，而AI可以取代的是大量重复性的绘画需求，画家可以充分利用AI完成这些工作，将精力更多放到创作本身。而AI绘画作品是否属于艺术品这个问题目前并没有一

个明确的答案。有些人认为AI绘画作品并不属于艺术品，艺术是包含人性与情感的，艺术的人性与情感不仅来源于艺术家的创作过程，也需要观赏者能够充分领悟艺术品所蕴含的情感，这是一个双向的过程。另一部分人认为AI绘画作品属于艺术品，因为目前的AI绘画仍然只能依赖已有的数据进行建模，也就是说AI绘画只能临摹已有的流派和表现形式，即无法开辟新的表现形式同时也无法自己创造出新的风格。

5 总结

AI绘画是一门兼具艺术与科学的技术，它不仅能够实现绘画创作的自动化和智能化，还能够拓展人类对于美和艺术的认知和理解。本文对AI绘画的研究现状及发展趋势进行了综述，首先介绍了AI绘画的发展历程与主要技术，然后探讨了AI绘画在技术层面和社会层面存在的问题。

以ChatGPT[82]为代表的AI语言模型，以GPT-4[83]为代表的多模态模型，以Midjourney为代表的AI绘画模型，让世界见识了AI的效率及能力。随着AI绘画的火爆，各大短视频平台也迅速上线了AI绘画特效，人们通过该特效可以看到自己的形象变成了动漫人物或古风人物、数字人物。AI绘画也被应用于艺术领域，它的出现大幅降低了绘画门槛，普通人轻而易举就可以创作出风格各异的作品，一个只会写作的作家，完全可以利用AI为自己的作品制作插图。企业也有望大幅提升制图的效率，减少在相关方面的支出。不仅如此，AI绘画还可以应用于设计、建筑和医疗等领域。在设计领域，AI绘画可以帮助设计师进行快速的原型设计和效果图生成；在建筑领域，AI绘画可以帮助建筑师进行建筑方案的生成和优化，生成建筑草图或效果图，模拟建筑环境和光照等；在医疗领域，AI绘画可以帮助医生进行医学图像的分析与诊断。

6 AI绘画的未来展望

本文认为在未来，AI绘画会有如下变化：(1)AI绘画和画家作画两者可能会朝着不同的方向发展。AI绘画不再是对现有流派和风格的组合，可能会形成自己特有的数据风格，使得模型更加具有审美意义和创造能力。画家的作品虽然在绘画技巧上不及AI绘画作品，但是它更多带给观众的是人性化的思考与表达，更加注重作品传递的情感与内涵。因此，两者可能会在艺术领域中形成一种互补而非替代的关系。(2)AI绘画将会具有互动性，人们在使用AI绘画的过程中，可以在模型运行的中间步骤与AI互动，像与ChatGPT那样对话的形式，模型尽可能多尽可能准确地理解用户的需求来进行创作。这种互动性可以让用户进一步参与到创作过程中，并且提高用户的满意度和归属感。(3)AI绘画将会催生出新职业与新产业。一种新职业表现在文本提示方面，这类职业从业者会对普通用户的需求进行高度提炼总结，找到能够使模型最大化满足需求的文本提示，然后交由用户；另一种新职业表现在图像再处理任务上，这类职业从业者就可以通过对AI生成的初稿进行人工精修，从而达到用户期待的效果。新产业既可以表现在售卖AI绘画作品方面，也可以表现在元宇宙、设计、文旅等行业。

未来的AI绘画模型生成图像的方式也会更加多元化，例如可以声音生图，通过与3D建模、虚拟现实技术相结合，通过语音指令来生成自己想要的图像，并可以通过VR设备做到实时预览。总之，AI绘画在未来将会越来越受到人们的关注，有越来越多的应用场景。

参考文献：

[1] 冯强. 人工智能绘画的艺术价值及未来发展研究[D]. 沈阳: 鲁迅美术学院, 2021.
FENG Q. Research on artistic value and future development of artificial intelligence painting[D]. Shenyang: Luxun Academy of Fine Arts, 2021.

[2] 列夫·马诺维奇, 埃马努埃莱·阿列利, 陈卓轩. 列夫·马诺维奇: 人工智能(AI)艺术与美学[J]. 世界电影, 2023(3): 4-24.
MANOVICH L, ARIELLI E, CHEN Z X. Lev Manovich: art and aesthetics of artificial intelligence (AI) [J]. World Cinema, 2023(3): 4-24.

