计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

文本图像生成模型研究综述

顾许磊[1)]

[1)](南京邮电大学 计算机学院、软件学院、网络空间安全学院, 南京 210003)

摘 要 文本图像生成是将文本描述转化为语义上与其内容一致的图像。随着自然语言处理和深度学习技术的迅速发展，人工智能生成内容逐渐成为热门研究话题。本文详细探讨了扩散模型在文本图像生成任务中的应用，并通过与生成对抗网络（GAN）和自回归模型的对比分析，揭示了扩散模型的独特优势与局限性。同时，本文基于CUB数据集对不同生成模型进行了实验对比，实验结果表明，扩散模型在零样本生成方面表现出色，且在根据复杂文本提示生成高质量图像时展现了其强大的潜力。最后，本文总结了扩散模型在文本图像生成任务中所面临的挑战及未来的发展趋势，为研究者深入推进该领域的探索提供了重要参考。

关键词 文本图像生成；扩散模型；生成对抗网络；自回归模型
中图法分类号 TP39

A Review of Research on Text-to-Image Generation Models

Xu Lei-Gu[1)]

[1)](Nanjing University of Posts and Telecommunications School of Computer Science, Nanjing 210003)

Abstract Text-to-image generation is the process of converting textual descriptions into images that are semantically consistent with the content. With the rapid development of natural language processing and deep learning technologies, AI-generated content has become a popular research topic. This paper provides a detailed exploration of the application of diffusion models in text-to-image generation tasks, and through a comparative analysis with Generative Adversarial Networks (GANs) and autoregressive models, reveals the unique advantages and limitations of diffusion models. Additionally, the paper conducts experimental comparisons of different generative models based on the CUB dataset. The experimental results show that diffusion models excel in zero-shot generation and demonstrate strong potential in generating high-quality images from complex text prompts. Finally, the paper summarizes the challenges faced by diffusion models in text-to-image generation tasks and discusses future development trends, providing valuable references for researchers to further advance their exploration in this field.

Keywords text-to-Image generation; diffusion models; generative adversarial networks; autoregressive models

1 引言

随着人工智能技术的迅猛发展，人工智能生成内容（Artificial Intelligence Generated Content，AIGC）已成为当今热门话题。AIGC依托深度学习技术，能够生成文章、图像、视频、代码和语音等多种形式的内容。在文本图像生成领域，稳定扩散模型（Stable Diffusion，SD）自2022年问世以来，迅速引起了学术界和产业界的广泛关注。这项技术的实际应用，不仅能够缩短生产周期、降低人工成本、提升工作效率，还为各行业提供了灵活、高效且富有创新性的解决方案，展现出巨大的发展潜力。

近年来，随着多模态学习的兴起，文本图像生成（Text-to-Image，T2I）逐渐成为人工智能研究的热点领域之一。文本图像生成任务旨的是在给定文本描述c的条件下，生成与文本内容一致的图像$x = f(c)$，实现文本到图像的高质量和语义一致性生成。文本图像生成主要分为传统方法和基于深度学习的方法。传统方法通常依赖于图像检索[1][2]，通过在数据库中搜索并选择最符合给定文本描述的图像作为生成结果，然而这种方法的局限性在于其生成结果受限于数据库中已有的图像，无法生成新的内容。随着人工智能和计算机视觉技术的快速发展，基于深度学习的文本图像生成方法逐渐成为主流。为了生成视觉上真实且语义准确的图像，研究者提出了多种模型和方法，其中主要的深度学习方法有生成对抗网络（Generative Adversarial Networks，GAN[3]）、自回归模型（Autoregressive Models，AR）和扩散模型（Diffusion Model，DM[4]），这些方法通过训练模型学习文本与图像之间复杂的映射关系，实现了更灵活的图像生成，不仅能够生成数据集中未出现过的全新视觉内容，还极大地推动了该领域的快速发展。

文本图像生成领域中基于生成对抗网络的方法[5][6][7]已经得到了详尽的梳理和分类，本文将主要着重介绍新兴的扩散模型在文本图像生成任务上面临的重要挑战和瓶颈，当前大部分综述[8][9][10]仅提供了扩散模型在文本图像生成任务上的粗略简介，缺乏对模型内部机制、优化策略以及实际应用挑战的深入剖析，其独特的优势和应用前景尚未得到充分揭示和探讨。因此，对基于扩散模型的T2I任务进行全面而深入的综述研究，对于推动该领域的发展具有重要意义。

