Pixels Aligned with Words: Technical Route and Horizons of Text-to-Image Generation

Xiaotian Hu*

School of Artificial Intelligence, Southeast University, Nanjing, China

*Corresponding author: 213232245@seu.edu.cn

Abstract. As a hot field in current society, artificial intelligence text-to-image gen-eration has received extensive attention in recent years. Text-to-Image generation task refers to the process of converting natural language descriptions to corresponding visual content such as pictures and illus-trations, and has demonstrated a powerful influence in fields such as education, economic models, and artistic creation. Based on different technical frameworks, the current mainstream Text-to-Image large models can be divided into diffusion-based models, generative adver-sarial networks, variational autoencoder-based models and other methods. Different technical architectures have their own advantages and characteristics. Based on the above representative frameworks, this paper introduces some of the latest technological developments, ex-pounds its innovation direction and operation process, and analyzes its shortcomings. This paper introduces a few classic datasets such as LAION-5B and COCO, and analysis the performance of representative methods on these datasets. This paper summarizes the current prob-lems in the field of Text-to-Image, looks forward to the future develop-ment direction, and hopes to bring some inspiration to future researchers.

Keywords: Pixels Aligned with Words; Text-to-Image; Generation.

1. Introduction

In 2016, StackGAN embedded a text encoder into the GAN for the first time, enabling end-to-end training from a sentence to an image, marking the entry of text-to-image synthesis tasks into the field of deep learning [1]. In 2019, DALL-E [2] and CogView [3] used 1-4 B-level transformers for the first time to run from text to discrete image tokens to high-fidelity images, pushing text-to-image generation tasks into the era of large models and big data. At present, text-to-image generation tasks have shown great value and potential in advertising and marketing, e-commerce, games, film and television and other fields, which leverage the cost structure, business model and social division of labor of the entire creative industry chain, creating a lot of value and social benefits.

Due to the high-dimensional, multi-modal, and strongly constrained characteristics of image generation, its technological development can be divided into different levels of evolution. In terms of data, the training data of image generation models has experienced an increase in data volume. As hardware performance grows, the development of models faces a game between computing resources and sampling efficiency. And the needs of different application scenarios also put forward different requirements. Constrained by the above development dimensions, text-to-image generation models form several main technical frameworks.

The diffusion model is a type of deep probabilistic model that realizes data generation through a two-step process of noising and denoising. The workflow is to gradually add Gaussian noise to the real image in the training stage until it becomes approximately pure noise, and then use a learnable neural network to reverse denoising in the inference stage to restore the random noise to a high-fidelity image. Representative work includes DDPM [4], Stable Diffusion [5], DALL-E2 [6], etc.

The generative adversarial network (GAN) includes at least two modules, as well as a generative model and a discriminant model. The generative model receives the noise and generates a picture, and the judgment model judges the possibility of the real image based on the image. Through the game mechanism of dual networks, generatorlearns to convert random noise into images, audio or text that are almost indistinguishable from the real data distribution in adversarial training, and finally

realizes the output of high-fidelity samples with a single forward inference. Its representative work includes DCGAN [7], StyleGAN [8], etc.

For Variational Autoencoder (VAE), as a form of deep generative model, the variational autoencoder is a generative network structure based on variational Bayesian inference proposed by Kingma et al. in 2014 [9]. The input data is encoded into a learnable distribution of probability latent variables, and then sampled and decoded from it, and trained jointly with the "reconstruction error KL regular", which not only retains the ability to generate data, but also provides explainable uncertainty estimation through explicit likelihood modeling.

2. Method introduction

2.1. Diffusion Model

In 2015, the Stanford team first proposed a diffusion probabilistic model, which turned "adding noise and removing noise" into a learnable generation process using the idea of non-equilibrium thermodynamics [10]. In the following three years, DDPM used variational inference to write ELBO as an optimizeable objective function [4], SGM used stochastic differential equations to give a continuous time perspective [11], and DDIM cut Markov chains into deterministic mapping [12], compressing 1,000-step sampling to less than 100 steps in one fell swoop, laying the "three cornerstones" of modern diffusion models.

In 2022, DALL-E 2 [6] uses CLIP to reverse text-image embedding, Stable Diffusion [5] moves computing into the latent space so that consumer graphics cards can also run, the open source community explodes overnight In 2023-2024, ControlNet [13] and T2I-Adapter [14] turned edge maps, depth maps, postures and even QR codes into "hard constraints." LoRA [15] and DreamBooth [16] allow ordinary users to fine-tune their exclusive models with dozens of images. DiT moves Transformer into the diffusion framework to open the era of "large models and large resolution" [17].

