中图法分类号:TP309.7; TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)06-1535-20

论文引用格式:Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(06):1535-1554(刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(06):1535-1554)[DOI:10.11834/jig.240003]

AIGC视觉内容生成与溯源研究进展

刘安安[1],苏育挺[1*],王岚君[1],李斌[2],钱振兴[3],张卫明[4],周琳娜[5],
张新鹏[3],张勇东[4],黄继武[2],俞能海[6]

1. 天津大学电气自动化与信息工程学院,天津　300072; 2. 深圳大学电子信息与工程学院,深圳　518060;
3. 复旦大学计算机科学技术学院,上海　200438; 4. 中国科学技术大学信息科学技术学院,合肥　230026;
5. 北京邮电大学网络空间安全学院,北京　100876; 6. 中国科学技术大学网络空间安全学院,合肥　230027

摘　要: 随着数字媒体与创意产业的快速发展,人工智能生成内容(artificial intelligence generated content, AIGC)技术以其在视觉内容生成中的创新应用而逐渐受到关注。本文旨在围绕AIGC视觉内容生成与溯源研究进展深入研讨。首先,针对图像生成技术进行探讨,从基于生成式对抗网络的传统方法出发,系统地分析了基于生成式对抗网络、自回归模型和扩散概率模型的最新进展。接着,深入探讨可控图像生成技术,突出了通过布局、线稿等附加信息以及基于视觉参考的方法来为创作者提供精确控制的技术现状。随着图像生成技术的革新和应用,生成图像的安全性问题逐渐浮现。而预先审核和过滤的技术手段已难以满足实际需求,故亟需实现生成内容的溯源来进行监管。因此,本文进而对生成图像溯源技术进行研讨,并聚焦水印技术在确保生成内容可靠性和安全性方面的应用。依据水印嵌入的流程节点,首先将现有的水印相关的生成图像溯源方法归为无水印嵌入的生成图像溯源、水印前置嵌入的生成图像溯源、水印后置嵌入的生成图像溯源以及联合生成的生成图像溯源并进行详细分析,然后介绍针对生成图像的水印攻击研究现状,最后对生成图像溯源技术进行总结和展望。鉴于视觉内容生成在质量和安全上的挑战,旨在为研究者提供一个视觉内容生成与溯源的系统研究视角,以促进数字媒体创作环境的安全与可信,并引导未来相关技术的发展方向。

关键词: 人工智能内容生成(AIGC);视觉内容生成;可控图像生成;生成内容安全;生成图像溯源

Review on the progress of the AIGC visual content generation and traceability

Liu Anan[1], Su Yuting[1*], Wang Lanjun[1], Li Bin[2], Qian Zhenxing[3], Zhang Weiming[4],
Zhou Linna[5], Zhang Xinpeng[3], Zhang Yongdong[4], Huang Jiwu[2], Yu Nenghai[6]

1. School of Electrical and Information Engineering, Tianjin University, Tianjin　300072, China;
2. College of Electronics and Information Engineering, Shenzhen University, Shenzhen　518060, China;
3. School of Computer Science, Fudan University, Shanghai　200438, China;
4. School of Information Science and Technology, University of Science and Technology of China, Hefei　230026, China;
5. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing　100876, China;
6. School of Cyber Science and Technology, University of Science and Technology of China, Hefei　230027, China

Abstract: In the contemporary digital era, which is characterized by rapid technological advancements, multimedia con-

收稿日期:2024-01-01;**修回日期**:2024-03-14;**预印本日期**:2024-03-21

*通信作者:苏育挺　ytsu@tju.edu.cn

基金项目:国家自然科学基金项目(U21B2024,U20B2047,U2336206,U20B2051,U23B2022,62371330,62202329,62172053)

Supported by: National Natural Science Foundation of China(U21B2024,U20B2047,U2336206,U20B2051,U23B2022,62371330, 62202329,62172053)

tent creation, particularly in visual content generation, has become an integral part of modern societal development. The exponential growth of digital media and the creative industry has attracted attention to artificial intelligence generated content (AIGC) technology. The groundbreaking applications of AIGC in visual content generation not only have equipped multimedia creators with novel tools and capabilities but also have delivered substantial benefits across diverse domains, which span from the realms of cinema and gaming to the immersive landscapes of virtual reality. This review comprehensive introduces the profound advancements within AIGC technology. Our particular emphasis is on the domain of visual content generation and its critical facet of traceability. Initially, our discussions trace the evolutionary path of image generation technology, from its inception within generative adversarial networks (GANs) to the latest advancements in Transformer auto-regressive models and diffusion probability models. This progression unveils a remarkable leap in the quality and capability of image generation, which underscores the rapid evolution of this field. This evolution has transitioned from its nascent stages to an era characterized by explosive growth. First, we delve into the development of GANs, encompassing their evolution from text-conditioned methods to sophisticated techniques for style control and the development of large-scale models. This type of technology pioneered the text-to-image generation. GANs can further improve their performance by expanding network parameters and dataset size due to their strong scalability. Furthermore, we explore the emergence of Transformer-based auto-regressive models, such as DALL·E and CogView, which have heralded a new epoch in the domain of image generation. The basic strategy of autoregressive models is to first use the Transformer structure to predict the feature sequence of images based on other conditional feature sequences such as text and sketches. Then, it uses a specially trained decoding network to decode these feature sequences into a complete image. They can generate realistic images based on the large-scale parameters. In addition, our discourse delves into the burgeoning interest surrounding diffusion probability models, which are renowned for their stable training methods and their ability to yield high-quality outputs. The diffusion models first adopt an iterative and random process to simulate the gradual transformation of the observed data into a known noise distribution. Then, they reconstruct the original data in the opposite direction from the noise distribution. This random process based on stochastic approach provides a more stable training process, while it also demonstrates impressive results in terms of generated quality and diversity. As the development of AIGC technology continues to advance, it encounters challenges, such as the enhancement in content quality and the need of precise control to align with specific requisites. Within this context, this review conducts a thorough exploration of controllable image generation technology, which is a pivotal research domain that strives to furnish meticulous control over the generated content. This achievement is facilitated through the integration of supplementary elements, such as intricate layouts, detailed sketches, and precise visual references. This approach empowers creators to preserve their artistic autonomy while upholding exacting standards of quality. One notable facet that has garnered considerable academic attention is the utilization of visual references as a mechanism to enable the generation of diverse styles and personalized outcomes by incorporating user-provided visual elements. This review underscores the profound potential inherent in these methodologies, which illustrates their transformative role across domains such as digital art and interactive media. The development of these technologies introduces new horizons in digital creativity. However, it presents profound challenges, particularly in the domain of image authenticity and the potential for malevolent misuse. These risks are exemplified by the creation of deep fakes or the proliferation of fake news. These challenges extend far beyond mere technical intricacies; they encompass substantial risks pertaining to individual privacy, security, and the broader societal implications of eroding public trust and social stability. In response to these formidable challenges, watermark-related image traceability technology has emerged as an indispensable solution. This technology harnesses the power of watermarking techniques to authenticate and verify AI-generated images, which safeguards their integrity. Within the pages of this review, we meticulously categorize these watermarking techniques into distinct types: watermark-free embedding, watermark pre-embedding, watermark post-embedding, and joint generation methods. First, we introduce the watermark-free embedding methods, which treat the generated traces left during model generation as fingerprints. The inherent fingerprint information is used to achieve model attribution of generated images and achieve traceability purposes. Second, the watermark pre-embedding methods aim to embed the watermark into input training data such as noise and image. Another aim is to use the embedded watermark data to train the generation model, which can also introduce traceability information in the generated image. Third, the watermark post-embedding

methods divide the process of generating watermark images into two stages: image generation and watermark embedding. Watermark embedding is performed after image generation. Finally, the joint generation methods aim to achieve adaptive embedding of watermark information during the image generation process, minimize damage to the image generation process when fusing with image features, and ultimately generate images carrying watermarks. Each of these approaches plays a pivotal role in the verification of traceability across diverse scenarios, which offers a robust defense against potential misuses of AI-generated imagery. In conclusion, while AIGC technology offers promising new opportunities in visual content creation, it simultaneously causes significant challenges regarding the security and integrity of generated content. This comprehensive review covers the breadth of AIGC technology, which starts from an overview of existing image generation technologies, such as GANs, auto-regressive models, and diffusion probability models. It then categorizes and analyzes controllable image generation technology from the perspectives of additional conditions and visual examples. In addition, the review focuses on watermark-related image traceability technology, discusses various watermark embedding techniques and the current state of watermark attacks on generated images, and provides an extensive overview and future outlook of generation image traceability technology. The aim is to offer researchers a detailed, systematic, and comprehensive perspective on the advancements in AIGC visual content generation and traceability. This study deepens the understanding of current research trends, challenges, and future directions in this rapidly evolving field.

Key words: artificial intelligence generated content (AIGC); visual content generation; controllable image generation; security of generated content; traceability of generated images

0 引　言

在数字化时代，随着科技的迅速迭代更新，多媒体内容创作无疑成为现代社会发展的核心要素之一。这样的发展趋势使得众多领域，从电影、游戏到虚拟现实，都受益于视觉内容的生成技术。其中，人工智能生成内容（artificial intelligence generated content，AIGC）技术因其在视觉内容生成中所展现出的创新性，为多媒体创作者开辟了全新的工具和可能性。

这类技术在数字内容创作领域展现出巨大潜力，然而，如何提升生成质量，以及如何精确控制生成内容使其符合创作需求，仍是当前研究者们面临的挑战。对此，可控图像生成技术作为一个前沿的研究方向，旨在通过引入布局、线稿等附加信息来对生成内容进行更为精确的控制（Mou 等，2023；Zhang 等，2023）。这种方法可以帮助创作者在保持创意自由的同时，确保生成的内容达到预期的质量。此外，基于视觉参考的图像生成技术（Gal 等，2023；Kumari 等，2023；Li 等，2023a）也受到了学术界的广泛关注，该类方法允许生成模型参考特定的视觉元素来生成不同风格的图像。这种方法的优势在于它可以将用户的参考视觉元素直接纳入生成过程中，从而获得更为满意的结果。

然而，正如每一枚硬币都有两面，技术的进步往往也伴随着新的挑战。这些针对特定主题生成的高度逼真图像可能被滥用于恶意目的，如深度伪造（Yu 等，2019）、虚假新闻制作（https://www.thepaper.cn/newsDetail_forward_22830685）等。这不仅对个人的隐私和安全构成威胁，还可能影响社会舆论的稳定性和公众对社媒信息的信任度。而此类生成图像的安全性问题，能够通过水印相关的图像溯源技术（Fernandez 等，2023；Liu 等，2023；Yu 等，2021）得到缓解，故该领域正逐渐成为研究者和工业界关注的焦点。这类技术旨在利用水印识别和验证由图像生成技术生成的图像，以确保其来源、真实性和可靠性。目前的水印技术可依据水印嵌入的流程节点，划分为无水印嵌入方法、水印前置嵌入方法、水印后置嵌入方法以及联合生成方法，能够针对不同生成场景进行图像的溯源验证，并通过水印信息对模型、用户等进行追责，能够在一定程度上缓解生成图像带来的安全性问题。