[3] 李白杨, 白云, 詹希旎, 等. 人工智能生成内容(AIGC)的技术特征与形态演进[J]. 图书情报知识, 2023, 40(1): 66-74.
LI B Y, BAI Y, ZHAN X N, et al. The technical features and aromorphosis of artificial intelligence generated content (AIGC) [J]. Library Intelligence Knowledge, 2023, 40(1): 66-74.

[4] GARCIA C. Harold Cohen and AARON—a 40-year collaboration[EB/OL]. (2016-08-23)[2023-03-18]. https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration.

[5] 周飞. 人工智能数字绘画的艺术性思辨[J]. 湖北经济学院学报(人文社会科学版), 2017, 14(7): 14-15.
ZHOU F. Thoughts on artistry of artificial intelligence digital painting[J]. Journal of Hubei University of Economics

(Humanities and Social Sciences), 2017, 14(7): 14-15.
[6] COLTON S. The painting fool: stories from building an automated painter[M]//Computers and Creativity. Berlin, Heidelberg: Springer, 2012: 3-38.
[7] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[8] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27, Montreal, Dec 8-13, 2014: 2672-2680.
[9] MARKOV A A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain[J]. Dynamic Probabilistic Systems, 1971, 1: 552-579.
[10] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems 33, Dec 6-12, 2020. Red Hook: Curran Associates, 2020: 6840-6851.
[11] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.
[12] BOURLARD H, KAMP Y. Auto-association by multilayer perceptrons and singular value decomposition[J]. Biological Cybernetics, 1988, 59(4/5): 291-294.
[13] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. Machine Learning, 2013, 106(9/10): 2979-3024.
[14] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
[15] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.
[16] VAN DEN OORD A, VINYALS KAVUKCUOGLU K. Neural discrete representation learning[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 6309-6318.
[17] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.
[18] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Pixel recurrent neural networks[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, Jun 19-24, 2016: 1747-1756.
[19] 陈淑環, 韦玉科, 徐乐, 等. 基于深度学习的图像风格迁移研究综述[J]. 计算机应用研究, 2019, 36(8): 2250-2255.
CHEN S H, WEI Y K, XU L, et al. Survey of image style transfer based on deep learning[J]. Application Research of Computers, 2019, 36(8): 2250-2255.
[20] 陈淮源, 张广驰, 陈高, 等. 基于深度学习的图像风格迁移研究进展[J]. 计算机工程与应用, 2021, 57(11): 37-45.
CHEN H Y, ZHANG G C, CHEN G, et al. Research progress of image style transfer based on deep learning[J]. Computer Engineering and Applications, 2021, 57(11): 37-45.
[21] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[J]. Computer Vision and Pattern Recognition, 2015, 29(2): 241-250.
[22] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-03-21]. https://arxiv.org/abs/1409.1556.
[23] LI C, WAND M. Combining Markov random fields and convolutional neural networks for image synthesis[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2016: 2479-2486.
[24] LIAO J, YAO Y, YUAN L, et al. Visual attribute transfer through deep image analogy[J]. ACM Transactions on Graphics, 2017, 36(4): 120.
[25] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 694-711.
[26] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Washington: IEEE Computer Society, 2017: 2223-2232.
[27] LI Y, FANG C, YANG J, et al. Universal style transfer via feature transforms[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 385-395.
[28] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2017: 5967-5976.
[29] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. [2023-03-21]. https://arxiv.org/abs/1411.1784.
[30] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 2015 International Conference on Medical Image Computing and Computer Assisted Intervention. Cham: Springer, 2015: 234-241.
[31] SANAKOYEU A, KOTOVENKO D, LANG S, et al. A style-aware content loss for real-time HD style transfer[C]//Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018: 715-731.
[32] ZHU J Y, ZHANG R, PATHAK D, et al. Toward multimodal image-to-image translation[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 465-476.
[33] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[J]. Nature Reviews Physics, 2021, 3(6): 422-440.
[34] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision

