Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Advancements and Challenges in Text-to-Image Synthesis: Exploring Deep Learning Techniques

Ge Chen *

Department of mathematics and physics, Beijing University of Chemical Technology, Beijing, China

* Corresponding Author Email: 2020080232@buct.edu.cn

Abstract. This paper presents an in-depth exploration of text-to-image synthesis, categorizing the field into three main areas based on deep learning methodologies. Each category undergoes meticulous analysis to chart its development and dissect its fundamental mechanisms. The emphasis is on the significant advancements in the field, with a special focus on major breakthroughs and the continual evolution of these technologies. Concurrently, the paper critically evaluates the challenges and limitations present in the current state of text-to-image synthesis, providing valuable insights into the obstacles hindering progress. The study also delves into potential applications, pondering the future impact of this technology across diverse industries. The implications of these advancements are examined, considering not only technological capabilities but also their wider societal and ethical consequences. This comprehensive review not only sheds light on the current landscape of text-to-image synthesis but also looks forward to future innovations, highlighting the dynamic and continuously evolving nature of this area. The fusion of deep learning with creative image generation is poised for groundbreaking applications, setting the stage for transformative shifts in the interaction and interpretation of visual and textual content.

Keywords: Image generation; deep learning; GAN; diffusion; autoregressive.

1. Introduction

The field of computer vision, propelled by the swift advancements in artificial intelligence, stands at the forefront of today's research endeavors. Notably, products developed by AI research entities, including OpenAI, have revolutionized image and video generation technologies, garnering global attention. Among these innovations, text-to-image synthesis has found practical applications, reflecting the dynamic nature of this field. Text-to-image generation involves creating images that align with a given text description via a computer network. This technology, evolving with advancements in AI, now leverages deep learning for image generation, marking a departure from traditional methods. Current deep learning-based image generation encompasses various models such as generative adversarial networks (GANs), diffusion models, and autoregressive models. This article focuses on text-to-image generation methods derived from these three model types. Firstly, the article explores text-to-image methods based on the GAN model. GAN, distinctly different from traditional text-to-image approaches, holds a dominant position in data generation. The discussion includes an overview of GAN's basic model structure, a brief review of enhanced models like CGAN and StyleGAN, and the latest developments, such as TediGAN [1]. Secondly, it examines text-to-image methods based on the diffusion model [2], a pivotal learning model since OpenAI's popularization and application in 2021. The article elucidates the fundamental structure and principles of the diffusion model, presents the innovative Diffstyler dual diffusion processing framework, and covers popular conditional guidance image generation methodologies based on this model [3]. Lastly, it delves into text-to-image methods based on the autoregressive model. This approach, distinct from the aforementioned models, has its unique strengths. The article details the autoregressive model's basic principles and structure, introduces classic implementations like Parti, and discusses the advantages and prospective developments of this model type [4].

2. GAN Models

2.1. GAN

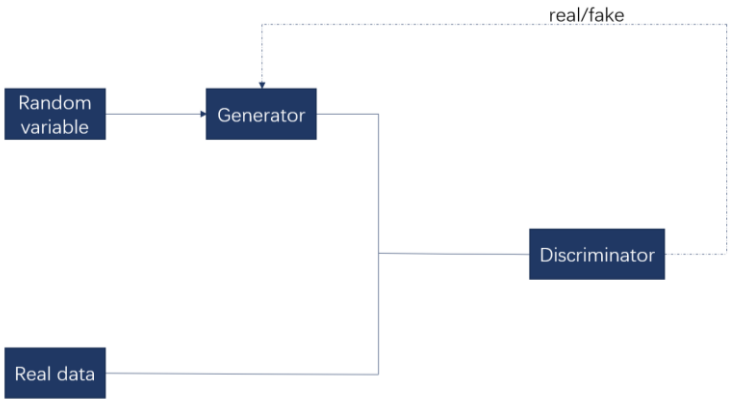

Fig.1 The structure of GAN (Photo/Picture credit: Original).

Among the learning models of text-to-image generation, GAN model is a classic model with absolute milestone significance. The traditional GAN model consists of two main parts: G(a generator) and D(a discriminator). When a z(random variable) is input to the G, the G outputs the $G(z)$(generated data); then the G(z) and y(the real data) are input to the D together, and the D will output $P(y)$(the probability of the authenticity of the real data) and $P(G(z))$(the probability of the authenticity of the G(z)). The closer the probability is to 1, the higher the authenticity.

