Advancements and technical challenges in generative models of artificial intelligence painting

Xuan Chen

College of Software Engineering, Sichuan University, Chengdu, 610200, China

xu_anc@163.com

Abstract. Artificial Intelligence Painting, as a frontier field of AI technology application, has gained wide attention and rapid development in recent years. With the advancement of generative models, especially the emergence of variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, the ability of AI in art creation has significantly increased. These techniques not only enable computers to generate realistic images, but also to simulate complex artistic styles, opening up new possibilities for artistic creation. Nevertheless, how to maintain a high degree of complexity and realism in the generated images, how to improve the data quality to enhance the generalization ability of the model, how to solve the pattern collapse problem to improve the stability of the generation, and how to improve the efficiency while ensuring the quality of the generation are all key issues in the current research. In addition, with the continuous evolution of AI technology, the future development direction and potential applications are also worth exploring in depth. The paper aims to systematically analyze the generative models and their applications in AI painting, to deeply discuss the technical challenges faced, and to look forward to the future development prospects.

Keywords: AI Painting, Generative Adversarial Networks,Variational Autoencoder, Diffusion Models.

1. Introduction

Artificial Intelligence (AI) painting is a part of computational art that expands and extends traditional art practices. Within the realm of deep learning research, related studies have made persistent efforts to develop image generation techniques, proposing diverse generative models and enhancement strategies with the objective of enabling machines to generate high-quality images in a creative manner. VAEs, GANs, and diffusion models that have gained much attention in recent years, have gradually improved the expected results of image generation. These technologies have undergone a gradual maturation process, with applications in areas such as image generation, style transfer, restoration, and enhancement, resulting in notable advancements. Concurrently, a series of AI painting platforms accessible to the general public have emerged, attracting considerable attention around 2022. As interest in AI painting has grown, the promotion of this field has encountered obstacles, which have been sparked by controversies rooted in technical limitations, copyright issues, and the public's relatively low level of understanding. This paper aims to provide an overview of the development of AI painting, to examine the main generative models, and to suggest ways of addressing the challenges that have arisen in the recent history of AI painting.

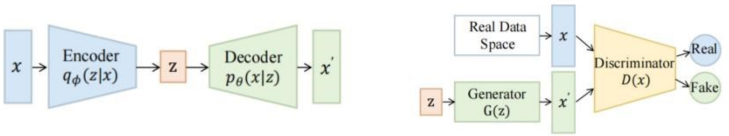
© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

2. Evolution of Artificial Intelligence Painting

In its infancy during the 1970s, Harold Cohen's AARON painting machine served as a pioneer in AI art, learning Cohen's drawing techniques and generating ever-changing images. Despite its simplicity, the device demonstrated autonomy, thereby bridging the gap between computer art and AI art. In 2006, Simon Colton's The Painting Fool advanced beyond the limitations of AARON's constrained style, striving for diversity, creativity, and independent judgment [1] . Advances in The Painting Fool can be attributed to its capacity to engage with both external and self-generated art forms, thereby introducing a "humanistic" dimension to the domain of machine painting. In the exploratory phase, AI painting concentrated on deep learning models for automated image creation, initiated by Ng and Dean's research on generating cat and human faces in 2012 [2]. This was further developed by research on unsupervised learning in 2013, thereby paving the way for AI-driven art [3] In 2014, VAEs and GANs emerged, and the GANs' adversarial framework significantly advanced image generation. GANs became the foundation for AI painting and evolved into StyleGANs, CycleGANs, BigGANs, and others. CAN is a noteworthy model derived from GANs that has had a significant impact on the development of AI painting. Its basic concept is an adversarial process between two networks, which limits creativity due to the necessity of maintaining verisimilitude. CAN, proposed in 2017, introduced an artistic style discriminator to facilitate creativity in GANs. In 2015, Google introduced the open source software DeepDream to shift the focus of neural networks from recognition to creation. This approach involved amplifying image features to demonstrate the potential of neural networks [4]. In the development phase, diffusion modeling has rapidly gained popularity since the breakthrough of DDPM in 2020 and has become a mainstay of image generation. In 2021, OpenAI's DALL-E fused diffusion, morphing, and CLIP, triggering text-to-image generation. In 2022, AI painting software such as disco diffusion augmented creativity with CLIP, and spawned platforms such as Midjourney and DALL-E 2. In July 2022, stabilized diffusion latent space reduction techniques led to memory and computation reductions, solving the problem of diffusion's resource requirements.

3. Analysis and Application of Generative Models

Three representative AI painting generation models are briefly introduced: Variational autoencoder (VAE), generative adversarial network (GAN), and diffusion model, as shown in Figure 1.



