Dean&Francis

ISSN 2959-6157

Comprehensive Investigation of Algorithmic Models and Prospects in Artificial Intelligence Generated Content

Bofeng Peng

School of Information Technology and Science, Fudan University, Shanghai, China
22307130291@m.fudan.edu.cn

Abstract:

Artificial Intelligence Generated Content (AIGC) has quickly evolved into a critical paradigm for automated content creation, supplementing traditional professionally and user-generated content. Empowered by deep learning, AIGC enables the generation of excellent text, pictures, audio, and video across diverse domains. This paper provides an exhaustive investigation of the core algorithmic models that underpin AIGC technologies, including Generative opposite Networks, Variational Automatic Encoders, Diffusion Models, and Large Language Models. This paper analyzes their principles, technical advancements, and representative applications in visual, textual, and speech generation. Furthermore, the current limitations related to controllability, computational overhead, domain generalization, and ethical considerations were discussed. Looking forward, this paper highlights emerging trends and research directions aimed at improving interpretability, efficiency, and trustworthiness in AIGC systems. By combining technical insight with application-oriented discussion, this paper aims to provide a comprehensive foundation for future research and guide the safe, effective and large-scale deployment of generative artificial intelligence across industries.

Keywords: Artificial intelligence generated content; variational automatic encoders; diffusion model; large language model

1. Introduction

As an emerging content creation paradigm, Artificial Intelligence Generated Content is becoming an important supplement to Professionally Generated Content and User Generated Content. AIGC means the use of artificial intelligence algorithms to automatically generate text, images, audio, video, and other multimodal content based on user demands, thereby meeting diverse task requirements. It is regarded as a key technology driving the transformation of the digital content industry. Therefore, studying the core methods and application potential of AIGC holds significant practical value for advancing the digital economy and improving production efficiency.

In recent years, the development of AIGC has relied heavily on deep learning technologies and has achieved significant breakthroughs in model algorithms, generative tasks, and industry-level products. Deep learning-based generative models can synthesize high-quality content tailored to various task demands [1]. Currently, there are four mainstream algorithmic models: 1) Generative Adversarial Networks (GANs) [2], which generate samples from random noise through opposite training between a generator and a discriminator, are mainly used for image synthesis, style transfer, and image inpainting; 2) Diffusion Models [3], which rely on a two-stage process of forward noise addition and reverse denoising, offer superior image quality and interpretability compared to traditional GANs, and enhance controllability and stability during generation; 3) Variational Autoencoders [4], which introduce randomness model mechanisms, can generate new samples in the meanwhile keeping the potential space of model structure. Recently, the fast rise of massive models has further accelerated AIGC development. Models such as GPT-4, DALL·E 2, and DeepSeek possess vast numbers of parameters [5], strong pretraining capabilities, and impressive generalization performance, achieving milestone results in multimodal generation tasks. The commercialization of these technologies is approaching maturity, with representative products including ChatGPT and DreamStudio, signaling that AIGC is moving from research to large-scale deployment.

However, despite significant improvements in generation quality, discrepancies between generated content and user expectations still exist, posing challenges in real-world deployment and downstream applications [6].

This paper mainly reviews the development of AIGC technologies and core algorithmic methods, analyzes their roles and prospects in various application domains, evaluates the current challenges and limitations, and explores future outlooks.

2. Algorithm Research Based on AIGC Technology

2.1 GAN Models

Goodfellow et al. proposed a frame containing a generator and a discriminator. The generator can learn the diffusion of collective data to generate realistic samples, the discriminator distinguishes if samples are from real data or from generative samples generated by the generator. These two networks are trained adversarially until they reach a Nash equilibrium. This framework is called the GAN because it aims to estimate the real data diffusion through generative modeling. Thanks to its ability to generate excellent samples, GAN has broad application prospects in fields such as image, text, and speech generation [7].

In the visual generation field, an increasing number of researchers use GANs to generate high-quality images or videos. Numerous improved GAN-based models have also emerged. For example, Deng et al. proposed a 3D-aware conditional generation model for controllable image synthesis that renders both images and pixel-aligned label feature maps [8]. Yin et al. introduced a novel 3D GAN inversion method with facial symmetry priors to address the ill-posed problem of reconstructing 3D portraits from monocular images [9]. Xu et al. proposed SH-GAN to address texture blur and structural distortion in image inpainting. SH-GAN incorporates a Spectral Hint Unit and two novel spectral strategies—heterogeneous filtering and Gaussian partitioning—for improved performance [10]. Jain et al. proposed the Fourier Coarse-to-Fine photos repairing framework, combining fast Fourier convolution with coarse-to-fine generator modulation to capture global texture repetition and generate realistic image structures [11].