This forms the phenomenon of the generator and the discriminator competing with each other, so the generation based on the GAN model is also called generative adversarial network. In the training process, the generator and the discriminator complete iterative optimization by constantly competing with each other. When the probability of authenticity of the generated data and the real data are both 0.5, this is the ideal situation. Therefore, the design of the loss function of the adversarial generative network is also very important. The classic GAN achieves the idealized goal by introducing adversarial loss.

$$L_G = -E_{x' \sim P_G}[log_2(D(x'))] \tag{1}$$

$$L_D = -0.5E_{x \sim P_{data}}\left[log_2\left(D(x)\right)\right] - 0.5E_{x' \sim P_G}[log_2(1 - D(x'))] \tag{2}$$

2.2. CGAN

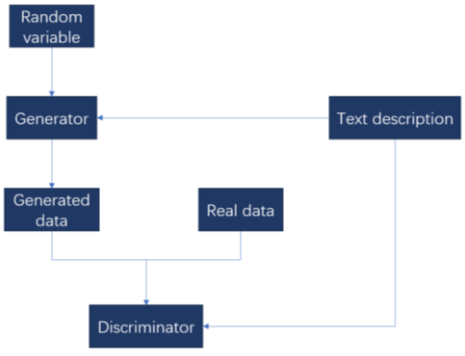

Fig.2 The structure of CGAN [5].

However, with the joint development of technology and demand, the original single GAN model cannot meet the requirements of generating images that match the text description. In order to control

the generation of images that match the text description instead of simply randomly generating images, the CGAN model was born. As shown in the figure, the structure introduces the text description in the generator and the discriminator, which can guide the image generation through the text description. This is also a typical conditional generative adversarial network. Compared with the GAN model, the CGAN model introduces the text description condition, which makes the generated images more accurate and meets the requirements [6].

The loss function of the CGAN structure also made corresponding changes, and the text condition was written into it as a new variable, such as φ.

$$L_G = -E_{x' \sim P_G, \varphi \sim P_{data}}[log_2(D(x', \varphi))] \tag{3}$$

$$L_D = -0.5E_{x \sim P_{data}}\left[log_2\left(D(x)\right)\right] - 0.5E_{x' \sim P_G}[log_2(1 - D(x'))] \tag{4}$$

2.3. StyleGAN

In the development process of GAN model, styleGAN has a milestone significance. Different from the traditional GAN, the biggest innovation of styleGAN is the introduction of "style vector", which maps the latent vector to different style vectors to control the high-level features of the image, such as facial features, eye color, etc., to generate highly personalized images. And the latent space W of styleGAN can be used to develop inversion technology, which can invert the real image back to the latent space to perform many meaningful operations [7].

The main body of styleGAN has two parts, namely Mapping network and Synthesis network. In the working process: first input the image, then convert it to a hidden code, then input the hidden code into the Mapping network to decouple, and get an intermediate vector w, these intermediate vectors will be passed to the generation network to get 18 control vectors, these 18 control vectors are grouped in pairs, and input into the AdaIN module of the 9 convolutional layers of the Synthesis network, before each AdaIN module, a scaled noise is added to each channel, and finally output an image of different styles.

2.4. TediGAN

After styleGAN attracted many people to study its inversion technology with its unique latent space, the GAN series of models have been greatly developed, and more and more models based on styleGAN have been developed. TediGAN is a recent model with advanced performance in virtual face generation, which is different from other GAN models' inversion methods. TediGAN focuses on text-based inversion [8]. The module is divided into StyleGAN model inversion module and visual language similarity module, which learn the language representation consistent with visual language by jointly inverting text and image into the latent space, and instance-level optimization module, which accurately operates the attributes consistent with text from the latent space for optimization. Its superiority is reflected in its ability to generate diverse and high-quality results with a resolution of up to 10242, and support image synthesis with multimodal input.

The main structure of TediGAN is divided into StyleGAN Inversion Framwork and visual language similarity module. StyleGAN Inversion Framwork is different from the general GAN model's inversion. TediGAN model's inversion is to map the real image to the latent space, thus obtaining the encoder code that is semantically consistent with the generation. The visual language similarity module projects the image and text description into a common embedding space W, and the text compiler trained by this module can achieve the learning of the association and alignment between text and image.

