Paintings Generated by Text-to-Image System

Subjects: Art

Contributor: Yanru Lyu, Xinxin Wang, Rungtai Lin, Jun Wu

Since the generative adversarial network (GAN) portrait painting titled “Edmond de Belamy” was constructed in 2018, AI art has already entered the public’s vision. One of the latest applications of AI is the generation of images based on natural language descriptions, which enhances the efficiency and effect of the transformation from creativity to visuality to a great extent. In the past, whether in traditional or digital painting creation, the author needed to be skilled in using tools and to have rich technical experience to accurately map the brain’s imagination to the visual layer. However, in co-creation with text-to-image AI generators, both artists and nonartists can input the text description to produce many high-quality images. During traditional painting creation, artists and nonartists in a painting task indicated quantitative and qualitative differences in some studies, such as artists spending more time on planning their painting, having more control over their creative processes, having more specific skills, and having more efficiency than nonartists. Whether such differences still exist in the new human–AI interaction mode and what new changes arise are worth discussing.

Keywords: AI painting ; human–AI interaction ; artistic perception

1. Introduction

A series of text-to-image AI systems, such as Disco Diffusion [1], Midjourney [2], Stable Diffusion [3], OpenAI’s DALL-E 2 [4], and Google’s Imagen [5], is making a big splash. The generation mechanism is to use a language–vision model to understand the “prompt” input by users, and then the generator is guided to produce high-quality images. They are capable of synthesizing images with any style and content based on a prompt. Besides, users can control the system to iterate more variations. With the rise of AI art, many artists have also started to use AI to assist in creation. According to the Colorado State Fair competition’s website [6], the art piece “Théâtre D’opéra Spatial,” which was generated by Midjourney, won first place in the digital art category. As the formation of generators using natural language text to create various styles of creative images occurs, the question that arises immediately is: what is the essence of artistic creation, and what is the core capability of artists? Though everyone thought art was one thing robots could never do, maybe we will face the challenges of emerging AI technology.

2. Text-to-Image Systems

With the successful application of transformer-based architectures in neural language processing (NLP), text-to-image systems based on deep generative models have become popular means for computer vision tasks [7][8]. They generate creative images combining concepts, attributes, and styles from expressive text descriptions [9]. The primary generation mechanism is that a language–vision model (i.e., CLIP) is adopted to guide the generator to produce high-quality images.

When OpenAI released CLIP in 2021 [10], it spurred immense technical progress in text-to-image generation. CLIP is a pre-trained language–vision model that enables zero-shot image manipulation guided by text prompts. Unlike traditional representation learning that is based mostly on discretized labels, the vision–language model aligns images and texts in a common feature space, allowing zero-shot transfer to a downstream task via prompting [11]. CLIP guides the generator to synthesize digital images when used as a discriminator in a generative system. Using its joint text–image representation space, people can control the synthesis process with natural language. At present, most programs use CLIP for text encodings, such as DALL-E 2 and Stable Diffusion. Differently, Google’s Imagen uses the T5-XXL language model to encode the text and then generate images directly without learning the priori model [5]. The text input, known as the prompt, plays a crucial role in downstream datasets. It is an important aspect for improving the quality and changing the aesthetics of images, which entails the practice and capabilities of interacting with the system. The term prompt engineering knows the practice and skill of writing prompts due to its iterative and experimental nature [12]. However, identifying the right prompt is a nontrivial task which often takes a significant amount of time for word tuning—a slight change in wording could make a huge difference in performance [11].

Currently, text-to-image generation models can be divided into two designs: sequence-to-sequence modeling and diffusion-based modeling [13]. The main idea of the sequence-to-sequence modeling design is to turn images into discrete image tokens via leveraging transformer-based image tokenizers and to employ the sequence-to-sequence architectures to learn the relationship between textual input and visual output from a large collection of text–image pairs, such as Vector Quantized Variational Autoencoder (VQ-VAE) and Vector Quantized Generative Adversarial Networks (VQ-GAN). VQ-VAE discretely incorporates ideas from vector quantization and encoder network outputs. Then, by pairing these representations with an autoregressive prior, the model with a PixelCNN decoder can generate high-quality images [14]. This model is used by the first vision of DALL-E [15]. More like a variant, VQ-GAN represents a variety of modalities with discrete latent representations by building a codebook vocabulary with a finite set of learned embeddings and using Transformer instead of the PixelCNN in VQ-VAE [8]. Anyway, the PatchGAN discriminator is used to add anti-loss in the training process. The representative work of this modeling is Parti [16]. Different from the above idea, the diffusion-based models, which are built from a hierarchy of denoising autoencoders, start from random noise and gradually denoise them, conditioned on textual descriptions, until images matching the conditional information are generated [17]. Based on the power of diffusion models in high-fidelity image synthesis, the text-to-image system is significantly pushed forward by the recent effort of Disco Diffusion [1], Midjourney [2], Stable Diffusion [3], DALL-E 2 [4], and Imagen [5].