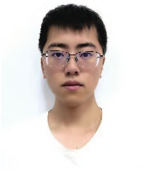
and Pattern Recognition. Piscataway: IEEE, 2019: 4401-4410.
[35] KARRAS T, LAINE S, AITTALA M, et al. Alias-free generative adversarial networks[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 852-863.
[36] SAUER A, SCHWARZ K, GEIGER A. StyleGAN-XL: scaling StyleGAN to large diverse datasets[EB/OL]. (2022-05-05)[2023-07-11]. https://arxiv.org/abs/2202.00273.
[37] JIANG Y, CHANG S, WANG Z. TransGAN: two pure transformers can make one strong GAN, and that can scale up [C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 14745-14758.
[38] TANG S. Lessons learned from the training of GANs on artificial datasets[EB/OL]. (2020-07-14)[2023-07-13]. https://arxiv.org/abs/2007.06418.
[39] ZHANG Y, ZHOU P, HUANG Z, et al. Off-policy reinforcement learning for efficient and effective GAN architecture search[C]//Proceedings of the 16th European Conference on Computer Vision. Cham: Springer, 2020: 175-192.
[40] WU Y, ZHOU P, WILSON A G, et al. Improving GAN training with probability ratio clipping and sample reweighting[C]//Advances in Neural Information Processing Systems 33, Dec 6-12, 2020. Red Hook: Curran Associates, 2020: 5729-5740.
[41] TRAN N T, TRAN V H, NGUYEN N B, et al. Self-supervised GAN: analysis and improvement with multi-class minimax game[C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019: 14761-14772.
[42] GONG X, CHANG S, JIANG Y, et al. AutoGAN: neural architecture search for generative adversarial networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 1-13.
[43] TERJÉK D. Adversarial lipschitz regularization[C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019. Red Hook: Curran Associates, 2019: 1-17.
[44] ZHANG D, KHOREVA A. Progressive augmentation of GANs [C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019. Red Hook: Curran Associates, 2019: 6249-6259.
[45] JABRI A, FLEET D J, CHEN T. Scalable adaptive computation for iterative Generation[EB/OL]. (2023-06-14)[2023-07-15]. https://arxiv.org/abs/2212.11972.
[46] HO J, SAHARIA C, CHAN W, et al. Cascaded diffusion models for high fidelity image generation[EB/OL]. (2021-12-17)[2023-07-15]. https://arxiv.org/abs/2106.15282.
[47] KIM D, LAI C H, LIAO W H, et al. Consistency trajectory models: learning probability flow ODE trajectory of diffusion[EB/OL]. (2023-10-01)[2023-11-07]. https://arxiv.org/abs/2310.02279.
[48] NICHOL A, DHARIWAL P. Improved denoising diffusion probabilistic models[C]//Proceedings of the 38th International Conference on Machine Learning, Jul 18-24, 2021:8162-8171.
[49] WANG Z, ZHOU P, HUANG Z, et al. Diffusion-GAN: training GANs with diffusion[EB/OL]. (2022-06-05)[2023-06-23]. https://arxiv.org/abs/2206.02262.
[50] WANG Z, JIANG Y, ZHENG H, et al. Patch diffusion: faster and more data-efficient training of diffusion models[EB/OL]. (2023-10-18)[2023-11-02]. https://arxiv.org/abs/2304.12526.
[51] DARAS G, DELBRACIO M, TALEBI H, et al. Soft diffusion: score matching for general corruptions[EB/OL]. (2022-10-05)[2023-06-20]. https://arxiv.org/abs/2209.05442.
[52] 赖丽娜, 米瑜, 周龙龙, 等. 生成对抗网络与文本图像生成方法综述[J]. 计算机工程与应用, 2023, 59(19): 21-39.
LAI L N, MI Y, ZHOU L L, et al. Survey about generative adversarial network and text-to-image synthesis[J]. Computer Engineering and Applications, 2023, 59(19): 21-39.
[53] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, Jun 19-24, 2016: 1060-1069.
[54] RADFORD A, METZ L, CHINTALA S, et al. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2016-01-07)[2023-03-16]. https://arxiv.org/abs/1511.06434.
[55] REED S, AKATA Z, MOHAN S, et al. Learning what and where to draw[C]//Advances in Neural Information Processing Systems 29, Barcelona, Dec 5-10, 2016: 241-250.
[56] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[C]//Proceedings of the 2018 International Conference on Learning Representations. Red Hook: Curran Associates, 2018: 1-26.
[57] ZHANG H, XU T, LI H, et al. StackGAN: text to photorealistic image synthesis with stacked generative adversarial networks [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 5907-5915.
[58] ZHANG H, XU T, LI H, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 694-711.
[59] YIN G, LIU B, SHENG L, et al. Semantics disentangling for text-to-image generation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2327-2336.
[60] XU T, ZHANG P, HUANG Q, et al. AttnGAN: finegrained text to image generation with attentional generative adversarial networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2018: 1316-1324.
[61] QIAO T, ZHANG J, XU D, et al. MirrorGAN: learning text-to-image gneration by redescription[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 1505-1514.
[62] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//Proceedings of the 38th International Conference on Machine Learning, Jul 18-24, 2021: 8821-8831.
[63] RAZAVI A, VAN DEN OORD A, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[C]//Advances in Neural Information Processing Systems 32, Vancouver,