总之，AIGC技术为视觉内容创作提供新机遇的同时，也带来了生成内容安全的挑战。为了充分了解、利用和发展视觉生成式技术，本文首先从生成对抗网络、自回归模型以及扩散概率模型等技术对现有图像生成技术做出总结，并从附加条件和视觉示例的角度对可控图像生成进行技术划分。此外，为进一步确保视觉生成内容的安全性和可靠性，本文

聚焦于水印相关的图像溯源技术，并依据水印嵌入的时间节点将其划分为无水印嵌入的生成图像溯源、水印前置嵌入的生成图像溯源、水印后置嵌入的生成图像溯源以及联合生成的生成图像溯源，并进行详细分析，然后探讨针对生成图像的水印攻击研究现状，最后对生成图像溯源技术进行总结和展望。本文旨在为研究者提供一个系统的、全面的关于AIGC时代视觉内容生成与溯源研究进展的视角，期望通过本文的阐述，研究者能更深入地了解当前领域的研究现状、面临的挑战以及未来的研究趋势。

1 图像生成技术

图像生成技术是计算机视觉领域备受关注的研究方向，具有广泛的应用潜力。本节将深入研究图像生成技术的演进，从回顾传统的基于生成对抗网络（generative adversarial network，GAN）（Hu 等，2018）的图像生成方法开始，逐步介绍当前备受瞩目的图像生成大模型，系统地探讨基于 GAN、Transformer（Vaswani 等，2017）自回归模型以及扩散概率模型（Ho 等，2020；Nichol 等，2022；Sohl-Dickstein 等，2015）的最新进展，以便更好地理解图像生成技术从最初的发展期到如今的图像生成技术在生成质量、生成能力都迅速发展的爆发期之间的技术演进和当前研究的前沿趋势（如图 1）。在 GAN 的领域，从最初的文本条件驱动方法到后来的风格控制和大型预训练模型，都体现了技术的日益成熟和提升。基于 Transformer 的自回归模型也为图像生成提供了新的视角，其中，DALL·E（Ramesh 等，2021）和 CogView（Ding 等，2021）等模型取得了令人瞩目的成果。最后，扩散概率模型作为一个新的研究方向，凭借其稳定的训练方法和高质量的生成结果，已经吸引了大量的研究兴趣。总的来说，图像生成技术在多个方向上都取得了显著的进展，预示着更为广泛的应用前景。

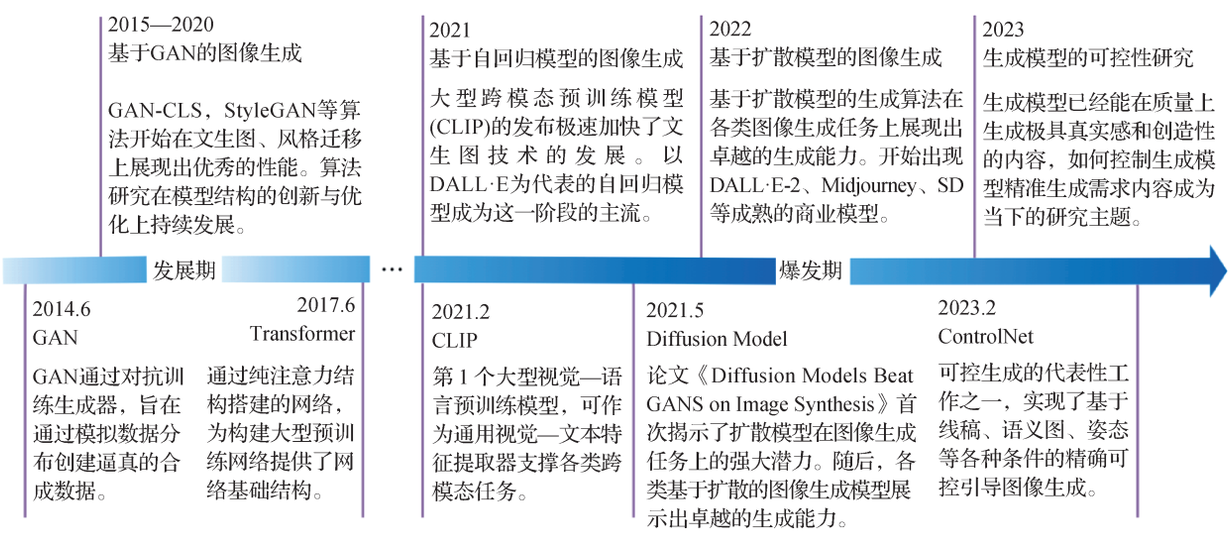

图 1 图像生成技术发展阶段

Fig. 1 The development stage of image generation technology

1.1 基于生成对抗网络的图像生成技术

基于深度学习的图像生成技术受到科研人员的广泛关注最早可以追溯到生成式对抗网络（GAN）（Goodfellow 等，2014）的出现，GAN 的核心思想是通过让两个神经网络相互对抗，即一个生成器网络和一个判别器网络，来实现生成具有真实感图像的效果。在早期的图像生成研究中，Reed 等人（2016b）率先引入了文本条件驱动图像生成的 GAN-INT-CLS 方法。这一方法核心在于使用文本嵌入技术，将描述性的文本编码成向量，并与生成器网络融合。如此，生成器便可以依据具体的文本描述，创造出相应的图像。为了进一步增强图像的细节与文本描述之间的一致性，Xu 等人（2018）后续推出了 AttnGAN。其独到之处在于整合了注意力机制，使得在图像的不同子区域中可以更为精准地反映文本的描述细节，从而产生更为高质量和高分辨率的图像。随后，Karras 等人（2019）再度推进了这一领域的研究，提出了 StyleGAN。不同于前者，StyleGAN 的亮点在于

其对风格的深度控制。通过替代传统的噪声向量输入，StyleGAN采用“风格适应”技术，在网络的多个层次中加入风格信息，从而实现了对图像生成中的细节进行深度控制，既保证了图像的高分辨率和高质量，又增添了对图像风格、内容的灵活调控。

虽然研究人员已尝试通过更新生成架构和算法来不断提升基于GAN的图像生成方法（Qiao等，2019；Reed等，2016a；Zhang等，2017）的性能，然而，受到数据集大小的限制和文本编码器性能的制约，这些方法仍然无法生成复杂且具备真实感的图像。近年来，随着模型参数的扩增和跨模态预训练模型的迅速发展，基于GAN的图像生成模型的性能已有显著提升。由Tao等人（2023）提出的GALIP（generative adversiral CLIPs）使用CLIP（contrastive language-image pre-training）模型设计生成器和判别器的损失函数，实现了GAN特征空间与预训练的CLIP对齐以引导视觉概念的生成。其不仅提高了训练效率，而且在较低的计算成本下达到了与大型预训练模型相似的性能。此外，扩大模型大小和使用更大的训练数据集也已逐渐成为提高性能的关键策略。例如，GigaGAN（Kang等，2023）通过增大模型参数至10亿级别，并结合像LAION-2B（Schuhmann等，2022）这种拥有20亿幅图像的庞大数据集进行训练，成功使其生成性能可与最先进的自回归模型或基于扩散概率模型的生成方法相抗衡。

1.2 基于自回归模型的图像生成技术

Transformer（Vaswani等，2017）模型以其独特的自注意力机制和能力强大的表示学习为特点，已在自然语言处理和计算机视觉领域得到了广泛应用并实现了一系列突破性应用。在跨模态图像生成任务中，基于Transformer的自回归模型也展现出了优越的性能。自回归模型的基本策略是利用自回归的Transformer结构来从文本、草图等其他条件的特征序列的基础上预测图像的特征序列，然后利用专门训练的解码网络将这些特征序列解码为完整的图像。

OpenAI发布的DALL·E（Ramesh等，2021）是这类方法中最初的代表性工作，其核心思想是利用大型的Transformer模型直接从文本描述生成图像。为了实现这一目标，DALL·E首先使用一个离散变分自编码器（discrete variational autoencoder，dVAE）（Rolfe，2017）将图像编码为标记序列，再结合字节对编码（byte-pair encoding，BPE）（Sennrich等，2016）技术对文本进行编码。通过这种方式，DALL·E将文本和图像标记结合，然后利用自回归Transformer模型学习文本到图像的联合分布。这种直接的生成策略使DALL·E能够根据各种文本描述生成高质量的、具有创意的图像。

基于DALL·E的成功实践，CogView（Ding等，2021）进一步完善并优化了相关技术。它采用基于VQ-VAE（vector quantized variational autoencoder）（van den Oord等，2017）的编解码器以增强图像特征的编解码能力，并引入了一种基于规范化的Transformer训练策略以增强模型的稳定性。CogView的应用领域不仅限于图像生成，还扩展到了多种下游任务，如样式学习、超分辨率、文本—图像排序和时尚设计。在该团队的工作CogView2（Ding等，2022a）中，其进一步解决了先前工作在生成效率低、训练复杂方面的缺陷。其通过训练跨模态通用语言模型CogLM对不同的下游任务进行微调。在生成过程中，其首先生成一批低分辨率的图像，再将生成的图像映射到从预训练的CogLM微调的直接超分辨率模块中，使用局部注意力来减少训练开销。最后，通过另一个从预训练的CogLM微调的迭代超分辨率模块来细化这些高分辨率图像。CogView2能够显著加速高分辨率图像的生成，将模型运行时间从3 600 ms减少到6 ms（仅为1/600），显著加速了高分辨率图像的生成。

由Google发布的Parti（Yu等，2022b）模型是近期较为先进的基于自回归的图像生成模型，其支持复杂的概念组合生成精细的高质量图像。首先，Parti使用基于Transformer的图像分词器ViT-VQGAN（vision Transformer with vector quantized generative adversarial network）（Yu等，2022a）将图像编码为离散标记的序列。其通过另外训练文本编辑器，实现了更加强大的文本表示能力。通过将Transformer编码器—解码器模型的大小扩展到200亿参数，Parti实现了高质量的逼真图像的生成，再次给此类方法的发展带来了推动。

1.3 基于扩散模型的图像生成技术

在近年的深度学习进展中，Diffusion Model（扩散模型）（Ho等，2020；Nichol和Dhariwal，2021；Sohl-Dickstein等，2015）逐渐成为图像生成领域的一个重要研究方向。不同于传统的生成对抗网络（GAN）

(Hu 等, 2018)和变分自编码器(VAE)(Kingma 和 Welling, 2022)等框架,扩散模型采用了一个迭代、随机的过程,模拟如何将观察数据逐渐转变为某种已知的噪声分布,再反向地从这种噪声中重构原始数据。这种基于随机过程的方法为模型提供了一个更为稳定的训练过程,同时在生成的质量和多样性上也展现了令人印象深刻的效果。