3. Diffusion Models

In the technological evolution of text-to-image synthesis, diffusion model-based techniques have demonstrated significant potential for development, characterized by high-quality generation and diverse image styles. Diffusion models generate data by simulating the diffusion process observed in the natural world. In the context of image generation, these models start with a distribution of random noise and gradually introduce structure to produce clear images. The core concept involves the forward process of a Markov chain and the reverse diffusion process. During the forward process, the model incrementally adds noise to the data until it becomes entirely noisy; in the reverse process, the model progressively recovers data from the noise.

Within the framework of text-to-image synthesis, image generation techniques based on diffusion models are generally divided into two stages: text encoding and conditional generation. The text encoding stage converts textual descriptions into a continuous vector representation; the conditional generation stage then employs the diffusion model to create images based on the text vectors obtained from the previous stage.

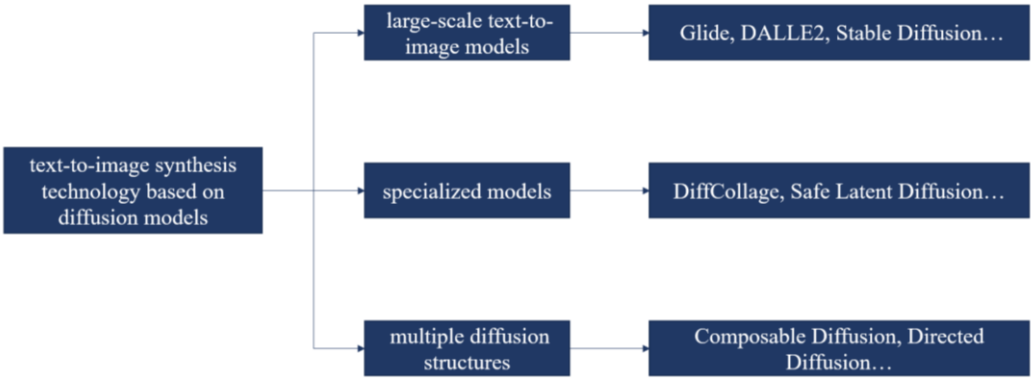

Fig.3 Condition-based classification of conditional image generation model (Photo/Picture credit: Original).

Presently, text-to-image synthesis technology based on diffusion models has reached a considerable level of maturity, as illustrated in the Fig.3. Various types of techniques have been developed for different application scenarios. Among large-scale text-to-image models, Glide was the first of its kind to be introduced. Glide utilizes a cascading architecture and an implicit classifier to initially generate a low-resolution, coarse result based on the text vector, followed by a text-conditioned super-resolution model to produce a high-resolution image. For scenarios requiring the generation of images with specific elements such as colors and quantities, specialized models are employed. For instance, Diff Collage represents the relationships of various parts within an image using nodes, with each node denoting a segment of the generated image. When the complexity of the desired image is substantial, models like Composable Diffusion, which combine multiple diffusion structures, emerge. In the generation process, different diffusion models are responsible for creating distinct content, which is then amalgamated to achieve the requisite complexity. This demonstrates the current diffusion models' sufficient advancement.

3.1. Diffstyler Framework

As image generation technology matures, the demand for its practical applications has significantly increased, surpassing the capabilities of basic image generation techniques. Text-driven style transfer in current images represents a novel research direction. Presently, tasks involving text-driven stylization are primarily based on Generative Adversarial Network (GAN) models. However, the adversarial structure of GANs, along with the patch style discriminator, limits their stylization capabilities, often failing to produce satisfactory results. Consequently, diffusion models, which hold

an advantage in generating artistic images, have been introduced for this task. Among these, the DiffStyler model has achieved a high degree of completion in terms of content preservation and style transfer in images.