本综述将专注介绍扩散模型在文本图像生成任务上的发展，并通过其与生成对抗网络及自回归模型的对比，深入的揭示扩散模型方法的优势和局限。本文首先在第1章分析文本图像生成领域研究现状，接下来在第2章对扩散模型、GAN和自回归模型的原理进行了详细介绍，第3章在通用CUB数据集对不同生成模型进行对比分析，进一步验证了扩散模型的有效性。在第4章对扩散模型在文本图像生成任务上的现状进行了总结，并对其未来发展进行了展望。

2 文本图像生成模型

生成模型是一类深度学习模型[11]，旨在学习数据的概率分布并生成新的数据样本。这类模型已在自然语言处理、计算机视觉等多个领域得到了广泛应用，特别是在文本图像生成任务中，扩散模型、生成对抗网络和自回归模型成为研究的重点，均表现出良好的应用前景并持续受到关注。

2.1 基础的文本图像生成方法

随着扩散模型的快速发展，其在文本图像生成任务上展现出了强大的潜力。研究者们根据架构提出了多种基于扩散模型的文本图像生成方法，主要有基于级联的扩散模型、基于unCLIP先验的扩散模型、基于离散空间的扩散模型和基于潜在空间的扩散模型。

基于级联的扩散模型主要通过引导策略来生成高分辨率图像。如Nichol等人[12]提出的GLIDE模型采用级联扩散的方法，首先，将文本条件输入到扩散模型中生成低分辨率的图像以捕捉图像整体结构内容，然后，通过上采样扩散模型生成高分辨率的图像，提升图像细节，并保持多样性和真实性。

基于unCLIP先验的文本图像生成通过先验模型将文本信息转化为图像嵌入，利用扩散模型将图像嵌入作为条件生成图像，显著增强了文本和图像之间的一致性。如Ramesh等人[13]提出的DALLE2，首先训练CLIP模型，使其能够编码图像文本对；然后训练一个扩散先验模型，可以将文本嵌入转化为图像嵌入，最后使用扩散模型作为解码器生成以图像嵌入为条件的图像。

基于离散空间的扩散模型，主要通过将图像和文本信息编码到离散空间中，利用离散扩散模型实现文本到图像生成，不仅可以提高生成效率，还可以增加生成过程的可控性和可解释性。如Gu等人[14]提出的两阶段生成模型VQ-Diffusion，可以解决自回归模型中存在的单向偏差问题，但有时仍然会产生低质量的样本或与文本输入相关性差的图像。

2.2 扩散模型

扩散模型是一种基于马尔科夫链的生成模型，

主要包括两个过程：前向扩散过程和反向扩散过程。前向扩散过程逐步将噪声添加到真实数据中，而反向扩散过程则学习如何逐步去除噪声，从而生成新的数据样本。最终，扩散模型通过从噪声中采样生成数据，具有较好的可解释性和多样性。根据是否有生成条件，扩散模型可以分为去噪扩散概率模型和条件扩散概率模型，其中后者通常用于结合额外信息生成与条件相关的数据。去噪扩散模型根据输入数据的特性，分为连续空间和离散空间下的去噪扩散概率模型。连续空间下的去噪扩散概率模型最初由Sohl-Dickstein[15]提出，主要用于处理图像、音频和视频等连续数据。在这种模型中，如图1所示，通过前向过程$q(x_t \mid x_{t-1})$不断向图像中添加噪声。当每一步添加的噪声幅度足够小时，后验概率$q(x_{t-1} \mid x_t)$会近似为对角高斯分布。因此，通过学习一个UNet模型$p_\theta(x_{t-1} \mid x_t)$，可以近似真实的后验概率分布，从而实现逐步去噪的过程。而离散空间下的扩散模型则专门用于处理离散数据，如文本和离散信号。在这种模型中，通过引入状态转移矩阵，并利用邻近关系传播信息，模型能够直接捕捉离散数据中的结构依赖关系，从而完成前向加噪和反向去噪的过程。这一方法能够更好地建模离散数据的特性，实现有效的生成和恢复任务。