为了在随机的采样过程中施加条件控制,Dhariwal 和 Nichol(2021)首先提出了分类器引导的扩散过程,该过程将扩散过程引导至类标签的方向,初步实现了一定程度上的图像生成的控制。在此基础上,Ho 和 Salimans(2022)进一步改进了分类器引导的扩散模型,引入了无分类器引导的扩散过程,使有条件的和无条件的扩散过程一起进行联合训练,同时根据比例因子调节采样。这个方法可以使得扩散模型在采样过程中仅仅依赖于扩散过程,而无需分类器模型的参与。

GLIDE(Nichol 等, 2022)是应用扩散模型进行图像生成这一领域早期的代表性工作。它采用了相对简单的方法设计实现基于文本引导的图像生成,将采样的随机高斯噪声与 CLIP 编码的文本变量一起插入到扩散模型中进行训练,并用一个基于 CLIP 特征的引导扩散损失计算特定扩散时间步上噪声图像和输入文本之间的相似性。这项损失的引入对提高文本到图像的生成性能很有作用,但它也使得生成图像的多样性降低。因此,GLIDE 也对在方法中引入无分类器引导的方法进行了研究,以在生成过程中产生更加多样的生成结果。

Google 发布的 Imagen(Saharia 等, 2022)则是在图像细节的生成能力以及复杂文本输入的理解能力表现突出的扩散图像生成模型。相比于对扩散概率模型本身的优化,该工作在优化 U-Net 架构的同时,深入探讨了扩展预训练语言模型对于引导图像生成的优势。该方法利用 T5-XXL(Raffel 等, 2020)编码器实现进行输入文本的特征映射,并通过两个超分辨率层将 64 × 64 像素图像放大至 1 024 × 1 024 像素。为提高采样质量,应用了噪声级条件化技术和级联扩散模型。同时,借鉴了无分类器引导(Ho 和 Salimans, 2022)方法来提升图像样本质量,但由于可能出现图像保真度问题,应用了阈值技术进行限制。Imagen 针对 U-Net 架构的推理时间、收敛速度和内存效率,做了多项改进,如通过常数值缩放 U-Net 跳过连接以加速收敛,优先考虑较低分辨率的模型设计以降低内存成本,并调整了下采样和上采样操作的顺序以提高推理速度。最终,Imagen 在生成性能和生成效率上得到了很好的权衡,实现了针对复杂文本输入的高质量图像生成能力。

继 DALL· E(Ramesh 等, 2021)和 GLIDE(Nichol 等, 2022)的成功实践之后,OPENAI 将扩散概率模型的理念进一步应用在 DALL·E-2(Ramesh 等, 2022)模型中。该模型主要由两大核心组件构成:先验模型和解码器模型。先验模型负责根据文本提示生成 CLIP 图像嵌入,而解码器则依据扩散概率模型,根据图像嵌入生成相应的图像。相较于前代技术,DALL·E-2 创新性的技术架构赋予了它支持多种创新生成任务的能力,包括图像的编辑与处理。例如,DALL·E-2 能够融合两图的显著特征以产生特定风格的图像生成;它能够无限扩展基于原始的图像,创造大型且复杂的组合图像;DALL·E-2 还具备对现有图像进行编辑的能力,能够依据自然语言的提示,直接修改现有图像的特定内容。

DALL·E-2 模型凭借其编解码器架构取得了令人瞩目的生成效果。与其设计理念近似,研究人员进一步探索发现,通过将图像压缩到低维潜在空间表示进行扩散概率模型的训练能够在有效降低计算复杂性的同时保持扩散模型的生成能力。在 LDM (latent diffusion model)(Rombach 等, 2022)这项工作中,研究人员使用一个预训练的自编码器将图像数据映射到一个分辨率为原始图像 1/16～1/4 的压缩空间中。这样做的初衷是将原始图像中的高频细节压缩到能够充分表达图像感知内容的潜在空间内,以保证扩散概率模型训练的高效性和稳定性。通过在潜在空间中进行采样和去噪操作,LDM 将最终生成的潜在特征通过解码器恢复为图像。

同时,LDM 还提出了一种利用特定条件信息来控制扩散过程的机制。具体而言,该机制通过特定领域的特征编码器将文本、语义图和图像等多种形式的生成条件投影到潜在表示中,然后在 U-Net 主干结构中实施交叉注意力融合,从而实现特定条件下的生成控制。这些技术创新使得 LDM 获得了令人印象深刻的生成性能。通过扩充训练数据集,该研究团队发布了一系列名为“Stable Diffusion”的基于 LDM 的预训练模型。目前,Stable Diffusion 已经成为图像生成领域中最被广泛讨论和应用的模型。

2 可控图像生成技术

可控图像生成技术是对图像生成技术的进一步深化与拓展，它允许研究者和创作者按照特定的要求和条件生成图像，为图像生成应用更多的可能性和灵活性。本节将对可控图像生成技术进行深入探索，首先，介绍基于附加条件的生成方法，这些方法主要通过引入外部条件，如布局和线稿，来实现对生成图像的精确控制。接着，转向基于视觉示例的生成技术，这些技术着重于如何利用现有的视觉信息，结合文本提示，实现个性化和定制化的图像生成。总的来说，可控图像生成技术为图像生成研究带来了更多的可能性和灵活性，同时也为未来的研究和应用开辟了新的方向。

2.1 基于附加条件的可控图像生成

大型图像生成模型已经在基于文本的图像生成中取得惊艳效果。然而，尽管这些模型在学习复杂结构和有意义的语义方面展现出强大的能力，但仅仅依赖文本提示并不能充分利用模型所学习的知识，特别是当需要灵活和准确地控制（例如颜色和结构）生成的内容时。为了解决这个问题，Zhang 等人（2023）提出了一种新的神经网络结构 ControlNet，该结构旨在为预训练的大型扩散模型添加额外的输入条件控制。图 2 展示了 ControlNet 基于草图、语义图、深度图和姿态图 4 种附加条件的生成结果。ControlNet 保留了预训练模型的参数，通过训练一个额外的与原始 U-Net 同结构的网络来向模型引入新的知识。特别地，ControlNet 使用“零卷积”结构设计保证了模型微调过程中不会受有害噪声干扰。ControlNet 的出色之处在于能够处理各种条件，将边缘、深度、分割和人体姿势的控制信息引入图像生成过程中，实现图像生成任务上空间布局的精确控制。

与 ControlNet 的目的类似，Mou 等人（2023）提出了一种简单、轻量级的附加模块 T2I-Adapter 来实现将预训练图像生成模型中的内部知识与外部控制信号进行对齐，且不需要对原始生成模型的参数进行训练。这样，可以根据不同的条件，包括线稿、语义图和姿态等各种附加条件为预训练图像生成模型训练各自的适配器，实现在生成结果的颜色和结构上进行丰富的控制和编辑。将两者进行对比，ControlNet 更侧重于通过为图像生成添加额外条件来增强预训练图像生成模型服从某种条件的能力，而 T2I-Adapter 则是为预训练的文本到图像模型提供额外的指导，以实现更准确和灵活的结构控制。

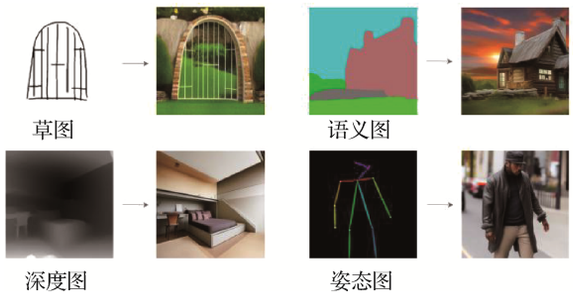

图 2 基于附加条件的可控图像生成技术示意

Fig. 2 Controllable image generation technique based on additional conditions

在通过附加条件控制图像生成这个任务上，与上两项工作关注于通过线稿、语义图控制图像的空间结构和色彩信息不同，另一种体现可控的思路是引导生成模型在特定的位置上生成特定的物体。在这方面，GLIGEN（grounded-language-to-image generation）（Li 等，2023c）技术提出了一种基于“grounding”的生成方法，能够对基于开放世界中的实体概念对生成图像的内容和布局进行控制。具体来说，GLIGEN 支持多种输入组合，例如文本实体 + 布局边框、图像实体 + 布局边框、图像风格 +文本 + 布局边框和文本实体 + 关键点等，实现在特定位置生成特定实体图像，或是组合各种布局控制条件的可控生成能力。

在最新的基于附加条件的可控性生成研究中，UniControl（Qin 等，2023）代表了扩散模型在图像生成领域的一项创新研究。它的核心设计整合多种附加条件到图像任务中，并将其与自然语言指令一起训练，从而使得生成模型能够在理解自然语言指令的基础上响应不同的基于附加条件的图像生成任务。为了使 UniControl 具备处理多样视觉条件的能力，作者扩充了预训练的文本到图像扩散模型，并引入了一个任务感知的 HyperNet，使网络适应不同的基于附加条件的图像生成任务。在 9 个独特的附加条件生成任务上进行训练后，UniControl 展示了令人印象深刻的零样本生成能力，其可以在自然语言的提示下，通过组合已训练任务的生成模式，基于未经训练过的附加条件形式进行图像生成。通过这种多

条件整合和高度控制的设计，UniControl为可控视觉生成领域带来了显著的进步，展示了更加灵活和多样的图像生成。

2.2 基于视觉示例的可控图像生成

随着文本到图像生成模型的不断进步，目前已经可以通过深度学习算法自动生成高质量、真实感强的图像。但如何在保持这些优点的同时，为生成模型引入更多的控制性和个性化元素仍是目前研究的一个热门议题。尤其是在用户只提供少量示例图像的场景下，如何有效地提取和利用这些视觉信息，将其与文本提示结合，以生成满足特定需求的图像（如图3所示），成为研究的核心挑战。

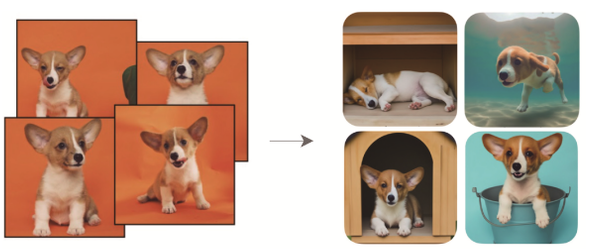

图3 基于视觉示例的可控图像生成技术示意

Fig. 3 Controllable image generation technique based on visual examples

Gal等人（2022）在研究中首次深入探讨了基于视觉示例的图像生成任务，着重研究了如何通过“文本反演（textual inversion，TI）”技术来实现个性化的文本到图像生成。该方法通过收集用户提供的3～5幅图像来表示特定的概念，如对象或风格，学习它们的视觉特征，将这些概念表示为嵌入空间中的“伪词”，以捕获高级语义和精细的视觉细节。这些伪词可以组成自然语言句子，通过这种方式，可以利用自然语言生成特定概念的图像，修改其外观，或将其组合在新颖的场景中，以直观的方式引导基于用户输入图像中物体的个性化图像生成。