(a) (b) (c)

Figure 1. Three Vision Generative Models (a. Variational Autoencoder (VAE), b. Generative Adversarial Network (GAN), and c. Diffusion Model) [5]

3.1. Variational Autoencoder (VAEs)

Variational Autoencoder (VAE) is a deep generative model that learns latent representations of data to generate new samples. It utilizes Bayesian variational inference with an encoder (mapping input to latent space) and a decoder (mapping latent to data space). The VAE minimizes reconstruction error and divergence between latent and prior distributions. With a learning latent space characterized by structural features, VAE exhibits a high degree of flexibility and robust control in generating new data. As a probabilistic generative model, VAE not only provides probabilistic explanations for the generated samples, but also learns the latent representations of the data and imposes probabilistic constraints to generate diverse data samples [6]. In a variety of domains, including image and text generation, VAE has exhibited robust capabilities, generating data samples that closely resemble the original data, either visually or semantically. However, it is important to consider the limitations of the model, including the need to carefully design the dimension and complexity of the latent space to optimize model performance; the relative complexity of the training process and the need to strike a balance between reconstruction error and latent space probability distributions; and the fact that, in some cases, the data samples generated may lack sufficient detail or diversity.

Various types of images such as faces, landscapes, and animals can be generated by training VAE models, and these generated images have high visual fidelity and diversity. VAE-generated images can be used for virtual scene construction in the entertainment industry and character design in games, movies, and other media. In areas where data is scarce or difficult to obtain, such as medical image analysis and automated driving, VAE can generate new data samples that are similar to real data, thereby increasing the size of the dataset and enhancing the generalization ability of the model. The generation of disparate data samples by VAE can enhance the predictive performance of the model on unknown data. The latent space of a VAE offers a comprehensive view of the underlying data representations. The analysis and visualization of the latent space allow for the discernment of the intrinsic relationships and structures between data. This not only facilitates a more profound comprehension of the intrinsic characteristics of the data set, but also offers invaluable assistance in subsequent data analysis and processing operations.

3.2. Generative Adversarial Networks (GANs)

Generative Adversarial Network is a deep generative model with a generator and discriminator. Inspired by Nash equilibrium, the generator mimics real data distribution, while the discriminator identifies data sources (real or generated). To reach the equilibrium state of the game, these two actors continuously perform optimization iterations to seek the Nash equilibrium point [7]. GANs can generate high-quality and diverse data samples, and show strong applicability in a variety of domains, including data generation, image restoration, and style transformation, greatly advancing the fields of computer vision and natural language processing. Notably unique, GANs do not require explicit modeling of data distributions but implicitly learn them through the ingenious design of an adversarial process, conferring greater flexibility on their training procedures. Furthermore, this model facilitate end-to-end training, further simplifying the complexity of model training. However, GANs present many challenges, including the complexity and time-consuming nature of their training process, which requires significant computational resources. Generated data samples may exhibit unnatural artifacts or distortions, and problems such as pattern crashes and instability during the training process. These issues indicate potential avenues for future research and improvements.

The training of GANs models enables the generation of a diverse range of images, including faces, landscapes, objects, and more. The generated images exhibit high visual fidelity and can be utilized for virtual scene construction in the entertainment industry and character design in games, movies, and other media. In the context of image inpainting, GANs are particularly adept at automatically filling in missing or damaged portions of an image. By analyzing and learning from the surrounding pixels, GANs are capable of generating content that is stylistically consistent with the original image, thereby enhancing the quality and visual appeal of the image. Moreover, GANs are capable of performing style transformation tasks, which entail applying the style of one image to another. By training specific GAN models, conversions between different styles can be achieved, such as applying an oil painting style to a sketch or a watercolor effect to a photograph. In domains where data is scarce or challenging to procure, such as medical image analysis and autonomous driving, GANs can generate novel data samples that are analogous to real data, thereby augmenting the size of the dataset and enhancing the model's capacity for generalization.

3.3. Diffusion Models

Diffusion models, probabilistic generative models rooted in physics' diffusion theory, have been innovatively applied in ML for data generation, particularly image, sound, and text synthesis. They emerge as a flexible, accurate image generation class, surpassing GANs in architecture and log-likelihood computation. As shown in Figure 2, these models contain two key processes: forward diffusion and backward diffusion. In the forward diffusion phase, the training data is progressively corrupted by adding Gaussian noise to the original data, gradually erasing the image details and transforming it into pure noise. Subsequently, the neural network is trained for the reverse process, which consists of gradually denoising and reconstructing the image, transforming the pure noise into a high-quality image.