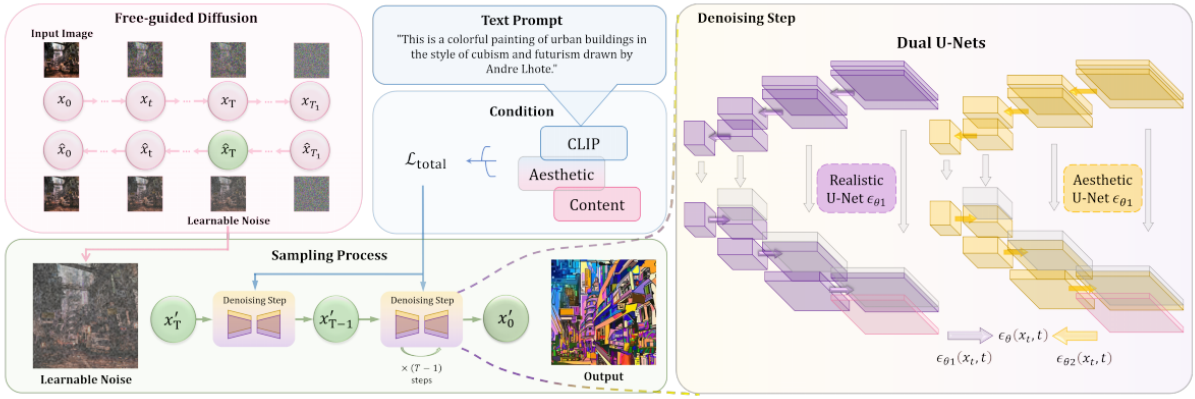

Fig.4 The three sections of Diffstyler[4] [9].

As depicted in the Fig.4, the workflow of DiffStyler[4] primarily consists of two steps. The first step is the free-guided diffusion process, where the input image x_0 is fed into the diffusion model to obtain learnable noise x_T. The second step involves taking the learnable noise x_T obtained from the first step as the input x_T' for the dual diffusion model, which then undergoes reverse sampling to yield the output result x_0'. Notably, during the reverse sampling process, guiding conditions are incorporated into the denoising steps, and each step employs a dual denoising U-Net architecture, which encompasses both realistic and aesthetic components.

4. Autoregressive Model

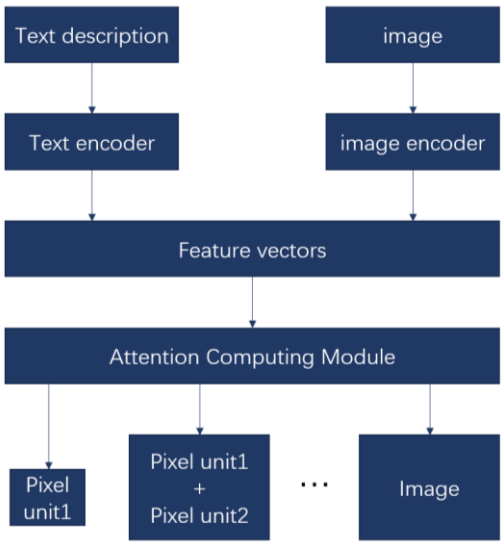

Fig.5 The process of Autoregressive model (Photo/Picture credit: Original).

The emergence of autoregressive models constitutes an indispensable part of the text-to-image synthesis technology domain. Google's recently proposed Parti model has elicited a significant response and possesses the remarkable capability to generate images that correspond with textual descriptions. Specifically, autoregressive models employ a step-by-step approach to generate each pixel unit of the image during the generation process. As illustrated in Fig.5, the process begins with a text encoder that produces feature vectors for the current pixel unit. These vectors are then used by a neural network, in conjunction with the textual description, to predict the conditional probability

distribution of the next pixel unit. This prediction yields the pixel, which is then used as the input for the subsequent pixel prediction and generation, and this cycle continues until the image is fully generated.

4.1. Parti

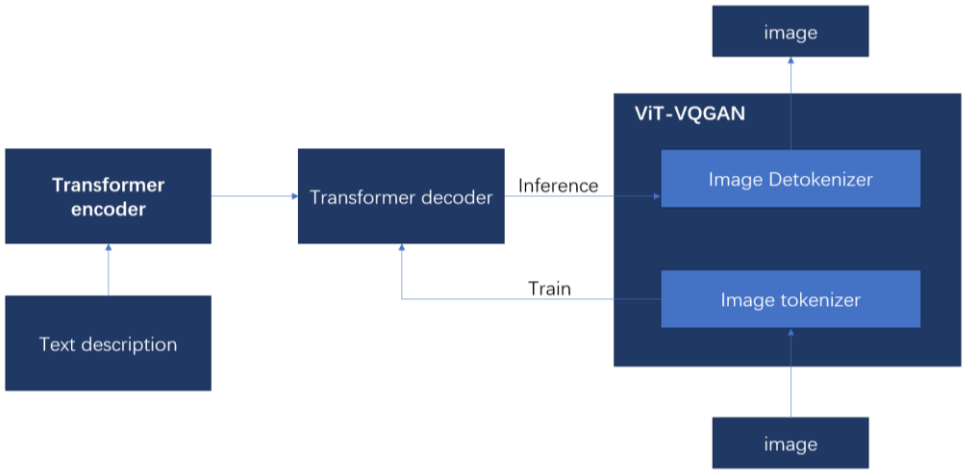

Fig.6 The main structure of Parti [10].