在与Textual Inversion的同期，Google也针对于它们的Imagen（Saharia等，2022）模型提出了DreamBooth（Ruiz等，2023）这项基于视觉示例驱动的图像生成方法，并将此类生成方法称为“主题驱动（subject-driven）”。该方法通过引入稀有标识符（通常是在词表中不会用到的稀有词）来表示特定物体，并采用一种新的类特定先验保留损失技术，以保持对实例核心视觉特征的高保真度，使得模型在不同的自然语言上下文中为该实例生成各种风格的图像。目前，DreamBooth和TI技术已经相结合使用在Stable Diffusion框架中，能够应用于多种文本引导的图像生成应用，包括特定实例的场景生成和艺术风格生成。

为了实现更加灵活的基于主题驱动的图像定制，特别是在多个实例的组合生成问题上，Kumari等人（2023）提出了Custom Diffusion技术，以增强主题驱动方法的性能。该方法能够在仅优化图像生成模型中的少量参数的条件下，实现对用户输入的新概念的快速微调。其核心在于优化文本到图像扩散模型的交叉注意力层中的参数，以高效地将输入实例的图像学习为新概念，并在面对多个概念组合时，可以先单独训练各个概念模型，再通过约束优化将多个微调模型合并成一个，这使得该方法可以在文本提示条件下生成包含多个新概念和现有概念视觉实例的定制图像。

不同于现有的基于实例的生成模型通过将视觉实例表示为伪词来引导生成，BLIP-Diffusion（Li等，2023a）引入了一个预训练的多模态编码器来提供实例表示，该编码器首先基于BLIP-2（Li等，2023b）文本—视觉大模型进行预训练，以产生与文本对齐的实例表示，然后构建生成指令使扩散模型能够利用这种视觉表示来生成各种个性化图像。与DreamBooth等之前的方法相比，BLIP-Diffusion模型既支持零样本的主题驱动生成，也能针对特定实例进行模型微调，且微调效率相比DreamBooth提高了20倍。

3 生成图像溯源技术

图像取证技术（Shi等，2023）是一种通过分析图像特征来确定图像内容的真实性、完整性和来源的技术，可分为主动溯源取证和被动溯源取证两种取证方式。主动溯源取证的方式主要有数字签名（Alam等，2015）和数字水印（Ding等，2022b）两种技术手段。数字签名是一个包含文件信息以及发送者身份，利用加密技术进行编码的字串；数字水印用信号处理的方法在原始信息中嵌入特定的标识信息。数字签名技术一般应用于电子邮件或者数字文档中；而数字水印是对数字图像进行溯源验证的主流技术，能够进一步适应AIGC时代下生成图像的溯源

场景中。

目前，尽管图像生成模型的快速发展逐渐满足使用者对高保真图像定制化、专门化的设计需求，但也使得社交媒体中传播图像的真实性难以辨别，从而导致生成图像存在被滥用于恶意目的的风险。例如，在推特上，一些账户发布有关灾情的AI合成图像，通过收取加密货币来诈捐。此类事件严重误导舆论走向和民众决策，对信息可靠性和社会稳定性带来不良影响。为保障生成内容的安全性，包括OpenAI和谷歌在内的多家科技巨头联合发表声明，承诺将在未来产品的生成图像中集成隐蔽的、可识别的水印，以开发安全、可靠的人工智能技术（Wu和Liu，2023）。因此，为缓解生成图像带来的安全性问题，展开结合图像生成与以图像水印为代表的溯源技术研究迫在眉睫，当前亟需完善技术手段来实现对生成图像的溯源保护。与传统的变换域图像水印技术不同，生成图像水印旨在利用深度学习技术使得水印嵌入适配生成模型的生成过程，使得生成图像携带溯源标识信息，并且在水印鲁棒性和图像生成质量上有良好表现。在图像生成过程中，要求生成语义一致、携带水印的高质量图像；在水印验证过程中，要求解码有效抵御外界攻击的强鲁棒水印。

作为当前图像生成安全的技术热点，本节将深入介绍和分析生成图像溯源的最近研究进展。首先，详细介绍无水印嵌入的生成图像溯源、水印前置嵌入的图像溯源、水印后置嵌入的图像溯源以及联合生成的水印溯源等4种技术类型，并对每种方法的特点和优劣进行深入分析，将代表性方法在不同维度下的特点进行详细对比，结果如表1所示。其次，详细介绍了目前针对生成图像的水印攻击研究现状。由于图像生成、风格迁移、文生图和图像编辑等生成式技术的迅速迭代导致针对生成图像水印技术的催生和发展，可以窥见现有的水印方法已经展现出巨大的研究潜力和应用价值，能够为生成式技术保驾护航。在未来的研究中，生成图像溯源技术将朝着图像高质量、水印强隐蔽和水印强鲁棒的方向进一步发展。这将为生成式技术的应用提供更多可能性，并在确认图像归因、保障生成内容安全等方面发挥重要作用。

表1　生成图像溯源代表性方法对比

Table 1　Comparison of representative methods for the traceability of generated image

嵌入类型	算法	任务	输入	溯源信息形式	侧重性质	溯源对象
无水印嵌入	Marra等人（2019）	图像生成	噪声采样	指纹	计算开销	模型
无水印嵌入	Yu等人（2019）	人脸生成	噪声采样	指纹	计算开销	模型
前置嵌入	Yu等人（2021）	人脸生成	噪声采样	二值序列	可扩展性、隐蔽性	模型
前置嵌入	Fernandez等人（2023）	图像生成	噪声采样	二值序列	隐蔽性、鲁棒性、图像质量	模型
后置嵌入	Fei等人（2022）	人脸生成	噪声采样	二值序列	鲁棒性、隐蔽性、图像质量、可扩展性	模型
联合生成	Yu等人（2022c）	人脸生成	噪声采样	二值序列	鲁棒性、图像质量、可扩展性	用户
联合生成	Liu等人（2023）	文生图	文本、创作元数据	图像水印	鲁棒性、隐蔽性、图像质量、可扩展性	用户、输入文本、生成时间

3.1　生成图像的水印相关溯源技术

本节依据水印嵌入的流程节点，将现有的水印相关生成图像溯源方法分为无水印嵌入的生成图像溯源、水印前置嵌入的生成图像溯源、水印后置嵌入的生成图像溯源以及联合生成的生成图像溯源等4类技术展开介绍。

生成图像溯源分析维度划分如图4所示，将生成图像溯源从溯源信息形式、评估标准、溯源对象、目标模型、方法划分、生成任务以及溯源目的等方面进行分析。其中，主要的评估标准（吴汉舟 等，2023）要求如下：

1）鲁棒性。溯源信息对外部攻击具有较强防御

能力，能够有效应对内容篡改、几何攻击和图像处理攻击。

2）隐蔽性。生成图像中的溯源信息具有较强的不可见性，难以被察觉和感知。

3）图像质量。嵌入溯源信息后尽量降低对图像质量的影响，将嵌入内容隐藏在空域、频域或者特征域中。

4）计算开销。对于溯源信息的嵌入和提取，图像生成模型产生较小的额外计算开销。

5）可扩展性。溯源信息的嵌入和提取能够泛化到其他生成模型中。

此外，生成图像的溯源目的具体如下：

1）内容监管（Cui等，2023）。是确保生成内容具备正常合理的语义表达的监管手段，维护生成内容的可信性，促进数字创作环境的发展和繁荣。

2）归属确认（Liu等，2023）。确定生成内容的创作用户，从而防止生成内容被复制、传播和修改而导致生成内容的原创性混淆，确保其合法权益得到保护。

3）模型归因（Yu等，2019）。追溯生成内容的模型来源，确定生成内容的生成模型，并提供相应的归因证据。

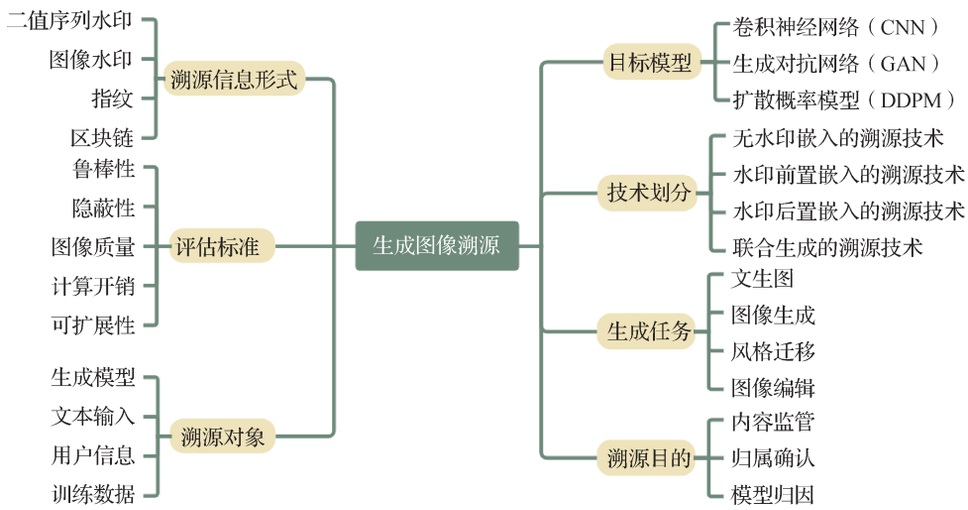

图4　生成图像溯源分析维度划分

Fig. 4　Division of dimensions for the generated image traceability analysis

3.1.1　无水印嵌入的生成图像溯源

无水印嵌入的生成图像溯源方法，将模型生成时所遗留的生成痕迹视做指纹，旨在利用此类固有指纹信息来实现生成图像的模型归因，从而达到溯源目的。该类方法表明，即使在图像生成过程中不主动嵌入水印标识信息，生成图像中也会存在某一类生成模型的特定指纹信息，本质上反映了模型的图像生成模式并且其不具有与图像语义的相关性。该类方法的示意如图5所示。

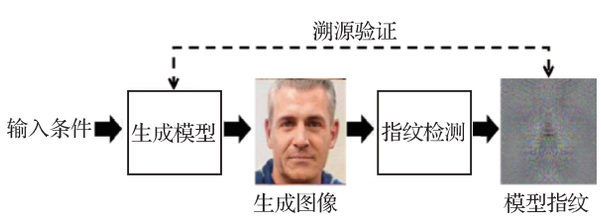

图5　无水印嵌入的生成图像溯源示意

Fig. 5　Watermark-free embedding for the traceability of generated images

为了应对逼真的生成图像对多媒体安全构成的严重威胁，需要对图像是否由模型生成进行判别。Marra等人（2019）首先提出并验证了不同GAN会在其生成图像中留下特定指纹的猜想，通过去噪滤波器从图像中提取噪声残差，该残差由指纹和零均值随机噪声叠加共同构成，再利用大数定理通过多次残差的平均值估计指纹，为后续多媒体取证以及图像溯源提供了理论依据。根据Marra等人（2019）的研究结论，Yu等人（2019）率先提出了对基于GAN生成的人脸图像进行模型归因的研究，设计端到端自编码器学习生成图像中遗留的稳定指纹来确定图像归属，并且能够免疫多种对抗性图像扰动。结果表明GAN中的指纹存在唯一性。与Yu等人（2019）的工作类似，针对实现深度伪造的人脸图像取证，Albright和McCloskey（2019）通过人脸图像来学习多