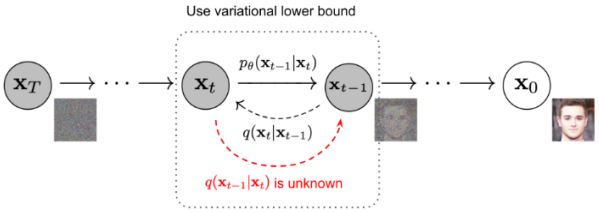

图1　扩散模型原理

2.3　生成对抗网络

在文本到图像生成任务中，生成对抗网络是一种关键的生成模型，由生成器和判别器组成。通过生成器与判别器之间的对抗训练，模型不断提升生成图像的质量和真实性。最早由Reed[16]等人提出的基于GAN的文本图像生成模型，将DCGAN[17]作为网络骨干，并引入文本特征作为生成器和判别器的约束条件，显著提高了图像生成的相关性和多样性。这一工作为文本图像生成领域奠定了重要基础。

基于GAN的文本图像生成的核心在于生成器和判别器之间的对抗训练过程。如图2所示，生成器以随机噪声和文本描述为输入，逐步学习如何生成细节丰富、语义精确且贴近真实的图像。与此同时，判别器的任务是评估生成的图像是否符合文本描述，并判断其为真实图像还是生成图像。通过这种动态博弈式的对抗训练，生成器的图像生成能力得以不断优化，而判别器在评估图像真实性和语义匹配度方面的辨别能力也随之提升，从而实现高质量的文本到图像生成效果。

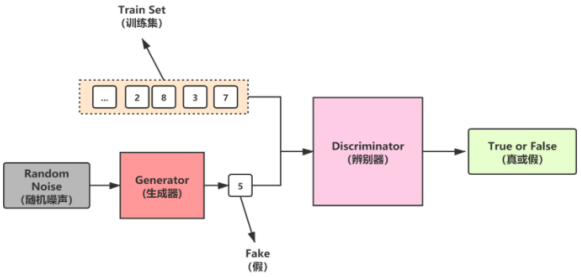

图2　生成对抗网络原理

2.4　自回归模型

自回归模型通过利用过去的数据点来预测新的数据点，受到生成式预训练Transformer模型（Generative Pre-trained Transformer，GPT）[18]在自然语言处理领域取得巨大成功的启发，逐渐被引入到图像生成任务中。该方法通过将图像数据转化为一维序列，并结合Transformer[19]架构，对图像进行自回归建模。通过这种方式，模型能够逐步生成图像的每个像素或块，从而实现高质量的图像生成，为文本到图像生成任务提供了新的解决方案。

基于自回归模型的DALLE在文本图像生成任务中取得了令人瞩目的成果。如图3所示，DALLE[20]首先利用dVAE将高分辨率图像编码为离散的图像tokens，同时采用BPE[21]将文本编码为离散的文本tokens。随后，通过Transformer模型在矢量空间中进行自回归建模，有效解决了直接在高分辨率像素空间中应用Transformer时序列长度过长的问题。在推理阶段，文本编码被输入到Transformer模型中以生成图像编码，接着通过dVAE解码器将图像编码转换为最终的生成图像。

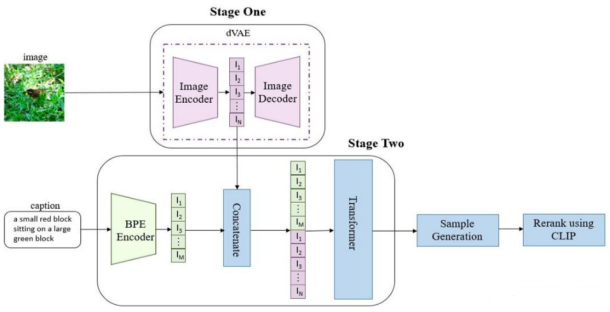

图3　自回归模型DALLE原理

3 生成模型的性能评估

本章首先介绍了文本生成图像任务中常用的评价指标及数据集，旨在为研究提供广泛的覆盖范围和可靠的评估基准。随后，针对图像生成的部分典型算法，在CUB数据集上的性能进行了详细的对比分析，以全面展示不同方法的生成效果和适用性。

3.1 数据集

CUB[22]鸟类数据集由加州理工学院在2010年提出的细粒度数据集，也是目前细粒度分类识别研究的基准图像数据集。该数据集共有11788张鸟类图像，包含200类鸟类子类，其中训练数据集有5994张图像，测试集有5794张图像，每张图像均提供了图像类标记信息，图像中鸟的bounding box，鸟的关键part信息，以及鸟类的属性信息。