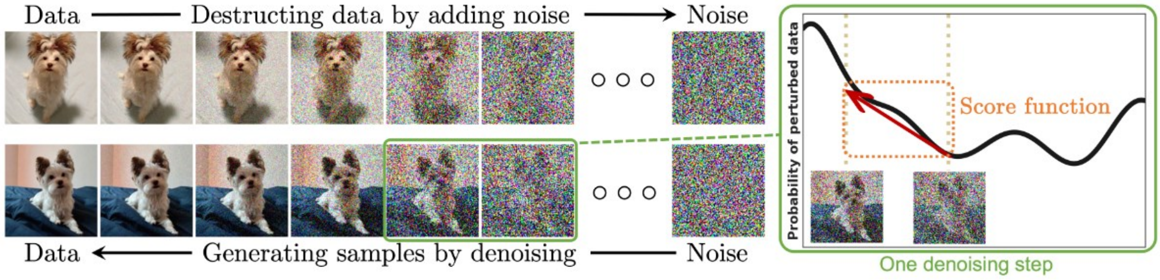

Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

Figure 2. Process of Diffusion Models [8]

Diffusion models generate high-quality samples that approximate any complex data distribution with good consistency, and show great potential for applications in many fields such as image generation, video generation, and natural language processing. They have a unique advantage over other generative models such as GAN and VAE by having more stable training targets and better generation results [9]. However, such models also face challenges, including high computational costs associated with repetitive inference evaluation during the sampling process, as well as potential problems such as instability, high-dimensional computational requirements, and complex likelihood optimization.

Diffusion models can generate high-quality image samples and support complex image editing tasks. For example, modifying textual descriptions or image features can result in the generation of images that satisfy particular requirements or facilitate image-to-image conversion. In addition, diffusion modeling has been applied to areas such as image super-resolution and image inpainting to improve the efficiency and effectiveness of image processing [10]. Video generation represents an important extension of the diffusion model beyond image generation. The introduction of a temporal dimension enables the diffusion model to generate continuous video sequences, which is a crucial aspect for the creation of dynamic content and the development of virtual and augmented reality applications. A number of diffusion model-based video generation methods have been proposed, including Make-A-Video and AnimatedDiff.

4. Key Technical Challenges Facing Artificial Intelligence Painting

4.1. Complexity and Authenticity of Images

AI painting, though adept at high-quality images, faces criticism for unrealistic human hands, revealing a gap between AI and human artistry. Its lack of 3D object comprehension limits rendering accuracy, light/shadow effects, and deformations. Limited training data for complex objects exacerbates this, impacting hand drawings. Solutions involve expanding datasets, incorporating human feedback, and enhancing AI's understanding of real-world objects. Current models struggle with detail processing, producing blurred or unrealistic images. Improving realism and detail in generated images is a crucial research direction.

4.2. Model Complexity and Computational Resources

Generative models, especially deep learning models, often require complex network structures and substantial computational resources. For example, models such as GANs and VAEs consume a lot of computational time and resources during the training process. With the increase of model complexity, the demand for computational resources also increases exponentially, which puts higher requirements for hardware and algorithm optimization. Meanwhile, current AI painting models usually require a large amount of training data, which not only reduces the efficiency of model training, leads to the instability of AI painting images, but also limits their performance in some specific styles or themes. Therefore, how to reduce the dependence on large amounts of data is a challenge[11].

4.3. Data Quality and Model Generalization

The effectiveness of the generative model depends greatly on the quality and diversity of the training data. If the training data are insufficient or biased, the generative model may produce wrong or duplicate images. It is therefore important to consider how to obtain high-quality and diverse training data and how to effectively utilize these data in the model training process, as this represents a significant technical challenge. The quality of data affects the model generalization ability to a certain extent. Generative models need to have strong generalization capabilities that enable them to generate unseen diverse images. However, current generative models still have some limitations in terms of generalization capabilities[12]. For example, they may be limited to generating images that are similar to the training data, while failing to generate completely novel and unique works.

4.4. Pattern Collapse and Stability

During the training process, the generative model may encounter the issue of pattern collapse, whereby the model is only capable of generating a restricted range of images that do not fully encompass the data distribution. Furthermore, the training process of the generative model may be unstable, resulting in significant fluctuations in the quality of the generated images. Consequently, current generative models still exhibit certain deficiencies in terms of controllability. For instance, they may not accurately capture the user's intent or requirements, leading to generated images that fail to meet expectations.