个预训练生成器与其最匹配的潜在编码，从而实现对人脸图像的归因。

然而，此类归因方法存在局限性，仅仅在某一类别（例如人脸）的图像数据集上训练分类器，这类方法很可能会与图像生成中使用的数据集类别产生过度关联，并且由于数据集偏差（Yu等，2016），可能无法泛化到新数据集上。为解决此问题，Wang等人（2020）研究得出，经由卷积神经网络生成的图像保留了可检测的指纹，并据此利用ProGAN（Wang等，2020）训练一种通用检测器，能够检测出任意架构的卷积神经网络和任意类别数据集生成的图像并对其溯源，防止恶意用户主动挑选超过现有技术检测阈值的虚假恶意图像，将其传播在社交媒体平台中。

此外，现有的研究已经可以高精度地将生成图像归因于相应的GAN模型，然而仅在封闭场景中有效，不能泛化到在训练中没有出现的GAN模型。因此，Girish等人（2021）将目光转移到GAN在开放世界中的生成图像归因，提出了一种迭代算法，由分布外检测、K均值样本聚类、相干簇融合和簇细化等步骤组成，最终实现来源相同的生成样本聚类。该方法通过利用不同GAN在其生成的图像上留下不同指纹的事实，来认证未参与训练的GAN生成的图像，具有较高的生成图像归因精度。与Girish等人（2021）工作类似，基于Marra等人（2019）已经证实不同GAN生成的图像中存在不同指纹的发现，Yang等人（2023）也意识到闭集分类设置限制了在现实世界场景中处理任意模型生成的图像的应用。所以Yang等人（2023）提出开放集模型归因，将图像同时归因于已知模型甚至是未知模型。该任务难点在于来自已知和未知模型的生成图像之间存在视觉上难以察觉的痕迹，据此Yang等人（2023）提出了一种渐进式开放空间扩展方法，利用一维傅里叶功率谱的方位角积分来计算指纹空间，依据已知模型的边界来模拟未知模型的潜在开放空间，在已知模型上添加双层卷积进行模型增强，从而模拟构建开集样本，使其保持与闭集样本相同的语义，但嵌入了不同的难以察觉的痕迹。实验结果表明，用人工构造的模型来模拟已知模型的边界，在开放集场景中是有效的。研究指纹如何受到网络架构的影响，在未来可能会导致更通用的指纹空间表示。

但是Girish等人（2021）的方法仅针对特定类别图像有效，并未解决多类别语义图像的归因问题，如人脸、动物、场景等多类别。针对该问题，Bui等人（2022）提出一种新的表征混合训练策略，通过在多个生成器之间进行生成图像的特征混合，实现对图像指纹识别的多类别泛化，从而具有追踪图像来源的能力。并且该模型不受图像语义内容的影响，对常见扰动也具有鲁棒性。

在生成模型参数和架构不可知的情况下，如何对生成图像进行模型归因和溯源仍未存在有效解决方案。依据Marra等人（2019）的生成图像包含水印指纹的研究结果，Asnani等人（2023）对生成图像进行逆向工程，通过指纹估计和生成模型聚类的方法，为生成图像反推生成模型网络架构和训练损失函数，这个方法也能够进行深度伪造检测和图像归因。

综上所述，无水印嵌入的生成图像溯源方法优势在于，无需在图像生成过程中额外嵌入信息，所以不会损害生成图像的质量。其局限性在于，随着生成式技术迅速发展，图像的生成痕迹越发微弱，从生成图像中挖掘指纹来进行归因验证变得更加困难，不能稳定地实现生成图像的溯源保护目的。

3.1.2　水印前置嵌入的生成图像溯源

水印前置嵌入的生成图像溯源方法，旨在将水印嵌入到噪声、图像等输入训练数据中，利用嵌入水印的数据来训练生成模型，能够使得生成图像也具备溯源信息。并且水印能够被特定解码器进行验证解码，与预先指定的水印进行一致性校验，最终确定图像归属。但由于水印前置嵌入方法都需要水印编解码器的预训练或者水印嵌入训练数据的预处理，所以构建代价较高。该类方法的示意如图6所示。

随着生成式技术逐渐走向成熟，生成图像逐渐对多媒体安全构成了严重威胁，虚假信息在社交网络中泛滥成灾。Hu等人（2018）首先展开对信息进行前置嵌入的探索。在生成阶段，在发送端将秘密信息映射成噪声向量，将其送入预训练生成器后，随机生成基于噪声的载体图像，在图像生成过程中无需进行修改和嵌入信息。在验证阶段，于接收端经由提取网络提取出噪声向量，再恢复为秘密信息。不同于Hu等人（2018）对秘密信息在输入端编码的方式，为解决图像生成模型所有权验证的问题，Ong等人（2021）提出了一种预先将所有权信息嵌入到具有正则化项的生成器中，能够泛化到各种GAN中，并且适用于白盒和黑盒场景。在黑盒场景中，该工

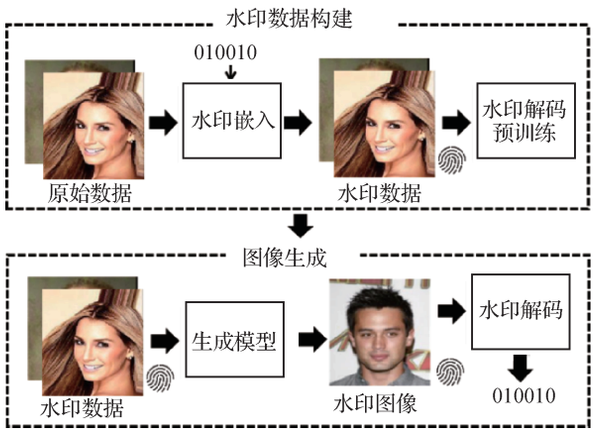

图6　水印前置嵌入的生成图像溯源示意

Fig. 6　Watermark pre-embedding for the traceability of generated images

作提出重构正则化方法，允许生成器在给定触发输入后，在合成图像的指定位置嵌入可见水印；而在白盒场景中，该工作采用并改进了Fan等人（2019）提出的符号损失，使得归一化的神经网络层携带水印信息。两种场景下的生成图像均包含所有权相关的溯源信息，能够实现生成图像的归属验证。

由于GAN模型的技术革新，人脸合成图像的真实性达到极高的质量水平，进一步引发了人们对虚假信息的担忧。尽管现有的检测方案具有较高准确率，但会受到生成技术迭代的影响。因此，Yu等人（2021）在训练数据中人工嵌入指纹序列并训练对应的指纹解码器，将携带指纹的数据送入生成模型训练后，指纹会转移到生成的图像中，并且几乎不会对生成图像的质量产生损害。该技术中生成模型对应的指纹具有唯一性，适用于基于GAN的人脸生成模型，并且不会随着生成技术迭代更新而失效。Yu等人（2021）通过二值序列指纹嵌入将合成图像检测和图像归因统一为一个任务。而Zhao等人（2023b）认为，目前的深度伪造检测研究仍处于起步阶段，因为当前主要依赖于捕捉深度伪造合成过程中留下的指纹作为检测线索，这些指纹可以通过各种失真（例如模糊）或更先进的深度伪造技术轻松去除。所以Zhao等人（2023b）不依赖于识别指纹，而是利用标识信息嵌入的机制来保护人脸图像免受深度伪造恶意篡改。具体来说，设计一个具有编码器—解码器结构的神经网络，先将人脸图像通过属性解码器和身份解码器分别解耦出属性和身份特征，水印作为标识信息嵌入到身份特征中，再与属性特征深度融合，经过深度伪造网络的修改后，身份解码器无法识别伪造图像中的水印。因此它要求对面部图像编辑（即深度伪造）敏感，对传统的图像修改（例如调整大小和压缩）鲁棒。Nadimpalli和Rattani（2023）同样意识到深度伪造的面部生成存在重大的安全隐患。目前的深度伪造被动检测是一种事后取证对策，无法提前阻止虚假信息的传播。因此，需要更为直观的方法来为人脸合成图像添加可见扰动，作为主动防御。Nadimpalli和Rattani（2023）提出一种新的基于GAN的可见水印的主动深度伪造检测技术，通过对输入图像指定位置添加触发器，并利用重建正则化来优化生成对抗损失函数，将唯一水印可见地嵌入到人脸生成图像的指定位置。该可见水印方法能够减轻公开的预训练GAN和相关智能手机应用程序生成的人脸伪造图像在社交媒体的快速传播造成的危害。

在当前用于GAN的基于后门的水印方法（Ramesh等，2021）中，水印的触发模式很容易被检测到，这可能无法通过溯源有效实现神经网络产权保护。为了解决这一问题，Zeng等人（2023）提出了一种针对GAN的黑盒水印方法，来挖掘不可见的模型后门。具体地，预先训练水印隐藏和提取网络，将输入数据进行水印嵌入的预处理。之后分别将隐藏水印的触发图像和原始图像送入GAN中，原始图像对应正常输出，而触发图像对应后门输出，此时水印变为可见，能够通过溯源生成图像的归属来确认生成网络版权。类似地，为针对性地保护GAN的产权，Qiao等人（2023）在训练阶段，将水印嵌入到校验图像中以构成触发集，再将训练数据和触发集整合以训练GANs，在验证阶段，提供水印密钥送入GAN，同时进行图像生成和所有权验证。

扩散概率模型展示了它们在学习和生成图像方面的强大能力，进一步推动了生成图像在各个领域的多元化应用。然而，这种不受限、不安全的生成内容传播引起了人们对作品版权保护的关注。例如，包括画家和摄影师在内的艺术家越来越担心生成扩散模型可以在未经授权的情况下轻而易举地复制和修改作品。为了应对这些挑战，Cui等人（2023）首先提出，不同图像的水印之间应该保证模式一致性，水印样式之间不能有太大偏差，所以引入了一种多样性水印归一化的方案，通过将所有权信息进行文本编码，以水印形式嵌入到输入图像中，即便经过扩散

模型的多次编辑，水印解码器也能从生成的图像中检测水印，这可以有效预防侵犯版权的行为。

随着Midjourney（https://www.midjourney.com/）和Stable Diffusion（Rombach等，2022）等大型生成扩散模型公开发布，生成模型受到广泛关注。由于其易访问性，针对数据自动收集和内容所有权的问题开始显现。为了解决这一问题，Ditria和Drummond（2023）提出了向公众安全共享图像的方法。首先，该工作证明在嵌入不可见水印的数据上训练的生成扩散模型将生成具有相关水印的新图像。其次，该工作进一步分析统计数据证明，如果给定的水印与训练数据的某个特征相关，则生成的图像也将具备这种相关性。具体地，与Yu等人（2021）提出方法的流程类似，训练水印编码器和解码器，将训练数据嵌入水印后，经过扩散模型生成的图像能够解码出水印。因此，该系统提供了一种在线共享内容时保护知识产权的解决方案。