3.2 评价指标

IS（Inception Score）[23]是一种常用的指标，用于评估生成图像的质量和多样性。该指标通常借助在ImageNet数据集上训练的Inception-v3图像分类器，对生成的图像进行分类评估。具体而言，IS通过分析生成图像的类别概率分布，利用信息熵来衡量图像的多样性：如果生成图像的类别分布具有较高的置信度且跨类别的分布较为均匀，则说明生成图像既具有高质量又具有丰富的多样性。IS分数越高，表示生成的图像质量越好，多样性越高。IS计算如式（1）所示：

$$IS = \exp\left(E_{x\sim p_{\text{data}}}\left[KL\left(p\left(y \mid x'\right)||p(y)\right)\right]\right) \quad (1)$$

其中x'表示生成的图像，y表示预训练的Inception-v3网络预测的类标签。

FID（Fréchet Inception Distance）[24]也是一种用来衡量生成图像质量和多样性的指标，它通过将Inception-v3模型用作特征提取器，计算生成图像和真实图像在特征空间中的Fréchet距离来评估图像质量。FID的值衡量了生成图像分布与真实图像分布之间的相似性，较低的FID值表示生成的图像更接近真实图像。FID计算如式（2）所示：

$$FID(x,x') = \left\|\mu_x - \mu_{x'}\right\|^2 + \operatorname{trace}\left(\Sigma_x + \Sigma_{x'} - 2\left(\Sigma_x\Sigma_{x'}\right)^{\frac{1}{2}}\right) \quad (2)$$

其中x，x'分别为真实图像和生成图像特征表示，μ_x为真实图像特征的均值，μ'_x为生成图像特征的均值，Σ_x为真实图像特征的协方差矩阵，Σ'_x为生成图像特征的协方差矩阵，trace为矩阵的迹。

除了以上两种常用的评估指标之外，还有SSIM、R-precision、CLIP等指标评估补充图像相似性和文本匹配度。此外，人工评估也是一种重要的补充方式，依据人类主观感受对生成图像的真实感以及与文本描述的一致性进行判断。这些指标与人工评估相结合，从不同角度确保了对生成模型性能的全面性和准确性的评价。

3.3 CUB数据集评价指标对比

为了系统地量化生成对抗网络和扩散模型在文本图像生成任务中的性能差异及其发展趋势，本文对部分典型算法在CUB数据集上的性能进行了整理和分析，如表1所示。通过对比这些算法的关键指标，可以更加直观地了解两种模型在生成质量、多样性以及文本匹配度等方面的表现，为进一步的研究提供参考依据。

从表1中的数据分析可以看出，无论是生成对抗网络还是扩散模型，在CUB数据集上的性能均取得了显著进展。这表明两类模型在文本图像生成任务中正不断深化对文本语义的理解，并逐步提升对图像生成过程的精细化控制。GAN通过对抗训练机制优化生成质量，而扩散模型则凭借逐步去噪策略实现了更高的图像细节还原与文本匹配度。两者的持续改进为文本到图像生成领域的发展提供了强大的技术支撑。

其中，Swinv2-Imagen通过结合预训练的大语言模型进行文本编码，并引入交叉注意力机制，实现了高质量的文本图像生成与文图一致性对齐。该方法在CUB数据集上的表现尤为突出，其IS高达8.44，FID值为9.78。这一结果充分展示了预训练大语言模型在文本编码与语义理解方面的优势，以及其对提升生成图像质量和语义一致性的关键作用。Swinv2-Imagen的成功进一步证明了跨模态建模中语言与视觉深度结合的潜力。

表1 部分方法在CUB 数据集上的评估结果

年份	模型	类型	IS 值	FID 值
2022	DF-GAN[25]	GAN	5.10	14.81
2022	LAFITE[26]	GAN	5.97	10.48
2023	GALIP[27]	GAN	-	10.08
2022	VQ-Diffusion[28]	DM	-	10.32
2023	Swinv2-Imagen[29]	DM	8.44	9.78
2023	ShiftDDPMs[30]	DM	4.42	14.26

4 文本图像生成相关领域研究现状

4.1 文本图像编辑

扩散模型通过逐步向图像添加噪声来破坏初始数据分布，并通过学习预测噪声和执行去噪过程，最终获得生成图像的概率分布。模型通过多次迭代，每次在输入噪声的基础上进行学习，从而逐步生成新图像。

DALL-E2[31]通过预测图像特征来实现扩散模型的生成过程，展现了强大的图像生成能力。Imagen 利用冻结的文本编码器提取文本特征，将其输入到图像扩散模型中以生成初始分辨率为64×64 的图像，随后通过两个超分辨率扩散模型逐步提升图像分辨率，最终生成1 024×1 024 的高质量图像。