In terms of computing efficiency, Huawei's Noah's Ark Lab and several universities have released the PixArt-Σ model, which can natively output high-quality images with 4K resolution with only 0.6 B parameters [18]. In order to achieve the lightweight of parameters and improve the training efficiency, the researchers have made two main improvements. In the DiT architecture, the researchers designed a local aggregation compression mechanism to reduce the amount of tokens in 2×2 steps of convolutional compression of keys and values in the self-attention layer to reduce the amount of tokens, In terms of training strategy, the researchers adopted a training strategy from weak to strong, gradually realized the resolution upgrade.

In terms of subject consistency, InstantID, developed by the Chinese team instantX, focuses on zero-shot identity retention generation [19]. With just a single reference face image, you can quickly generate high-fidelity, multi-style personalized images while maintaining a high degree of consistency in character identity characteristics. In order to achieve this function, the research group introduced three core components. ID Embedding extracts face semantic features through InsightFace, and projects them into the text embedding space. Image Adapter lightweight decoupled cross-attention module, independently integrating text prompts and face features to avoid style and identity conflicts. IdentityNet combines strong semantic face embedding and weak spatial key points . It guides generation under the premise of freezing the pre-trained model

2.2. Generate Adversarial Models

The exploration of text-to-image generation with GANs began in 2016, Reed et al. first concatenated a GAN with an RNN text encoder, proving the feasibility of "text → 64^2 image" [20]. Subsequently, StackGAN proposed a two-stage "sketch-refinement" approach for gradually upsampling from low to high resolution [1], StackGAN++ introduced tree-based generation and multi-scale discriminators [21], and AttnGAN further added cross-modal attention [22], with significant improvements in detail and semantic consistency.

From 2019 to 2021, the field entered the "deep semantic alignment" stage: DM-GAN used a memory network to iteratively complete missing details [23], DF-GAN simplified training with a deep and narrow fusion module [24], and XMC-GAN pushed the FID below 10 using cross-modal contrastive loss [25]. StyleGAN-Tinjected text conditions into the latent space of StyleGAN to achieve 1024^2 high-resolution images [26].

After 2022, diffusion models rose to prominence, and GANs were in a bottleneck for a long time. Researchers repositioned GANs as "diffusion accelerators", GigaGAN used GAN distillation of Stable Diffusion to achieve ultra-fast sampling in 1-4 steps [27]. A representative example of this idea is SDXL-Lightning [28]. The core technology of SDXL-Lightning lies in "progressive adversarial distillation": first, a "teacher distribution" is generated through multi-step diffusion, and then a discriminator with the same UNet structure and 1-8 step "student generators" are adversarially trained in the latent space. Without the need for additional networks, 50-step diffusion can be compressed to 1-8 steps. Its value is reflected in "second-level 1024^2 high-definition image generation"

In this context, in 2025, Brown University and Cornell University jointly released R3GAN [29], a landmark for the field of GAN. Its core innovation is the introduction of the regularized relativistic loss function (RpGAN + R1/R2 gradient penalty), which solves the problem of model collapse and training instability of traditional GAN through mathematical proof, and abandons complex empirical techniques. The model adopts a minimalist modern architecture, which generates high-quality images comparable to diffusion models with only half the number of parameters, surpasses mainstream GANon FFHQ, ImageNet and other datasets providing efficient solutions for mobile and edge computing scenarios

2.3. Variational Auto-Encoders:

In 2013, VAE proposed to use probabilistic potential space to improve diversity for the first time. In 2015, CVAE introduced text as a condition to achieve controllable synthesis. Subsequently, VQ-VAE discretized the continuous latent space from 2017 to 2019, significantly improving clarity and controllability, and paving the way for subsequent large models.

The real breakthrough was OpenAI's DALL· E [2], which uses VQ-VAE to turn images into discrete tokens, and then uses a Transformer to learn text-image joint distribution on 250 million image-text pairs, achieving high-quality and large-scale text generation for the first time. Since then, DALL· E 2 [6] and Stable Diffusion [5] continue to retain the VAE encoder but introduce diffusion models and CLIPs to efficiently generate within the latent space, further pushing the level of detail and realism to new heights.

By the end of 2023, Peking University and ByteDance proposed VAR [30], which simplifies the traditional two-stage process of "first training VQ-VAE and then training diffusion" into a one-time end-to-end autoregressive generation: first, VQ-VAE compresses 256^2 images into 32×32 discrete tokens, and then all tokens are autoregressively predicted in a chain sequence of resolutions $1\rightarrow2\rightarrow4\rightarrow\ldots\rightarrow32$ within 20 steps. This framework provides a new high-speed and high-fidelity paradigm for "direct image generation by large language models".

At the same time, researchers began experimenting with VAE as an encoder. Qwen-Image uses a diffusion-transformer architecture [31], but deploys a 3D-VAE encoder at the forefront, compressing 1024×1024 RGB images into a latent tensor of only 16×128×128, which not only preserves texture details but also significantly reduces the memory footprint. The discrete codebook of the 3D-VAE also provides regular, continuous latent variable inputs for the diffusion transformer, making text rendering and image generation more accurately aligned in the latent space.