In text generation, many scholars have also used GANs. For instance, Kanagawa et al. proposed a style predictor with GAN to reduce mismatches between global style tokens and style embeddings from reference speech [12]. Qu et al. proposed MOSTEL, an throughout trainable framework for scene article editing that produces readable edited text images [13].

In the speech generation domain, Yoneyama et al. developed a rapid, pitch-controllable, high-accuracy neural vocoder. By hierarchically conditioning a resonator filter network on noise excitation information and incorporating the source filter into HiFi-GAN, their method improves both speech quality and synthesis speed [14].

2.2 VAE Models

Researchers have proposed many improvements to VAE models. In speech synthesis, Melechovsky et al. introduced a TTS model using multi-stage VAEs and opposite learning to handle accented speech topic and conversion. Chen et al. addressed expressiveness limitations in audiobook TTS using a self-supervised VQ-VAE style augmentation method. Li et al. proposed a cross-utterance conditional VAE speech synthesis framework that extracts acoustic, speaker, and textual features from nearby sentences to generate context-sensitive prosody for more natural speech. Xiang et al. combined deep complex convolutional recurrent networks with VAE for speech enhancement, modeling latent variables as complex Gaussians to better capture signal behavior[15] .

In image tasks, VAEs are widely used for generation and avatar creation. They are critical in spatiotemporal compression of video in models like OpenAI's SORA. Models based on discrete tokens from 3D VAEs or continuous latent variables from 2D VAEs underpin various diffusion video models. Zhao et al. proposed a training method for video VAEs compatible with pre-trained image VAE latent spaces.

In unsupervised learning, learning a useful representation space remains a challenge. VAE (with continuous latent variables) and VQ-VAE (with discrete latent variables) are important representation learning methods. VQ-VAE replaces continuous latent spaces with discrete token embeddings but introduces non-differentiability issues. To address this, Pucci et al. proposed the Capsule Vector-VAE, which replaces discrete bottlenecks with differentiable capsule layers [16].

2.3 Diffusion Models

Ho et al. introduced diffusion models whose inspiration is non-equilibrium thermodynamics. Diffusion models include a forward process (progressively adding Gaussian noise) and an opposite process (progressively denoising back to the original data) . Compared to GANs, diffusion models exhibit greater robustness and flexibility, and are now widely adopted in image, text, and speech generation.

In image/video generation, diffusion models handle complex deformations and occlusions well. They break down complex transformations into multiple denoising steps, effectively learning mappings from source to target images while preserving textures and details. In video generation, temporal alignment mechanisms extend static image diffusion models into temporally consistent video generators [17].

In text generation, diffusion models serve as alternatives to traditional autoregressive models. Discrete diffusion-based masked language modeling allows non-autoregressive generation of diverse, semantically complete text. Sequence-to-sequence structures with bidirectional encoders and denoising decoders refine noisy initial text. A "continuous paragraph denoising" strategy further improves context comprehension and richness.

In speech synthesis, diffusion models enable multi-level neural vocoders to produce high-quality, natural speech. These models support multiple speakers and singing voices, showing superiority in audio clarity and frequency modeling.

2.4 Algorithm Research Based on Large Language Models

Since ChatGPT's release in December 2022,large language models have received widespread attention for their practical value across industries. Their rise stems from key algorithmic breakthroughs. This section reviews the algorithmic development of generative foundation models, summarizing their strengths and limitations.

The precursor of foundation models was the Transformer-based BERT model, which applied self-supervised learning on unlabeled datasets to produce high-quality representations. However, BERT and its variants require fine-tuning for downstream tasks. In contrast, GPT models perform well in small-sample and zero-sample article generation scenarios.

Radford et al. introduced GPT-1, which consists of 12 layers of Transformers and 117M parameters. Despite its size, GPT-1 excelled in various NLP tasks. GPT-2 followed with improvements in word recognition, long-time dependency, common-sense inference, reading understanding, summarization, and translation [18].