随着扩散概率模型的进一步发展，主题驱动的可控生成逐渐成为研究热点。目前主题驱动的生成模型可以利用来自特定主题（例如人脸、艺术风格等）的一些图像进行模型微调来实现特定主题的图像个性化生成。然而，滥用主题驱动的图像合成可能会侵犯主题所有者的权益。针对主题滥用风险问题，Ma等人（2023）首先使用大规模数据集预训练水印生成器和水印检测器，使用预训练水印生成器在特征层级为训练数据嵌入扰动作为水印。随后在主题驱动生成阶段，对输入图像数据进行主题微调，生成图像送入水印检测器之后仍能解码出水印，从而防止原始图像的篡改和滥用。

在生成模型分发和部署的场景中，由于生成图像的高逼真性，有研究者对用户是否会滥用生成模型产生了道德担忧，需要溯源手段对其进行监管。所以Fernandez等人（2023）提出了一种结合图像水印和潜在扩散模型的内容溯源方法，预先训练水印嵌入的编解码框架，再将水印解码器接入扩散模型后与图像解码器一同微调，使部署的某个扩散模型生成的图像具有特定的二值序列水印，并允许检测和识别。

综上所述，水印前置嵌入的生成图像溯源方法优势在于，预先嵌入的信息能够在生成过程中得到有效保留，并且对生成图像质量的损害较小。其局限性在于，需要对水印编解码器进行预训练或者水印嵌入训练数据进行预处理，在大规模数据集上的构建代价较高。并且由于现有方法大多通过生成图像溯源来实现对模型的归因，所以生成图像中的水印具有单一性。

3.1.3 水印后置嵌入的生成图像溯源

水印后置嵌入的生成图像溯源方法，将水印图像的生成过程分为图像生成和水印嵌入两阶段的组合，水印嵌入是在图像生成之后进行。在图像生成阶段，生成器根据给定输入来生成图像；在水印嵌入阶段，将已有的生成图像嵌入水印信息。与其他生成图像溯源方法相同，设置特定水印解码器将图像中的水印进行溯源验证。该方法的示意如图7所示。

为保护GAN的知识产权，Fei等人（2022）提出了一种水印后置嵌入的方法，将预训练二值序列水印的编码器接入生成模型之后，使得GAN输出的任意图像都包含一个不可见的水印标识，通过预训练解码器进行图像所有权的溯源验证。此外，在训练过程中也会微调水印编码器使其适应图像的生成模式。该方法具有泛化性，可用于任何基于GAN结构的溯源保护。

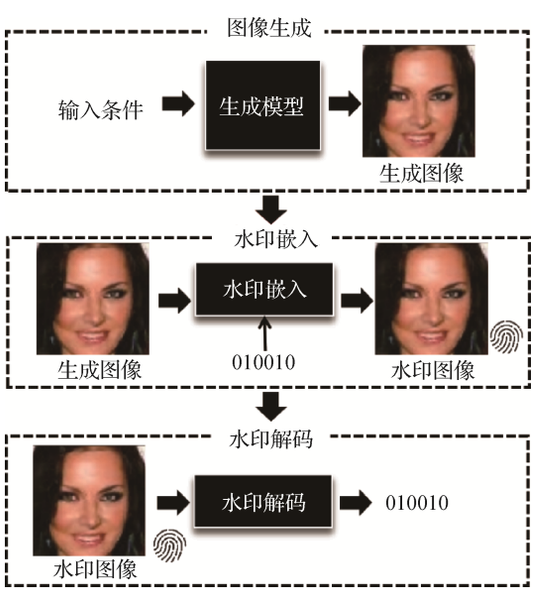

图7 水印后置嵌入的生成图像溯源示意

Fig. 7 Watermark post-embedding for the traceability of generated images

为防止对生成模型恶意微调而植入后门，Yin等人（2022）提出了一种针对生成网络的脆弱水印方法，设计正则化项约束生成器训练以生成脆弱触发

集图像，将该类触发集图像样本送入后置分类模型来降低分类准确性。所以在推理过程中，如果生成模型被篡改，生成图像输入后置分类模型的分类准确率会大大降低。该方法本质上是实现对分类模型的标记，并通过脆弱生成样本的追溯实现对模型篡改的检测。

针对现有方法在图像质量、抗扰动鲁棒性等方面存在不足，并且模型训练代价较高的问题，Bui等人(2023)提出一种在潜在空间下隐藏水印的自编码器，将二值序列信息作为偏移量，与生成图像的隐变量进行特征叠加，实现水印嵌入。该模型优势在于较少的参数量与灵活的模块化设计，并且具备较强的水印信息恢复能力。

综上所述，水印后置嵌入的生成图像溯源方法具有以下优势：首先，图像生成和水印嵌入过程是分离的，因此能够确保生成和嵌入之间相互独立，避免了相互干扰的问题；其次，水印嵌入对生成图像质量的影响是比较微弱的。其局限性在于，在图像生成和水印嵌入的两个阶段之间存在信息被篡改的风险，这可能导致生成图像溯源的不精确性。因此，在使用水印后置嵌入的方法进行生成图像溯源时，还需采取相应的技术手段来减轻这些问题。

3.1.4　联合生成的生成图像溯源

联合生成的生成图像溯源方法，旨在图像生成过程中实现水印信息的自适应嵌入，在与图像特征融合时尽量减少对图像生成过程的损害，最终生成携带水印的图像。在验证过程中，通过特定的水印解码器能够将图像中的水印进行解码验证。该方法的示意如图8所示。

图8　联合生成的生成图像溯源示意

Fig. 8　Joint generation for the traceability of generated images

随着生成模型表现出强大的生成能力，研究者逐渐注意到除了保护生成图像的归属权，生成模型的版权保护同样应该受到关注。为通过生成图像溯源来保护生成式神经网络的产权，Wu等人(2021)在生成模型中预设唯一的水印信息，并设置专门的水印提取网络通过密钥验证来获取水印。每个GAN模型由唯一水印进行标识，可以通过检测输出图像含有的水印信息来识别归属于何种生成网络。该工作能够有效应用于图像上色、图像编辑等生成任务。

深度生成模型已经达到高性能水平，人脸生成数据很难与真实数据区分开来，人们开始担心人脸图像可能会被滥用于深度伪造并大规模传播虚假信息。并且目前的深度伪造检测方法是不可持续的，无法应用于特定用户模型分发的场景。据此，Yu等人(2022c)提出了一种可扩展的白盒方法，将二值序列指纹嵌入到生成模型的卷积层参数中，以生成携带指纹的人脸图像，对指纹进行验证后能够对生成图像的模型持有者进行追溯归因。

生成式任务神经风格迁移逐渐兴起，但现有的神经风格迁移方案在创造多样性和个性化的艺术风格方面能力较弱。另外，生成的风格化图像在互联网上共享时容易被窃取和非法转发。为此，Wang等人(2023)提出个性化和水印引导的风格迁移网络，使用多种个性化向量引导内容和风格的图像对，自粗粒度向细粒度的风格迁移过程中，将二值序列水印嵌入到深层特征空间中最终生成具有版权信息的多样性风格化图像。

文本到图像的跨模态生成是生成式任务的热点之一，这也导致了侵犯隐私和虚假信息传播风险的增加，故依据文本生成的图像存在溯源需求。然而，由于水印前置嵌入数据难构建，并且后置水印嵌入方法在阶段之间存在信息泄露的风险，现有的文本到图像生成方法缺乏将可溯源信息与图像生成联系起来的能力。Liu等人(2023)提出一种文本到嵌入水印的图像跨模态生成任务，并提出一种泛化到多种生成模型的跨模态生成方法，主要包含图像和水印联合生成以及图像和水印协同解耦两个部分。在图像和水印联合生成过程中，利用用户信息、输入语料等维度的溯源元数据，学习溯源信息的联合水印表示，再通过基于卷积网络构建的特征耦合模型进行语义及水印特征协同表示的水印隐蔽嵌入，实现携带隐蔽水印、语义一致的图像生成。在图像和水印协同解耦过程中，利用非合作博弈理论解耦无水印图像和水印，达到两者质量权衡，并采用后处理数据增强策略，进一步保证水印鲁棒性，实现生成图像的溯源验证。此外，Liu等人(2023)从图像质量、水印隐蔽性和水印鲁棒性的角度提出一套全面的评估体系。实验结果表明，该方法生成的图像具有高质

量，嵌入水印具有强隐蔽性，解码水印具有强鲁棒性。

针对扩散模型仅仅能够嵌入固定水印序列以及后置嵌入时容易遭受逃逸攻击的问题，Xiong等人(2023)提出一种基于编码器—解码器和消息矩阵嵌入的端到端方法。在图像生成的前向扩散中，通过融合消息矩阵和潜在编码，可以将水印序列嵌入到生成的图像中。因此，可以通过利用消息编码器生成消息矩阵来灵活地更改消息，而无需再次训练目标模型。此外，该方法设计了一种安全防御机制，用来防御图像生成过程中的消息矩阵逃逸，破坏未嵌入消息矩阵的生成图像。

综上所述，联合生成的生成图像溯源方法优势在于无需事先构建携带水印的训练数据，并且不存在后置水印嵌入方法存在的多阶段间逃逸攻击风险，模型泛化性较强，能够以较小的训练代价应用于各类生成模型中。其次，该类技术不会随着生成技术的革新或者数据集的变更而失效，能够有效实现生成图像的溯源，保障生成图像的安全性。该类联合生成方法尚处于起步阶段，随着生成式技术的不断迭代，生成图像的可溯源能力对水印容量、鲁棒性及隐蔽性将会提出更高要求，所以该类技术具有广阔的研究前景。

3.2 生成图像的水印攻击

随着生成图像水印技术的发展，水印攻击可以作为一种测试水印算法鲁棒性的手段，帮助评估和改进水印算法的性能。除了常见的图像处理攻击(例如裁剪、旋转和高斯噪声等)，生成图像的水印攻击研究更多侧重于利用深度神经网络来破坏生成图像中隐藏的水印信息，增加水印解码的误码率，并且减少对生成图像视觉效果的破坏，使得生成图像在攻击后仍然能够保持较好的视觉质量。此类研究对于提高生成图像水印的安全性和可靠性具有重要意义。

研究者们提出了许多鲁棒的盲水印算法和模型，并实现了良好的效果。然而，目前针对生成图像的水印攻击算法的研究尚属于起步阶段，其研究广度还不能与水印添加算法相媲美。并且许多水印攻击算法仅仅关注干扰水印的正常提取，而忽视了对图像造成的严重视觉损害。为此，Li(2023)提出了一种用于水印攻击的条件扩散模型，去除嵌入水印的同时恢复图像。该方法的核心在于，在未标记图像上训练图像到图像的条件扩散模型，在图像加噪和去噪的采样过程中，使用距离引导算法保证图像恢复的完整性，以便生成与原始图像相似的攻击图像。Zhao等人(2023a)提出的攻击方法首先在生成图像中添加随机噪声以破坏水印，再进行图像的重建，结果表明所有基于像素的水印都容易受到该种攻击而被擦除。对于水印方法RivaGAN(Zhang等，2019)，重建攻击能够去除93%～99%的不可见水印。此外，该方法能够保证重新生成外观几乎不变的图像。