At present, the programs that use diffusion models for a better generation effect, Disco Diffusion, Midjourney, Stable Diffusion, and DALL-E 2, are open to the public, but the programs of Imagen are not. Disco Diffusion is a clip-guided diffusion model that is good at generating pretty abstract art, which can be run in Google Colab now [1]. Midjourney was created by an independent research lab with the same name. It is currently in open beta and is accessible on Discord, where users type in the textual prompt in the chat, and then the artwork is generated by the AI system [2]. Stable diffusion was released by Stability AI in 2022, which uses a latent diffusion mode trained on 512 × 512 images from a subset of the LAION-5B database. Similar to Google's Imagen, this model uses a frozen CLIP ViT-L/14 text encoder to condition the model to text prompts [18]. Furthermore, it has a better balance between speed and quality and can generate images within seconds [3]. The main novelty of DALL-E 2 seems to be an extra layer of indirection with the prior network, which predicts an image embedding based on the text embedding from CLIP. Specifically, this repository will only build out the diffusion prior network, as it is the best-performing variant [4].

With the emergence of such open-source implementations, the use of advanced text-to-image synthesis for generating images is becoming more widespread, which represents a relevant trend in the AI Art community [19].

3. Communication between Artists and Audiences

Artistic creation is a process for artists to explore and express ideas and concepts. A great painting has much more below the surface than is first seen on the surface. Therefore, it must access the mind as well as the senses [20]. Similar to how humans do not really know how they breathe, artists do not truly know how they create: while they may rely on a set of fundamental principles, such as how to arrange elements, light, colors, and other components, most of their creative decisions happen intuitively [21]. The experimental result of Eindhoven and Vinacke demonstrated that artists have more control over their creative activities and produce better results than nonartists in the creative process of painting [22]. Kay also found that nonartists, semiprofessional artists, and professional artists differed on certain process-related variables [23].

The interplay between the internal (cognitive) representation and the external (physical) representation is a fascinating problem in cognitive psychology, art, science, and philosophy [24]. The various painting attributes, such as colors, shapes, and boundaries, are selectively redistributed to the brain for processing. For example, color may be experienced as warm or cold or as cheerful or somber [25]. Audiences can also perceive the painter's actions by observing the brushstroke of the painting [26]. Apart from that, from a psychological viewpoint, Kozbelt examined various experiments on artists' perception and depiction skills and showed evidence suggesting possible perceptual differences between artists and nonartists [27][28]. Aesthetic appreciation is an active process influenced by several objective features: external and subjective factors that engage both bottom-up and top-down processes [29]. In the series of studies on experimental aesthetics by Lyu et al. [30][31][32], the perception of artistic style was affected by individual attributes such as knowledge background and gender. Thus, the perception of art is a complex interaction process between the top and bottom levels, which is affected by various subjective and objective factors.

According to communication theory, the process of artist expression is called encoding, and the way the artwork is perceived by the audience is regarded as decoding [33][34]. Jakobson proposed six constitutive factors with six functions in communication: the addresser, addressee, context, message, contact, and code [34]. For example, an artist (addresser) sends a message to an audience (addressee) through his/her painting. The artist's work, as the message with a story

(context), plays a role in the connection between himself/herself and the audience (contact). Finally, his/her message must be based on a shared meaning system (code) by which his/her work is structured [20]. There are three levels of problems, namely technical, semantic, and effectiveness levels, that were identified in the study on the communication of paintings [31][35]. Among them, the technical level focuses on letting the addressee receive a message through visual attraction, and the semantic level requires that the addressee is allowed to understand the message's meaning without misinterpreting it. The effectiveness level concerns the effect of the audience's feelings. During the creative process of AI art, the artists choose AI algorithms according to their intentions for creating the artwork, and audience acceptance is a critical defining step in deciding whether it is "art" [36]. Studying the process of art perception can help build a bridge between artists and the audience [37][38].