GPT-3's debut marked an explosive growth in foundation models. GPT-3 is a model which has a batch size of 3.2 million, 96 attention layers, 175 billion parameters, trained on diverse online content, achieving remarkable performance in text generation and translation. GPT-3.5—best known through ChatGPT—added fine-tuning for better instruction-following and commercial usability.

On March 14, 2023, OpenAI released GPT-4, a multimodal model that supports both words and pictures input. Compared to GPT-3.5's text-only input/output, GPT-4 is significantly more capable.

In January 2025, DeepSeek-V3 was released, featuring 671 billion parameters with a 37B active parameter MoE architecture, incorporating multi-head latent attention and novel load-balancing strategies without auxiliary losses. It was pretrained on 148T high-quality tokens and fine-tuned using supervised and reinforcement learning in a stable training process requiring only 2.79M H800 GPU hours [19]. DeepSeek-R1, designed for complex reasoning, excels at logic and math problems.

In image generation, models like OpenAI's DALL·E 2 (an extension of GPT) can generate unique, high-quality images from text in minutes. Stable Diffusion offers fast text-to-image generation and AI art tools. Bai et al. proposed a method for training large vision models without using any language data, solving multiple visual generation tasks .

In speech generation, GPT-based methods are also gaining traction. The VALLE model can generate personalized high-quality speech from text. AudioPaLM combines PaLM-2 and AudioLM into a unified multimodal model,

ISSN 2959-6157

capable of tasks such as speech recognition and speech-to-speech translation.

In multimodal generation, Google released Gemini, a model designed for seamless reasoning across text, image, video, audio, and code. Benchmarks show Gemini outcompetes the cutting-edged technology in various tasks.

3. Discussion

Despite the impressive progress in AIGC algorithms, several limitations and unresolved challenges remain that restrict their full-scale deployment and industrial applicability. This section discusses these core issues and outlines prospective directions for future research and development.

3.1 Challenges and Limitations

3.1.1 Misalignment between generated content and user intent

While AIGC models such as GPT-4 and Stable Diffusion demonstrate remarkable generative capabilities, a significant gap still exists between user expectations and model outputs. Generated text may contain hallucinated facts, and synthesized images may lack precise semantic alignment with prompts. This misalignment is particularly problematic in applications requiring high accuracy, such as legal or medical content generation. The current prompt-based control is limited, and even with prompt tuning or classifier guidance, achieving fine-grained, consistent control remains a bottleneck.

3.1.2 High computational costs and accessibility barriers

Training and deploying large-scale generative models involve extensive computational resources. Models like DeepSeek-V3 require hundreds of billions of parameters, vast token datasets, and millions of GPU hours for training. This leads to high development and maintenance costs, creating an entry barrier for smaller research labs and startups. Moreover, the inference speed of diffusion-based models remains a constraint, particularly in real-time applications. While model distillation and optimization techniques are being explored, efficient AIGC systems for edge devices are still under development.

3.1.3 Generalization and domain adaptation challenges

Although LLMs show strong performance in general tasks, their adaptability to specific domains with limited data is often inadequate. Fine-tuning with small domain-specific datasets can result in overfitting or loss of general knowledge. Moreover, models struggle with incorporating domain-specific logic, temporal consistency in video generation, and structured information, limiting their utility in enterprise and scientific scenarios. Lack of grounding in external knowledge bases or physical principles further restricts reasoning and interpretability.

3.1.4 Ethical, privacy, and security concerns

AIGC raises substantial ethical concerns. Models trained on public data risk unintentional reproduction of copyrighted or sensitive material. They may propagate societal biases, stereotypes, or misinformation embedded in training corpora. Furthermore, malicious actors can exploit generative tools for deepfakes, disinformation, or automated spam generation. Privacy breaches, unauthorized identity replication, and the lack of transparent auditability have emerged as serious issues. While current watermarking and moderation mechanisms are helpful, they remain insufficiently robust against advanced misuse tactics.

3.2 Future Prospects

3.2.1 Infusion of domain knowledge and physical logic

One key direction for overcoming current limitations is integrating symbolic reasoning, external knowledge bases, and physics-informed learning into AIGC systems. By embedding structured knowledge or scientific constraints, models can improve their interpretability, consistency, and logical coherence. Hybrid approaches combining neural and rule-based modules may yield better results in critical applications such as education, healthcare, and scientific discovery.