ERNIE-ViLG2.0[32]是中文领域首个文本到图像生成模型，通过在训练过程中注重图像元素与文本语义的对齐，提升了模型对对象和属性关系的理解与生成能力，使其能够更准确地生成符合复杂文本描述的图像。这些方法进一步拓展了扩散模型的应用领域，并推动了文本图像生成技术的不断进步。

4.2 视频内容生成

2022年以来，扩散模型的快速发展推动了视频内容生成方法的突破性进展。Ho 等人[33]于2022 年提出了Video Diffusion Model，这是视频生成领域的一项重要创新。该方法采用3D U-Net 对视频数据中的时空噪声进行建模，将扩散模型成功扩展至视频生成任务，并探索了图像与视频数据的联合训练。该模型不仅是标准图像扩散架构的一种扩展，还通过有效降低小批量梯度的方差，加速了整体优化过程，显著提升了视频生成的效率和质量。

同年，Ho 等人[34]进一步提出了基于文本条件的视频生成系统Imagen Video。该系统基于级联视频扩散模型构建，可从文本描述中生成高清视频内容。Imagen Video 引入了渐进式蒸馏技术，并结合无分类器指导的方法，显著提高了视频采样的速度和生成质量。这些研究成果不仅展示了扩散模型在视频生成领域的广阔潜力，也为基于文本条件的视频内容生成提供了高效且创新的解决方案。

4.3 医学图像生成

随着扩散模型的快速发展及持续改进，其在图像生成领域的表现日益突出，展现了强大的功能与广阔的应用潜力。在医学图像生成领域，扩散模型也取得了令人瞩目的成果。

2022 年，Boah Kim 和Jong Chul Ye提出了扩散可变形模型（Diffusion Deformable Model, DDM），用于生成具有时间维度的4D 医学图像。该模型通过适配去噪扩散概率模型，生成源体积与目标体积之间的中间时间体积，成功将扩散模型应用于医学图像的时间序列生成任务。

2023 年，Alper Gungor 等人提出了名为AdaDiff 的自适应扩散先验方法，用于加速磁共振成像重建。该方法借助扩散模型在生成多样性和先验学习上的优势，为医学图像重建任务提供了高效的解决方案，显著改善了重建速度与图像质量。

扩散模型凭借其生成多样性和灵活性，在医学图像生成中展现出巨大的潜力。这些研究不仅为医学影像处理带来了新的可能性，还为解决复杂的医学图像生成与重建问题提供了创新的技术路径。尽管面临计算效率、模型复杂度等挑战，但随着研究的深入和技术的进步，扩散模型在医学图像生成领域的应用前景无疑十分广阔。

5 总结与展望

5.1 本文总结

在当前基于人工智能技术生成内容的背景下，生成模型受到广泛关注，而作为一种新兴的生成模型，扩散模型展现了巨大的发展潜力。本文首先介绍了文本图像生成方法的基本理论框架，并分别阐述了三种生成式模型的理论。接着，通过在CUB数据集上的实验分析，强调了扩散模型相较于生成对抗网络在实现简便性和训练稳定性方面的显著优势。此外，本文还介绍了其他生产类任务中扩散模型的应用，包括文本图像编辑、视频内容生成、医学图像生成。最后，本文展望了未来的研究方向，讨论了扩散模型在文本图像生成任务中的进一步优化和应用前景，为后续研究提供了有价值的参考。

5.2 未来工作展望

尽管扩散模型在文本图像生成实验中取得了显著成果，但问题和挑战依然存在，未来的研究工作可以在以下几个方向进行深入探索：

（1）基础理论不够成熟。当前的扩散模型更多地依赖于实践中的经验，而缺乏深入的理论分析。未来的研究可能会深入探讨扩散过程的数学原理，寻找更加通用的理论框架来解释模型的表现。此外，随着扩散模型的广泛应用，其可解释性也变得更加重要，未来工作应逐渐完善扩散模型的基础理论。