针对生成图像的水印攻击目前仍处于起步阶段，现有研究都是基于生成模型对目标生成图像进行重建，并且主要针对图像形式的水印攻击，对于二值序列形式的水印攻击尚未存在有效的策略。为了增强水印的鲁棒性，还需要进一步探究水印攻击模式，以便推进具备更强防御能力的生成图像鲁棒水印研究，并为鲁棒水印的发展提供更加可靠和安全的解决方案。

4 展　望

4.1 图像生成技术展望

随着生成模型的迅速发展，图像生成技术日益成熟并且广泛应用于多个领域。然而，其未来的发展仍然面临着挑战和机遇。以下是本文对图像生成技术未来发展的展望，主要从3个方面进行探讨：

1)交互式生成技术。目前的图像生成技术往往需要复杂提示工程技能(给模型输入精心设计的提示词组、特殊标签和附加条件)才能够获得令人满意的生成结果，这为普通用户带来了极大的门槛，限制了图像生成技术的应用和推广。DALL·E-3(Betker等，2023)的出现为上述问题提供了一个有效的解决方案。作为一个与人类语言交互的文生图模型，DALL·E-3允许用户通过自然语言描述与模型交互，进行图像的生成、编辑和细化。这为用户提供了一个直观和友好的交互界面，极大地降低了使用门槛，使得非专家用户也能轻松地获得满意的生成结果。此外，DALL·E-3还支持与图像和文本内容相关的问题回答，进一步增强了模型的交互能力和实用性。在未来，交互式生成也将成为重要的研究方向。通过实时的用户反馈，生成模型可以不断调整生成结果，确保生成内容完全符合用户的需求和意图。

2)高分辨率和高质量生成。随着计算能力的增强和模型结构的优化,未来的图像生成技术将能够生成更高分辨率和高质量的图像内容。当前的生成模型在生成高分辨率图像时仍然面临许多挑战,如细节丢失、生成噪声等。随着技术的进步,这些问题有望得到解决。例如,结合多尺度的生成策略,生成模型可以从粗到细地生成图像内容,确保生成结果既有宏观的结构完整性,又有微观的细节丰富性。此外,高质量的生成也需要考虑生成内容的真实性、一致性和多样性。未来的研究需要不断探索和优化生成模型的结构和策略,以实现更为高质量的图像生成。

3)可解释的生成控制。为了增强用户对生成过程的信任和理解,未来的图像生成技术需要提供更为直观和具有解释性的控制机制。当前的生成模型多数基于深度学习技术,其工作原理和决策过程往往难以解释和理解。但随着可解释性研究的深入,预期未来的生成模型将能够提供更为明确和直观的控制策略。例如,用户可以通过直观的界面和工具,指定生成内容的风格、结构和细节,而生成模型则可以根据这些指示,提供相应的生成结果。此外,生成模型也需要提供对其决策过程的解释,帮助用户理解生成结果的来源和原因,确保生成过程的透明性和可信度。

综上所述,图像生成技术正经历快速发展,其中交互式生成、高分辨率和高质量生成以及可解释的生成控制都是关键研究方向,本文对这些方向进行了展望,期待它们进一步推进技术进步,满足更广泛的应用需求。

4.2 生成图像溯源技术展望

通过对生成图像溯源技术的阐述和分析,本文对该技术的未来发展前景做展望,主要包括3个方面:

1)联合生成式的可扩展水印技术。首先,生成技术的迅速迭代导致图像的生成痕迹越发微弱,而生成图像中遗留的指纹信息将会更加难以提取,所以依赖固有指纹信息的无水印嵌入溯源方法的可靠性将会大幅降低。其次,目前的生成模型都在大规模数据集上进行训练和微调,水印前置嵌入溯源方法的数据构建代价将会越来越高昂,水印的提取和嵌入不够高效,而水印后置嵌入溯源方法存在逃逸攻击风险,此类风险难以规避。在未来的工作中,研究者可以注重联合生成的溯源水印方法,深入探究更适配生成模型的水印信息嵌入架构。此外,考虑到目前生成模型参数量较大难以进行训练和微调,所以应当侧重于轻量级附加模块的研究,通过训练低计算开销、强可扩展性的水印编解码器达到生成图像高质量、水印强隐蔽以及水印强鲁棒。

2)水印攻击模式。针对生成图像的水印攻击研究刚刚起步,且都是利用生成模型进行图像重建的方式来去除水印,但这种攻击方式仅仅对图像水印有效,尚未有一种攻击方式能够破坏各种形式(例如二值序列)的溯源水印,面向水印攻击模式的研究还有很大的前景和空间。

3)面向基于神经网络攻击的水印鲁棒性。现有关于水印鲁棒性的评测仅仅局限于图像后处理攻击,而基于生成模型构建的攻击手段,并没有很好的防御和评估机制。因此,随着水印攻击模式的深入研究,现有的水印技术可能无法确保强鲁棒性。所以,无论是针对图像处理攻击,还是基于神经网络的攻击,关于水印的鲁棒性需要进一步探究。

综上所述,生成模型的蓬勃发展导致生成图像溯源技术将成为未来的研究热点,本文展望了未来该技术的发展方向,希望能够引导和推动溯源技术的革新和推广,为生成图像的安全提供强有力的技术保障。

5 结 语

随着数字化时代的到来,多媒体内容生成已经渗透到日常生活的方方面面,深刻影响了电影、游戏、设计和虚拟现实等应用领域。在此背景下,AIGC技术为内容创作者提供了无尽的可能性,推动了视觉内容生成技术的巨大进步。当前AIGC带来的机会与挑战并存,生成内容的质量和安全性问题一直是研究者关注的焦点。

在图像生成技术方面,从基于GAN的传统图像生成方法到目前的基于扩散概率模型的大型生成模型,相关技术的研究已经取得了突破性的进展。这些技术不仅允许以前所未有的效率快速生成高质量的视觉内容,还提供了更为精确和灵活的控制手段,如引入布局、线稿等附加信息以及基于视觉参考的技术。

然而,生成技术的发展衍生出新的安全性问题。

生成图像有可能被用于如深度伪造和虚假新闻制作等恶意目的，影响社交媒体传播内容的可靠性。为缓解这些问题，研究者针对各类场景开发了以水印为主的多样化生成图像溯源技术，以确保生成图像的真实性和可靠性，实现防范恶意攻击和生成图像溯源。本文将现有生成图像溯源技术划分为无水印嵌入、水印前置嵌入、水印后置嵌入和联合生成等4类方法，并详细介绍每类方法的特点、优势以及局限性。同时，针对生成图像的水印攻击研究也在不断发展，旨在推进具备更强防御能力的生成图像鲁棒水印研究。

综上所述，AIGC时代的图像生成技术展现了巨大的研究潜力和应用价值，而生成图像溯源技术为这些图像生成技术的应用和进一步发展提供了安全保障。目前仍需继续深入研究和完善这些技术，以充分发挥其潜在应用价值，同时确保其安全和有效性。随着技术的迅速发展，未来将更好地平衡创新与安全，为数字内容创作者和用户带来更加丰富和安全的体验。

致　谢　本文由中国图象图形学学会数字媒体取证与安全专业委员会组织撰写，该专委会链接为https://www.csig.org.cn/16/201704/49326.html。

参考文献（References）

Alam S, Jamil A, Saldhi A and Ahmad M. 2015. Digital image authentication and encryption using digital signature//Proceedings of 2015 International Conference on Advances in Computer Engineering and Applications. Ghaziabad, India: IEEE: 332-336 [DOI: 10.1109/icacea.2015.7164725]

Albright M and McCloskey S. 2019. Source generator attribution via inversion//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 8: #3

Asnani V, Yin X, Hassner T and Liu X M. 2023. Reverse engineering of generative models: inferring model hyperparameters from generated images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12): 15477-15493 [DOI: 10.1109/TPAMI.2023.3301451]

Betker J, Goh G, Jing L, Brooks T, Wang J F, Li L J, Ouyang L, Zhuang J T, Lee J, Guo Y F, Manassra W, Dhariwal P, Chu C, Jiao Y X and Ramesh A. 2023. Improving image generation with better captions [EB/OL]. [2023-11-05]. https://cdn.openai.com/papers/dall-e-3.pdf

Bui T, Agarwal S, Yu N and Collomosse J. 2023. RoSteALS: robust steganography using autoencoder latent space//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 933-942 [DOI: 10.1109/cvprw59228.2023.00100]

Bui T, Yu N and Collomosse J. 2022. RepMix: representation mixing for robust attribution of synthesized images//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 146-163 [DOI: 10.1007/978-3-031-19781-9_9]

Cui Y Q, Ren J, Xu H, He P F, Liu H, Sun L C, Xing Y and Tang J L. 2023. DiffusionShield: a watermark for copyright protection against generative diffusion models [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2306.04642.pdf

Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis//Advances in Neural Information Processing Systems, 34: 8780-8794

Ding M, Yang Z Y, Hong W Y, Zheng W D, Zhou C, Yin D, Lin J Y, Zou X, Shao Z, Yang H X and Tang J. 2021. CogView: mastering text-to-image generation via Transformers//Advances in Neural Information Processing Systems, 34: 19822-19835

Ding M, Zheng W D, Hong W Y and Tang J. 2022a. CogView2: faster and better text-to-image generation via hierarchical Transformers//Advances in Neural Information Processing Systems. New Orleans, USA: 35: 16890-16902.