3.2.2 Efficient and scalable generative architectures

Reducing the computational and environmental cost of AIGC is critical for widespread adoption. While GANs are capable of producing high-quality samples with relatively few iterations, they often suffer from instability or collapse in later training stages. Moreover, GAN-generated data may not always be practical for downstream tasks, increasing training costs. Diffusion models provide greater robustness and output fidelity, yet require lengthy Markov chain sampling, resulting in high computational costs and slower generation. Recent research focuses on improving efficiency via faster sampling techniques, model compression, and hybrid architectures. In parallel, the development of lightweight and real-time models—through methods like quantization and sparsity—will be essential for edge deployment.

3.2.3 Trustworthy, transparent, and regulated aigc systems

Building trust in generative models requires progress on explainability, transparency, and verifiability. Mechanisms such as self-reflection modules, confidence calibration, and human-in-the-loop feedback can enhance reliability. LLMs, while powerful, still face limitations such as data security and privacy concerns, particularly in specialized or sensitive applications. Additionally, in tasks requiring logical reasoning or structured output, their performance remains limited. Therefore, regulatory frameworks, open benchmarking, and responsible AI practices must be established to promise safe, aligned, and ethical use of AIGC technologies.

4. Conclusion

AIGC has been seen as a revolutionary force across content creation, scientific modeling, and human-computer interaction. Driven by progress in generative models such as GANs, VAEs, Diffusion models, and LLMs, AIGC systems have achieved impressive performance in multimodal generation tasks. This paper reviewed core algorithmic developments, analyzed representative applications, and identified key limitations including controllability, computational efficiency, domain adaptation, and ethical risks. As research continues to evolve, the future of AIGC will depend on breakthroughs that enable a more interpretable, efficient, and trustworthy generation. Integrating domain knowledge, reducing resource demands, and enhancing safety mechanisms will be essential for transitioning AIGC from experimental models to dependable tools across industries. Ultimately, realizing the full potential of AIGC requires a balanced focus on algorithmic innovation, responsible governance, and interdisciplinary collaboration.

References

[1] Hao Y, Liu Y, Mou L. Teacher forcing recovers reward functions for text generation. Advances in Neural Information Processing Systems (NeurIPS), 2022, 35: 12594–12607.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144.

[3] Croitoru FA, Hondru V, Ionescu RT, et al. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850–10869.

[4] Doersch C. Tutorial on variational autoencoders. arXiv Preprint, 2016, arXiv:1606.05908.

[5] Wan T, Wang A, Ai B, et al. WAN: Open and advanced large-scale video generative models. arXiv Preprint, 2025, arXiv:2503.20314.

[6] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by backpropagating errors. Nature, 1986, 323(6088): 533–536.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 2014, 27.

[8] Deng K, Yang G, Ramanan D, et al. 3D-aware conditional image synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 4434–4445.

[9] Yin F, Zhang Y, Wang X, et al. 3D GAN inversion with facial symmetry prior. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 342–351.

[10] Xu X, Navasardyan S, Tadevosyan V, et al. Image completion with heterogeneously filtered spectral hints. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023: 4591–4601.

[11] Jain J, Zhou Y, Yu N, et al. Keys to better image inpainting: Structure and texture go hand in hand. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023: 208–217.

[12] Kanagawa H, Ijima Y. Enhancement of text-predicting style token with Generative Adversarial Network for expressive speech synthesis. ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023: 1–5.

[13] Melechovsky J, Mehrish A, Sisman B, et al. Accent conversion in text-to-speech using multi-level VAE and adversarial training. arXiv Preprint, 2024, arXiv:2406.01018.

[14] Qu Y, Tan Q, Xie H, et al. Exploring stroke-level modifications for scene text editing. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 2119–2127.

[15] Yoneyama R, Wu YC, Toda T. Source-filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder. ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023: 1–5.

[16] Bai Y, Geng X, Mangalam K, et al. Sequential modeling enables scalable learning for large vision models. arXiv Preprint, 2023, arXiv:2312.00785.

[17] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. OpenAI Blog, 2018.

[18] Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. arXiv Preprint, 2024, arXiv:2412.19437.

[19] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents. arXiv Preprint, 2022, 1(2): 3.