4. Artworks Generated by Human–AI Co-Creation

Artworks are increasingly being created by machines through algorithms with little or no input from humans. At the Christie's auction in 2018, the portrait "Edmond de Belamy", generated by generative adversarial networks (GAN), was auctioned for $432,500, which indicates that AI has begun to enter the field of vision at a rapid speed [39]. Recent works have addressed a variety of tasks such as classification, object detection, similarity retrieval, multimodal representations, and computational aesthetics, among others [19]. The neural style transfer in which AI technology first intervened in the field of art has been widely used in the platforms such as Prisma, Deep Dream Generator, and other art content production platforms. In 2022, text-to-image AI art generators are much more popular and have been applied to creating conceptual scenes, creative designs, and fictional illustrations. In this case, it can be seen that the processes in various art creations are changing. Meanwhile, some new jobs have also been immediately emerging, such as prompt sale [40].

With the explosion of AI-related technologies and their continuous application in the field of art, there is a growing body of research initiatives and creative applications arising at the intersection of AI and art. Artistic creation is embedded with cultural, historical, and institutional frameworks that directly interact with the artist's own creative process [21]. Lacking human consciousness, AI does not understand what it is doing and is merely a suite of statistical models calculating favorable odds through enormous variations. Considering that, AI cannot create art, but it can create patterns that an audience will likely perceive as art [41]. The human artist, as the author, is always the mastermind behind the work, and the computer is a tool [42]. However, AI technology is not like traditional tools. Its randomness changes the way humans control it. As a sparking trigger of inspiration, artists collaborate with AI agencies to augment the artistic process [41].

As for text-based generative art, it is also argued that creativity does not lie in the final artifact but rather in the interaction with the AI and the practices that may arise from the human–AI interaction [43]. It is not hard to imagine a future where text prompts could be generated by language models, thereby completely dehumanizing the creative artistic process and severely distorting the human perception of the meaning behind an image [44]. Most studies reported that visual artworks can be recognized to some extent by humans, especially by experts of a specific art field [45][46], but other experimental results showed that individuals are unable to accurately identify AI-generated artwork [32][47]. Based on the researchers' previous research, the deep learning model, trained by large amounts of data on paintings, can simulate human painting skills on the technical level. In contrast, people prefer paintings connecting the semantic and emotional levels [31].

References

1. Disco Diffusion. Available online: https://github.com/alembics/disco-diffusion (accessed on 10 June 2022).
2. Midjourney. Available online: www.midjourney.com (accessed on 25 August 2022).
3. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
4. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv 2022, arXiv:2204.06125.
5. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; et al. Photorealistic Text.-to-Image Diffusion Models with Deep Language Understanding. arXiv 2022, arXiv:2205.11487, 2022.
6. State Fair's Website. Available online: https://coloradostatefair.com/wp-content/uploads/2022/08/2022-Fine-Arts-First-Second-Third.pdf (accessed on 25 August 2022).

7. Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10696–10706.
8. Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; Raff, E. Vqgan-clip: Open domain image generation and editing with natural language guidance. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 88–105.
9. Lee, H.; Ullah, U.; Lee, J.S.; Jeong, B.; Choi, H.C. A Brief Survey of text driven image generation and maniulation. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Gangneung, Korea, 1–3 November 2021; pp. 1–4.
10. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
11. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. Int. J. Comput. Vis. 2022, 130, 2337–2348.
12. Liu, V.; Chilton, L.B. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–23.
13. Wu, Y.; Yu, N.; Li, Z.; Backes, M.; Zhang, Y. Membership Inference Attacks Against Text-to-image Generation Models. arXiv 2022, arXiv:2210.00968.
14. Van Den Oord, A.; Vinyals, O. Neural discrete representation learning. In Proceedings of the Neural Information Processing Systems Annual Conference, Long Beach, CA, USA, 4–9 December 2017; pp. 1–10.
15. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8821–8831.
16. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv 2022, arXiv:2206.10789.
17. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2256–2265.
18. Stable-Diffusion. Available online: https://github.com/CompVis/stable-diffusion (accessed on 2 September 2022).
19. Cetinic, E.; She, J. Understanding and creating art with AI: Review and outlook. ACM Trans. Multimed. Comput. Commun. Appl. 2022, 18, 1–22.
20. Lin, C.L.; Chen, J.L.; Chen, S.J.; Lin, R. The cognition of turning poetry into painting. J. US-China Educ. Rev. B 2015, 5, 471–487.
21. Audry, S. Art in the Age of Machine Learning; MIT Press: Cambridge, MA, USA, 2021; pp. 30, 158–165.
22. Eindhoven, J.E.; Vinacke, W.E. Creative processes in painting. J. Gen. Psychol. 1952, 47, 139–164.
23. Kay, S. The figural problem solving and problem finding of professional and semiprofessional artists and nonartists. Creat. Res. J. 1991, 4, 233–252.
24. Solso, R.L. Cognition and the Visual Arts; MIT Press: Cambridge, MA, USA, 1996; pp. 34–36.
25. Steenberg, E. Visual Aesthetic Experience. J. Aesthet. Educ. 2007, 41, 89–94.
26. Taylor, J.; Witt, J.; Grimaldi, P. Uncovering the connection between artist and audience: Viewing painted brushstrokes evokes corresponding action representations in the observer. J. Cogn. 2012, 125, 26–36.
27. Kozbelt, A. Gombrich, Galenson, and beyond: Integrating case study and typological frameworks in the study of creative individuals. Empir. Stud. Arts 2008, 26, 51–68.
28. Kozbelt, A.; Ostrofsky, J. Expertise in drawing. In The Cambridge Handbook of Expertise and Expert Performance; Ericsson, K.A., Hoffman, R.R., Kozbelt, A., Eds.; Cambridge University Press: Cambridge, UK, 2018; pp. 576–596.
29. Chiarella, S.G.; Torromino, G.; Gagliardi, D.M.; Rossi, D.; Babiloni, F.; Cartocci, G. Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. Comput. Hum. Behav. 2022, 137, 107406.
30. Lyu, Y. A Study on Perception of Artistic Style Tansfer using Artificial Intelligance Technology. Unpublished Doctor's Thesis, National Taiwan University, Taipei, Taiwan. 2022. Available online: https://hdl.handle.net/11296/grdz93 (accessed on 23 October 2022).