Ding W P, Ming Y R, Cao Z H and Lin C T. 2022b. A generalized deep neural network approach for digital watermarking analysis. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(3): 613-627 [DOI: 10.1109/tetci.2021.3055520]

Ditria L and Drummond T. 2023. Hey that's mine imperceptible watermarks are preserved in diffusion generated outputs [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2308.11123.pdf

Fan L X, Ng K W and Chan C S. 2019. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 4714-4723

Fei J W, Xia Z H, Tondi B and Barni M. 2022. Supervised GAN watermarking for intellectual property protection//2022 IEEE International Workshop on Information Forensics and Security. Shanghai, China: IEEE: 1-6 [DOI: 10.1109/wifs55849.2022.9975409]

Fernandez P, Couairon G, Jégou H, Douze M and Furon T. 2023. The stable signature: rooting watermarks in latent diffusion models [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2303.15435.pdf

Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano A H, Chechik G and Cohen-Or D. 2022. An image is worth one word: personalizing text-to-image generation using textual inversion [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2208.01618.pdf

Girish S, Suri S, Rambhatla S and Shrivastava A. 2021. Towards discovery and attribution of open-world GAN generated images//Proceedings of 2021 IEEE/CVF International Conference on Computer

Vision. Montreal, Canada: IEEE: 14074-14083 [DOI: 10.1109/iccv48922.2021.01383]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Ho J and Salimans T. 2022. Classifier-free diffusion guidance [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2207.12598.pdf

Hu D H, Wang L, Jiang W J, Zheng S L and Li B. 2018. A novel image steganography method via deep convolutional generative adversarial networks. IEEE Access, 6: 38303-38314 [DOI: 10.1109/access.2018.2852771]

Kang M, Zhu J Y, Zhang R, Park J, Shechtman E, Paris S and Park T. 2023. Scaling up GANs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10124-10134 [DOI: 10.1109/cvpr52729.2023.00976]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/cvpr.2019.00453]

Kingma D P and Welling M. 2022. Auto-encoding variational bayes [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/1312.6114.pdf

Kumari N, Zhang B L, Zhang R, Shechtman E and Zhu J Y. 2023. Multi-concept customization of text-to-image diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1931-1941 [DOI: 10.1109/cvpr52729.2023.00192]

Li D X, Li J N and Hoi S C H. 2023a. BLIP-diffusion: pre-trained subject representation for controllable text-to-image generation and editing [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2305.14720.pdf

Li J N, Li D X, Savarese S and Hoi S. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2301.12597.pdf

Li X Y. 2023. DiffWA: diffusion models for watermark attack [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2306.12790.pdf

Li Y H, Liu H T, Wu Q Y, Mu F Z, Yang J W, Gao J F, Li C Y and Lee Y J. 2023c. GLIGEN: open-set grounded text-to-image generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22511-22521 [DOI: 10.1109/CVPR52729.2023.02156]

Liu A A, Zhang G K, Su Y T, Xu N, Zhang Y D and Wang L J. 2023. T2IW: joint text to image and watermark generation [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2309.03815.pdf

Ma Y H, Zhao Z Y, He X L, Li Z, Backes M and Zhang Y. 2023. Generative watermarking against unauthorized subject-driven image synthesis [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2306.07754.pdf

Marra F, Gragnaniello D, Verdoliva L and Poggi G. 2019. Do GANs leave artificial fingerprints?//Proceedings of 2019 IEEE Conference on Multimedia Information Processing and Retrieval. San Jose, USA: IEEE: 506-511 [DOI: 10.1109/MIPR.2019.00103]

Mou C, Wang X T, Xie L B, Wu Y Z, Zhang J, Qi Z A, Shan Y and Qie X H. 2023. T2I-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2302.08453.pdf

Nadimpalli A V and Rattani A. 2023. Proactive deepfake detection using GAN-based visible watermarking. ACM Transactions on Multimedia Computing, Communications, and Applications: #3625547 [DOI: 10.1145/3625547]

Nichol A and Dhariwal P. 2021. Improved denoising diffusion probabilistic models//Proceedings of the 38th International Conference on Machine Learning. Virtual Event, PMLR: 139: 8162-8171

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I and Chen M. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models//Proceedings of 2022 International Conference on Machine Learning. Baltimore, Maryland, USA: PMLR: 16784-16804

Ong D S, Chan C S, Ng K W, Fan L X and Yang Q. 2021. Protecting intellectual property of generative adversarial networks from ambiguity attacks//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3629-3638 [DOI: 10.1109/cvpr46437.2021.00363]

Qiao T, Ma Y Y, Zheng N, Wu H Z, Chen Y L, Xu M and Luo X Y. 2023. A novel model watermarking for protecting generative adversarial network. Computers and Security, 127: #103102 [DOI: 10.1016/j.cose.2023.103102]

Qiao T T, Zhang J, Xu D Q and Tao D C. 2019. MirrorGAN: learning text-to-image generation by redescription//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1505-1514 [DOI: 10.1109/CVPR.2019.00160]

Qin C, Zhang S, Yu N, Feng Y H, Yang X Y, Zhou Y B, Wang H, Neibles J C, Xiong C M, Savarese S, Ermon S, Fu Y and Xu R. 2023. UniControl: a unified diffusion model for controllable visual generation in the wild [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2305.11147.pdf

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y Q, Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. The Journal of Machine

Learning Research, 21(1): 5485-5551

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Virtual-only: PMLR: 8821-8831

Reed S, Akata Z, Mohan S, Tenka S, Schiele B and Lee H. 2016a. Learning what and where to draw//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 217-225

Reed S, Akata Z, Yan X C, Logeswaran L, Schiele B and Lee H. 2016b. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR: 1060-1069

Rolfe J T. 2017. Discrete variational autoencoders [EB/OL]. [2024-01-07]. https://arxiv.org/pdf/1609.02200.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/cvpr52688.2022.01042]

Ruiz N, Li Y Z, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22500-22510 [DOI: 10.1109/cvpr52729.2023.02155]

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Lopes R G, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Advances in Neural Information Processing Systems. New Orleans, USA: 35: 36479-36494.

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson, K, Schmidt L, Kaczmarczyk R and Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models//Advances in Neural Information Processing Systems. New Orleans, USA: 35: 25278-25294.

Sennrich R, Haddow B and Birch A. 2016. Neural machine translation of rare words with subword units [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/1508.07909.pdf

Shi C Y, Chen L, Wang C Y, Zhou X and Qin Z L. 2023. Review on image forensic techniques based on deep learning. Mathematics, 11: #3134 [DOI: 10.20944/preprints202306.1179.v1]

Sohl-Dickstein J, Weiss E, Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR: 2256-2265

Tao M, Bao B K, Tang H and Xu C S. 2023. GALIP: generative adversarial CLIPs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14214-14223 [DOI: 10.1109/cvpr52729.2023.01366]

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Q, Li S, Zhang X P and Feng G R. 2023. Rethinking neural style transfer: generating personalized and watermarked stylized images//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 6928-6937 [DOI: 10.1145/3581783.3612202]

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot... for now//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8692-8701 [DOI: 10.1109/cvpr42600.2020.00872]

Wu H Z, Liu G, Yao Y W and Zhang X P. 2021. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology, 31(7): 2591-2601 [DOI: 10.1109/tcsvt.2020.3030671]

Wu H Z, Zhang J, Li Y, Yin Z X, Zhang X P, Tian H, Li B, Zhang W M and Yu N H. 2023. Overview of artificial intelligence model watermarking. Journal of Image and Graphics, 28(6): 1792-1810 (吴汉舟, 张杰, 李越, 殷赵霞, 张新鹏, 田晖, 李斌, 张卫明, 俞能海. 2023. 人工智能模型水印研究进展. 中国图象图形学报, 28(6): 1792-1810) [DOI: 10.11834/jig.230010]

Wu W Y and Liu S S. 2023. A comprehensive review and systematic analysis of artificial intelligence regulation policies [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2307.12218.pdf

Xiong C, Qin C, Feng G R and Zhang X P. 2023. Flexible and secure watermarking for latent diffusion model//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 1668-1676 [DOI: 10.1145/3581783.3612448]

Xu T, Zhang P C, Huang Q Y, Zhang H, Gan Z, Huang X L and He X D. 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1316-1324 [DOI: 10.1109/cvpr.2018.00143]

Yang T Y, Wang D D, Tang F, Zhao X Y, Cao J and Tang S. 2023. Progressive open space expansion for open-set model attribution//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15856-15865 [DOI: 10.1109/cvpr52729.2023.01522]

Yin Z X, Yin H and Zhang X P. 2022. Neural network fragile watermarking with no model performance degradation//Proceedings of 2022 IEEE International Conference on Image Processing. Bordeaux, France: IEEE: 3958-3962 [DOI: 10.1109/ICIP46576.2022.9897413]

Yu F, Seff A, Zhang Y D, Song S R, Funkhouser T and Xiao J X. 2016. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/1506.03365.pdf

Yu J H, Li X, Koh J Y, Zhang H, Pang R M, Qin J, Ku A, Xu Y Z, Baldridge J and Wu Y H. 2022a. Vector-quantized image modeling with improved VQGAN [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2110.04627.pdf

Yu J H, Xu Y Z, Koh J Y, Luong T, Baid G, Wang Z R, Vasudevan V, Ku A, Yang Y F, Ayan B K, Hutchinson B, Han W, Parekh Z, Li X, Zhang H, Baldridge J and Wu Y H. 2022b. Scaling autoregressive models for content-rich text-to-image generation [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2206.10789.pdf

Yu N, Davis L and Fritz M. 2019. Attributing fake images to GANs: Learning and analyzing GAN fingerprints//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7555-7565 [DOI: 10.1109/iccv.2019.00765]

Yu N, Skripniuk V, Abdelnabi S and Fritz M. 2021. Artificial fingerprinting for generative models: rooting deepfake attribution in training data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14428-14437 [DOI: 10.1109/iccv48922.2021.01418]

Yu N, Skripniuk V, Chen D F, Davis L and Fritz M. 2022c. Responsible disclosure of generative models using scalable fingerprinting [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2012.08726.pdf

Zeng Y W, Tan J X, You Z X, Qian Z X and Zhang X P. 2023. Watermarks for generative adversarial network based on steganographic invisible backdoor//Proceedings of 2023 IEEE International Conference on Multimedia and Expo. Brisbane, Australia: IEEE: 1211-1216 [DOI: 10.1109/icme55011.2023.00211]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5908-5916 [DOI: 10.1109/iccv.2017.629]

Zhang K A, Xu L, Cuesta-Infante A and Veeramachaneni K. 2019. Robust invisible video watermarking with attention [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/1909.01285.pdf

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824 [DOI: 10.1109/iccv51070.2023.00355]

Zhao X D, Zhang K X, Su Z H, Vasan S, Grishchenko I, Kruegel C, Vigna G, Wang Y X and Li L. 2023a. Invisible image watermarks are provably removable using generative AI [EB/OL]. [2023-11-05]. https://arxiv.org/pdf/2306.01953.pdf

Zhao Y, Liu B, Ding M, Liu B P, Zhu T Q and Yu X. 2023b. Proactive deepfake defence via identity watermarking//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 4591-4600 [DOI: 10.1109/wacv56688.2023.00458]

作者简介

刘安安，男，教授，主要研究方向为多媒体内容理解与安全。E-mail：anan0422@gmail.com

苏育挺，通信作者，男，教授，主要研究方向为多媒体内容理解与安全。E-mail：ytsu@tju.edu

王岚君，女，研究员，主要研究方向为多媒体内容理解与安全。E-mail：wang.lanjun@outlook.com

李斌，男，教授，主要研究方向为图像隐写。E-mail：libin@szu.edu.cn

钱振兴，男，教授，主要研究方向为多媒体与人工智能安全。E-mail：zxqian@fudan.edu.cn

张卫明，男，教授，主要研究方向为多媒体与人工智能安全。E-mail：zhangwm@ustc.edu.cn

周琳娜，女，教授，主要研究方向为多媒体与人工智能安全。E-mail：zhoulinna@tsinghua.edu.cn

张新鹏，男，教授，主要研究方向为多媒体与人工智能安全。E-mail：zhangxinpeng@fudan.edu.cn

张勇东，男，教授，主要研究方向为多媒体内容理解与安全。E-mail：zhyd73@ustc.edu.cn

黄继武，男，教授，主要研究方向为多媒体与人工智能安全。E-mail：jwhuang@szu.edu.cn

俞能海，男，教授，主要研究方向为多媒体与人工智能安全。E-mail：ynh@ustc.edu.cn