（2）评价方式不统一。当前文本到图像生成方法的评估主要依赖于特定的指标或人工评估，但这些方法存在一些局限性。例如，FID值虽然广泛应用于图像生成质量的评估，但并不总能准确反映图像的感知质量；而人工评估受限于评估者的主观审美差异，且效率较低。此外，不同研究通常采用各自独特的基准和文本提示生成图像，可能导致评估结果的不可比性，从而增加了对复杂场景进行公正评估的难度。针对上述问题，作者建议构建一个统一的评估框架，该框架应明确且多样化地涵盖各种评估指标，全面且客观地衡量文本到图像生成模型的性能。通过这种统一的评估框架，不同研究之间可以进行公平的比较，从而促进模型通用性和泛

化能力的提升，为该领域的发展提供更加稳固的基准支持。

参 考 文 献

[1] ALKHAWLANI M, ELMOGY M, EL BAKRY H. Text-based, content-based, and semantic-based image retrievals: a survey[J]. International Journal of Computer and Information Technology, 2015, 4(01): 58-66.

[2] LI W, DUAN L X, XU D, et al. Text-based image retrieval using progressive multi-instance learning[C]// Proceedings of the 2011 International Conference on Computer Vision. IEEE, 2011: 2049-2055.

[3] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]. Advances in Neural Information Processing Systems, 2014, 27.

[4] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[5] 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.

[6] 作者. 王威,李玉洁,郭富林,等.生成对抗网络及其文本图像合成综述[J].计算机工程与应用, 2022, 58(19):14-36.

[7] 陈佛计,朱枫,吴清潇,等.生成对抗网络及其在图像生成中的应用研究综述[J].计算机学报, 2021, 44(02):347-369.

[8] CROITORU F A, HONDRU V, IONESCU R T, et al. Diffusion models in vision: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[9] YANG L, ZHANG Z L, SONG Y, et al. Diffusion models: a comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2023, 56(4): 1-39.

[10] ZHAN F N, YU Y C, WU R L, et al. Multimodal image synthesis and editing: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[11] 胡铭菲,左信,刘建伟.深度生成模型综述[J].自动化学报,2022,48(01):40-74.DOI:10.16383/j.aas.c190866.

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

[13] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv:2204.06125, 2022, 1(2): 3.

[14] GU S, CHEN D, BAO J, et al. Vector quantized diffusion model for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10696-10706.

[15] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Procedings of the International Conference on Machine Learning. PMLR, 2015: 2256-2265.

[16] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//Procedings of the International Conference on Machine Learning. PMLR, 2016: 1060-1069.

[17] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv: 1511.06434, 2015.

[18] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[C]//Advances in Neural Information Processing Systems, 2018, 31: 8735-8745.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017, 30.

[20] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning. PMLR, 2021: 8821-8831.

[21] GAGE P. A new algorithm for data compression[J]. C Users Journal, 1994, 12(2): 23-38.

[22] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset[D]. California Institute of Technology, 2011:1-8.

[23] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[C]. Advances in Neural Information Processing Systems, 2016, 29.

[24] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]. Advances in Neural Information Processing Systems, 2017, 30.

[25] TAO M, TANG H, WU F, et al. Df-gan: a simple and effective baseline for text-to-image synthesis[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16515-16525.

[26] ZHOU Y, ZHANG R, CHEN C, et al. Lafite: towards language-free training for text-to-image generation[J]. arXiv: 2111.13792, 2021, 2.

[27] TAO M, BAO B K, TANG H, et al. GALIP: generative adversarial CLIPs for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14214-14223.

[28] GU S, CHEN D, BAO J, et al. Vector quantized diffusion model for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10696-10706.

[29] LI R, LI W, YANG Y, et al. Swinv2-imagen: hierarchical vision transformer diffusion models for text-to-image generation[J]. Neural Computing and Applications, 2023: 1-16.

[30] ZHANG Z, ZHAO Z, YU J, et al. ShiftDDPMs: exploring conditional diffusion models by shifting diffusion trajectories[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(3): 3552-3560.

[31] RAMESH A,DHARIWAL P,NICHOL A,et al.Hierarchical text-conditional image generation with CLIP latents[EB/OL]..https://arxiv.org/abs/2204.06125.

[32] FENG Z D,ZHANG Z Y,YU X T,et al.ERNIE-ViLG 2.0:Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Vancouver,BC,Canada.IEEE,2023:10135-10145.

[33] Ho J, Salimans T, Gritsenko A, et al. Video Diffusion Models[J]. ar Xv preprint ar Xi iv:2204.03458, 2022.

[34] Ho J, Chan W, Saharia C, et al. Imagen video:High definition video generation with diffusion models[J]. ar Xiv preprint ar Xv:2210.02303, 2022.i