Dec 8-14, 2019: 14761-14772.
[64] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017: 5998-6008.
[65] 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, Jul 18-24, 2021: 8748-8763.
[66] PATASHNIK O, WU Z, SHECHTMAN E, et al. StyleCLIP: text-driven manipulation of StyleGAN imagery[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 2085-2094.
[67] CROWSON K, BIDEMAN S, KORNIS D, et al. VQGAN-CLIP: open domain image generaton and editing with natural language guidance[C]//Proceedings of the 17th European Conference on Computer Vision. Cham: Springer, 2022: 88-105.
[68] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022-04-13)[2023-06-20]. https://arxiv.org/abs/2204.06125.
[69] YANG L, ZHANG Z, SONG Y, et al. Diffusion models: a comprehensive survey of methods and applications[EB/OL]. (2023-10-11)[2023-11-02]. https://arxiv.org/abs/2209.00796.
[70] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[EB/OL]. (2022-10-05)[2023-11-02]. https://arxiv.org/abs/2010.02502.
[71] LU C, ZHOU Y, BAO F, et al. DPM-Solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps[EB/OL]. (2022-10-13)[2023-04-02]. https://arxiv.org/abs/2206.00927.
[72] LU C, ZHOU Y, BAO F, et al. DPM-Solver++: fast solver for guided sampling of diffusion probabilistic models[EB/OL]. (2023-05-06)[2023-09-22]. https://arxiv.org/abs/2211.01095.
[73] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2022-03-08)[2023-09-28]. https://arxiv.org/abs/2112.10741.
[74] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[EB/OL]. (2022-05-23)[2023-08-18]. https://arxiv.org/abs/2205.11487.
[75] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[EB/OL]. (2022-04-13)[2023-08-17]. https://arxiv.org/abs/2112.10752.
[76] FENG Z, ZHANG Z, YU X, et al. ERNIE-ViLG 2.0: improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts[EB/OL]. (2023-03-28)[2023-05-18]. https://arxiv.org/abs/2210.15257.
[77] CHEN W, HU H, SAHARIA C, et al. Re-Imagen: retrieval-augmented text-to-image generator[EB/OL]. (2022-11-22)[2023-08-09]. https://arxiv.org/abs/2209.14491.
[78] ZHENG H, HE P, CHEN W, et al. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders [C]//Proceedings of the 11th International Conference on Learning Representations, Kigali, May 1-5, 2023: 1-28.
[79] LI R, LI W, YANG Y, et al. Swinv2-Imagen: hierarchical vision transformer diffusion models for text-to-image generation[EB/OL]. (2022-10-18)[2023-08-14]. https://arxiv.org/abs/2210.09549.
[80] XIA W, YANG Y, XUE J H, et al. Towards open-world text-guided face image generation and manipulation[EB/OL]. (2021-04-18)[2023-08-20]. https://arxiv.org/abs/2104.08910.
[81] XIA W, YANG Y, XUE J H, et al. TediGAN: text-guided diverse face image generation and manipulation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 1-13.
[82] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33, Dec 6-12, 2020: 1877-1901.
[83] OPENAI. GPT-4 technical report[EB/OL]. (2023-12-19)[2023-12-23]. https://arxiv.org/abs/2303.08774.

张泽宇(2000—),男,陕西宝鸡人,硕士研究生,CCF学生会员,主要研究方向为AI绘画。
ZHANG Zeyu, born in 2000, M.S. candidate, CCF student member. His research interest is AI painting.

王铁君(1981—),女,甘肃陇西人,博士,教授,博士生导师,CCF会员,主要研究方向为知识图谱、自然语言处理、图像识别。
WANG Tiejun, born in 1981, Ph.D., professor, Ph.D. supervisor, CCF member. Her research interests include knowledge graph, natural language processing and image recognition.

郭晓然(1981—),女,河北藁城人,博士,副教授,CCF会员,主要研究方向为知识图谱、知识抽取。
GUO Xiaoran, born in 1981, Ph.D., associate professor, CCF member. Her research interests include knowledge graph and knowledge extraction.



龙智磊(1999—),男,湖南凤凰人,硕士研究生,主要研究方向为目标检测。
LONG Zhilei, born in 1999, M.S. candidate. His research interest is object detection.

徐魁(1999—),男,四川广元人,硕士研究生,主要研究方向为三维视觉。
XU Kui, born in 1999, M.S. candidate. His research interest is 3D vision.