4.5. Generation Speed and Efficiency

In practical applications, generative models need to produce high-quality images within a relatively short period. However, current generative models still face bottlenecks in terms of generation speed. To enhance generation speed, optimizations and accelerations of the models are necessary. Commonly used methods include model simplification, where simpler model architectures such as MobileNet or ShuffleNet are utilized to reduce the number of parameters and computational complexity[13], thereby increasing the generation speed; knowledge refinement, where smaller “student” models are trained to mimic the behavior of a larger “teacher” model, reducing model size and computational requirements; application of quantization techniques and low-level training to reduce model size and computational load; and the use of pruning techniques to eliminate unimportant weights from the model, reducing computational requirements.

5. Future Prospects

The AI painting has been widely used in a variety of fields, including poster generation and citation, fashion design, video editing, and architectural design, highlighting the versatility and potential of AI-driven image creation to transform the creative industries. To further this momentum and enhance the fidelity of image content, future research efforts must move in ambitious directions beyond current AI drawing capabilities. One key research direction is to empower AI painting to understand the complex physical principles of real-world objects, requiring the development of advanced algorithms capable of understanding properties such as light, reflection, refraction, and material behavior. By incorporating these principles into the generation process, AI paintings can produce images that more accurately mimic the visual complexity and fidelity of the real world, thereby improving the realism and applicability of the images in a wider range of scenarios. Another significant research area is the enhancement of AI-generated artwork's capacity to discern and reflect emotional nuances from textual or real-world sources. Emotion is a fundamental aspect of the human experience and plays a pivotal role in the development of artistic expression. By employing natural language processing techniques and machine learning models trained on extensive emotion datasets, artificially intelligent paintings can be endowed with the capacity to discern emotional subtleties in textual cues or real-world observations and translate them into visually evocative images. Such emotional intelligence not only enhances the emotional depth of the resulting artwork but also broadens its appeal and relevance to users seeking to convey or evoke a specific emotional response.

6. Conclusion

Artificial intelligence (AI) painting has been in its infancy since the mid-to-late 20th century, continuously undergoing extensive research and exploration. In recent years, AI painting technology has seen rapid advancements, accompanied by the emergence of related tools and platforms into the public eye, garnering widespread attention. This paper briefly outlines three prevalent AI painting generation models, among which the diffusion model stands out for its superior accuracy and speed in image generation, thus becoming a new direction for current research and application. However, the development of AI painting still faces challenges such as reducing reliance on training data and improving model stability. Future research should focus on model optimization and expanding market applications of AI painting to ensure a balance between technological development and application promotion.

References

[1] Colton, S. (2012) The Painting Fool: Stories from Building an Automated Painter. In: McCormack, J., d'Inverno, M. (eds) Computers and Creativity. Springer, Berlin, Heidelberg.

[2] Dean, J., Corrado, G, et al. (2012) Large Scale Distributed Deep Networks[C]//Advances in Neural Information Processing Systems. Curran Associates, Inc.

[3] Le, Q.V. (2013) Building high-level features using large scale unsupervised learning. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada, IEEE, 8595–8598.

[4] Suzuki, K., et al. (2017) A Deep-Dream Virtual Reality Platform for Studying Altered Perceptual Phenomenology. Scientific Reports, 7(1): 15982.

[5] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., & Sun, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. ArXiv, abs/2303.04226.

[6] Caterini, A.L, et al. (2018) Hamiltonian Variational Auto-Encoder. Neural Information Processing Systems

[7] Goodfellow, I., et al. (2020) Generative adversarial networks. Communications of the ACM, 63(11): 139-144.

[8] Yang, L., Zhang, Z., et al. (2023) Diffusion Models: A Comprehensive Survey of Methods and Applications. ACM Computing Surveys, 56(4): 1-39.

[9] Dhariwal, P., et al. (2021) Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems. Curran Associates, Inc. 8780–8794.

[10] Saharia, C., et al. (2022) Palette: Image-to-Image Diffusion Models. ACM SIGGRAPH 2022 Conference Proceedings. New York, NY, USA:Association for Computing Machinery: 1-10.

[11] Anantrasirichai, N. and Bull, D. (2022) Artificial Intelligence in the Creative Industries: A Review. Artificial Intelligence Review, 55(1): 589–656.

[12] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu,Y. and Song, M. (2020) Neural Style Transfer: A Review. IEEE Transactions on Visualization and Computer Graphics, 26(11): 3365–3385.

[13] Wang, W., Li, Y., Zou, T., Wang, X., You, J. and Luo, Y. (2020) A Novel Image Classification Approach via Dense-MobileNet Models[J]. Mobile Information Systems, 2020(1): 7602384.