3. Analysis and Discussion

Dataset. The mainstream datasets in the field of Text-to-Image are centered on large-scale image-text pairs, the most representative of which includes LAION-5B with a scale of 5.85 billion samples,

which is currently the largest dataset in the world, with a total of 100T data. At the same time, large-scale datasets such as COYO, LAION-COCO, and CC-12M are also commonly used. On the other hand, with the continuous refinement of tasks and different research priorities, some field-specific datasets have gradually been formed, including MangaZero, a comic generation annotation set, and Alchemist, a fine-tuning optimization set.

Evaluation indicators. The evaluation indicators in the field of the Wensheng diagram are divided into multiple dimensions. The evaluation indicators of distribution consistency mainly include FID↓, IS↑, SceneFID ↓, etc. The most common one is FID, which uses a pre-trained deep neural network to extract image features and evaluates the generation quality by comparing the distribution differences of these features in high-dimensional space. In the field of image and text consistency, common indicators include CLIP Score↑, and VQAScore↑, among which CLIP Score measures the similarity between a given text description and the generated image in the CLIP semantic space, often used in conjunction with FID

Table 1. Comparison table of Performance of Classic Madels

Model	Core architecture	Number of parameters	512*512 Inference speed	trainGPU-days	FID-50K↓	GenEval↑	Rendering of the Chinese text↑
Qwen-Image	MMDiT + 3D-VAE	20B	0.12s	2400	8.2	0.91	92.5%
VAR	VQ-VAE + AR Transformer	20B	0.08s	310	8.5	0.81	72%
R3GAN	GAN (RpGAN+R1+R2)	2.5M	0.01s	4	15.7	—	—
SDXL-Lightning	Latent Diffusion + LCM-LoRA	3.5B	0.04s	240	8.3	0.82	65%
InstantID	IP-Adapter + SDXL	2.5B	0.15s	80	8.4	0.75	60%
PixArt-Σ	DiT + T5-XXL	0.8B	0.15s	180	9.1	0.80	70%
DALL·E 3	Diffusion + GPT-4V	12B	0.35s	6500	6.5	0.93	88
Stable Diffusion 3.5 Large	MMDiT + Flow Matching	8B	0.20s	1200	8.1	0.85	75
Flux.1-schnell	DiT + Flow Matching	12B	0.05s	900	6.8	0.90	80
CogView 2.0	CogLM + Layering AR	7.2B	0.22s	550	9.0	0.77	85
Muse	MaskGIT + T5-XXL	3B	0.08s	720	7.9	0.87	68

Based on the information provided in Table 1, three conclusions can be drawn, corresponding to the value and future direction of the three architectures:

The diffusion model is the most versatile technical architecture at present. Almost all models with FID<8, GenEval>0.8 in the table are diffusion routes (Qwen-Image[31], SDXL-Lightning [28]). With Transformer- or Flow-matched noise prediction, diffusion architectures establish a standard of fact for visual quality and text consistency. However, the high computing power threshold has also given rise to two sub-directions: lightweight distillation (SDXL-Lightning 4 steps, Flux-schnell 4 steps) reduces 20 steps to ≤4 steps, approaches GAN, and video memory drops to less than 6 GB; The vertical model (Qwen-Image Chinese typesetting, InstantID face consistency [19]) is plugged in through LoRA [15]/ControlNet [13], proving that "big base small expert" is more economical than retraining.

The limit of pure autoregressive (VAR) efficiency is still being pushed. VARprovedthat autoregression in discrete submersible space can reach the "single-step limit" earlier than diffusion[30]. Its 310 GPU-days training cost is only 1/5 of the spread, providing a new paradigm of "controllable training and flying reasoning" for small and medium-sized teams. As LLM compression, codebook expansion, and multi-scale parallel sampling mature, AR has the potential to compete with diffusion on 1024^2 and even video frame-level tasks.

The advantage of GAN is that it is small in size and single-step generation, and it is still an irreplaceable and important tool. R3GANproves that GAN is still very cost-effective in closed domains such as faces, animations, and product images [29]. In the future, the role of GANs will shift from general generative regression to specialized fields, complementing Diffusion/AR.