31. Lyu, Y.; Lin, C.-L.; Lin, P.-H.; Lin, R. The Cognition of Audience to Artistic Style Transfer. Appl. Sci. 2021, 11, 3290.
32. Sun, Y.; Yang, C.H.; Lyu, Y.; Lin, R. From Pigments to Pixels: A Comparison of Human and AI Painting. Appl. Sci. 2022, 12, 3724.
33. Fiske, J. Introduction to Communication Studies, 3rd ed.; Routledge: London, UK, 2010; pp. 5–6.
34. Jakobson, R. Language in literature; Harvard University Press: Cambridge, MA, USA, 1987; pp. 100–101.
35. Lin, R.; Qian, F.; Wu, J.; Fang, W.-T.; Jin, Y. A Pilot Study of Communication Matrix for Evaluating Artworks. In Proceedings of the International Conference on Cross-Cultural Design, Vancouver, BC, Canada, 9–14 July 2017; pp. 356–368.
36. Mazzone, M.; Elgammal, A. Art, creativity, and the potential of artificial intelligence. Arts 2019, 8, 26.
37. Gao, Y.-J.; Chen, L.-Y.; Lee, S.; Lin, R.; Jin, Y. A study of communication in turning “poetry” into “painting”. In Proceedings of the International Conference on Cross-Cultural Design, Vancouver, BC, Canada, 9–14 July 2017; pp. 37–48.
38. Gao, Y.; Wu, J.; Lee, S.; Lin, R. Communication Between Artist and Audience: A Case Study of Creation Journey. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, FL, USA, 26–31 July 2019; pp. 33–44.
39. Yu, Y.; Binghong, Z.; Fei, G.; Jiaxin, T. Research on Artificial Intelligence in the Field of Art Design Under the Background of Convergence Media. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Ulaanbaatar, Mongolia, 10–13 September 2020; p. 012027.
40. Promptbase. Available online: https://promptbase.com/ (accessed on 25 August 2022).
41. Hageback, N.; Hedblom, D. AI FOR ARTS; CRC Press: Boca Raton, FL, USA, 2021; p. 67.
42. Hertzmann, A. Can Computers Create Art? Arts 2018, 7, 18.
43. Oppenlaender, J. Prompt Engineering for Text-Based Generative Art. arXiv 2022, arXiv:2204.13988.
44. Ghosh, A.; Fossas, G. Can There be Art Without an Artist? arXiv 2022, arXiv:2209.07667.
45. Chamberlain, R.; Mullin, C.; Scheerlinck, B.; Wagemans, J. Putting the art in artificial: Aesthetic responses to computer-generated art. Psychol. Aesthet. Crea. 2018, 12, 177.
46. Hong, J.-W.; Curran, N.M. Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. ACM Trans. Multimed. Comput. Commun. Appl. 2019, 15, 1–16.
47. Gangadharbatla, H. The role of AI attribution knowledge in the evaluation of artwork. Empir. Stud. Arts 2022, 40, 125–142.