The core structure of the Parti model consists of an image tokenizer and an autoregressive model. As Fig.6 illustrated, the commonly used image tokenizer is the ViT-VQGAN architecture, which converts images into a sequence of visual tokens for training and concurrent image reconstruction. The other component is the training of a sequence-to-sequence autoregressive model.

Researchers are also exploring how to optimize this model, and discussions regarding the scale of Partihave become a research direction. Four different scales of Parti have been obtained through scaling operations, namely Parti-350M, Parti-750M, Parti-3B, and Parti-20B. Utilizing these four scales of Parti in various scenarios has proven that as the number of parameters increases, the images generated by the model become more realistic.

5. Challenges and Prospects

In considering the trajectory and future prospects of technology, two critical perspectives emerge for discussion: the absence of a universal standard evaluation system and the revolutionary goal of technology, such as the widespread application of text-to-video synthesis technology akin to sora.

A universal evaluation system is crucial for advancing text-to-image synthesis technology. Currently, the lack of standardized criteria for evaluating such technology complicates the comparison and validation of research outcomes. Commonly used evaluation metrics include the Inception Score (IS) and the Fréchet Inception Distance (FID), which, despite their respective assessment capabilities, suffer from severe limitations in generalizability and lack assessments for semantic consistency, thus focusing solely on visual quality while neglecting semantic alignment. A comprehensive evaluation system should encompass multiple dimensions, including image quality, generation speed, diversity, creativity, and relevance to textual descriptions. Moreover, the system should account for the model's generalizability and applicability across various domains. Establishing such a system would enable researchers and developers to objectively measure and compare the performance of different models, thereby fostering technological advancement.

On the horizon of application prospects is the goal of revolutionizing technology, such as extending text-to-image synthesis to the video domain, which represents an important trend for future development. With the continuous advancement of deep learning technology, it is foreseeable that, in the near future, technology capable of generating high-quality video content based on textual descriptions will be realized. This technology has the potential to significantly impact multiple fields,

including film production, game design, and virtual reality, and could even catalyze a market revolution in the entertainment industry. For instance, text-to-video synthesis technology like 'sora' not only generates static images based on textual descriptions but also creates dynamic video sequences, greatly enhancing the efficiency and diversity of content creation.

In summary, to propel the advancement of text-to-image synthesis technology based on deep learning, there is an urgent need for a universal evaluation system to measure and compare the performance of various technologies. Concurrently, the revolutionization of technology is burgeoning, with its application in video generation poised to meet the escalating demand for multimedia content. These directions of development and application prospects will lay a solid foundation for future research and implementation.

6. Conclusion

This article offers an in-depth analysis of text-to-image synthesis, a rapidly evolving field within deep learning. It systematically classifies the domain into three core technological approaches: Generative Adversarial Networks (GANs), diffusion models, and autoregressive models. The paper meticulously charts the evolution of each approach, with a focus on the advent of novel methodologies in recent times. Furthermore, it underscores a critical gap in the field: the absence of a standardized, objective framework for evaluating these technologies. This lacuna highlights the need for further development in this area. Additionally, the paper emphasizes the future impact of text-to-image synthesis, particularly its significant potential to revolutionize text-to-video synthesis. This advancement is poised to dramatically influence the next generation of the film and entertainment industry, suggesting an impending technological transformation that is both urgent and essential.

Reference

[1] Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

[2] Xia, W., Yang, Y., Xue, J. H., & Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2256-2265).

[3] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[4] Huang, N., Zhang, Y., Tang, F., Ma, C., Huang, H., Dong, W., & Xu, C. (2024). Diffstyler: Controllable dual diffusion for text-driven image stylization. IEEE Transactions on Neural Networks and Learning Systems.

[5] Zhu, X., Zhao, Z., Wei, X., & others. (2021). Action recognition method based on wavelet transform and neural network in wireless network. In 2021 5th International Conference on Digital Signal Processing (pp. 60-65).

[6] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

[7] Zhang, Q., Song, J., Huang, X., Chen, Y., & Liu, M. Y. (2023, June). Diffcollage: Parallel generation of large content with diffusion models. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10188-10198). IEEE.

[8] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z.... & Wu, Y. (2022). Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2(3), 5.

[9] Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J.... & Wu, Y. (2021). Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627.

[10] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.