4. Conclusion

Based on different mainstream technical architectures, this review expounds the technical evolution route of text-to-image generation and representative technologies in recent years, and compares and analyzes the performance of advanced models, analyzes the advantages and disadvantages of different architectures, and finds that the diffusion framework still has obvious advantages in terms of generation quality, text consistency, and other aspects. Meanwhile, other frameworks can play a role as components or in certain specific fields. For instance, VAE can improve computational efficiency as a latent space compression component and GAN, relying on its advantages in one-step generation and lightweight, can play a role in real-time video processing and interactive applications, etc. At the same time, the current model has problems such as insufficient text compatibility, lack of physical logic, copyright and security. In terms of copyright issues, in the future, optimization can be achieved through methods such as establishing and using datasets with clear Copyrights, embedding traceability information, clarifying legal frameworks for copyright ownership, and establishing compensation mechanisms. For safety issues, built-in filters and the establishment of a sound accountability mechanism are both helpful. Introducing an external knowledge base is undoubtedly effective for making the generated results conform to physical logic. At the same time, guiding users to use appropriate guidance and prompts will also be very helpful. Meanwhile, methods such as using multilingual Clips, introducing translation-based fine-tuning, and employing cross-language transfer learning can enhance multilingual adaptability. Through efforts in multiple aspects, the generation of text to image will develop in a more perfect and healthier direction.

References

[1] Zhang Han, Xu Tao, Li Hongsheng, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 5907–5915.

[2] Ramesh Aditya, Pavlov Mikhail, Goh Gabriel, Gray Scott, Voss Chelsea, Radford Alec, Chen Mark, Sutskever Ilya. Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning, 2021, 139: 8821–8831.

[3] Ding Ming, Yang Zheng, Hong Wenyi, Zheng Wendi, Zhou Chang, Yin Da, Lin Jie, Zou Xu, Shao Zhihua, Yang Hongxia, Tang Jie. CogView: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 2021, 34: 19822–19835.

[4] Ho Jonathan, Jain Ajay, Abbeel Pieter. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840–6851.

[5] Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer Björn. High-resolution image synthesis with latent diffusion models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 10684–10695.

[6] Ramesh Aditya, Dhariwal Prafulla, Nichol Alex, Chu Casey, Chen Mark. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint, arXiv:2204.06125, 2022.

[7] Radford Alec, Metz Luke, Chintala Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint, arXiv:1511.06434, 2015.

[8] Karras Tero, Laine Samuli, Aila Timo. A style-based generator architecture for generative adversarial networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 4401–4410.

[9] Kingma Diederik, Welling Max. Auto-encoding variational Bayes. arXiv preprint, arXiv:1312.6114, 2013.

[10] Sohl-Dickstein Jascha, Weiss Eric, Maheswaranathan Niru, Ganguli Surya. Deep unsupervised learning using nonequilibrium thermodynamics. International Conference on Machine Learning (ICML), Lille, France, 2015: 2256–2265.

[11] Hirschmüller Heiko. Accurate and efficient stereo processing by semi-global matching and mutual information. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2005, 2: 807–814.

[12] Song Jiaming, Meng Chenlin, Ermon Stefano. Denoising diffusion implicit models. arXiv preprint, arXiv:2010.02502, 2020.

[13] Zhang Li, Agrawala Maneesh. Adding conditional control to text-to-image diffusion models. IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 3836–3847.

[14] Mou Chong, Chen Tao, Zhang Yuxin, et al. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. AAAI Conference on Artificial Intelligence, 2024, 38: 4296–4304.

[15] Hu Edward, Shen Yelong, Wallis Phillip, et al. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR), 2022.

[16] Ruiz Nataniel, Li Yuanzhen, Jampani Varun, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 22500–22510.

[17] Peebles William, Xie Saining. Scalable diffusion models with transformers. IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 4195–4205.

[18] Chen Jie, Ge Chen, Xie Enze, Wu Yufei, Luo Ping. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. European Conference on Computer Vision (ECCV), Heidelberg: Springer, 2024: 74–91.

[19] Wang Qing, Bai Xinyu, Wang Haoran, Qin Zhen, Chen Anbang. InstantID: Zero-shot identity-preserving generation in seconds. arXiv preprint, arXiv:2401.07519, 2024.

[20] Reed Scott, Akata Zeynep, Yan Xinchen, Logeswaran Lajanugen, Schiele Bernt, Lee Honglak. Generative adversarial text to image synthesis. International Conference on Machine Learning (ICML), New York, USA, 2016: 1060–1069.

[21] Zhang Han, Xu Tao, Li Hongsheng, et al. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947–1962.

[22] Xu Tao, Zhang Han, Huang Xu, et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 1316–1324.

[23] Zhu Minfeng, Pan Pingbo, Chen Wei, Yang Yi. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 5802–5810.

[24] Tao Ming, Tang Hao, Wu Fei, Jing Xiaoyuan, Bao Bin, Xu Changsheng. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. arXiv preprint, arXiv:2008.05865, 2020.

[25] Zhang Han, Zhang Long, Xu Tao, et al. Cross-modal contrastive learning for text-to-image generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 833–842.

[26] Sauer Axel, Karras Tero, Laine Samuli, Geiger Andreas, Aila Timo. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. International Conference on Machine Learning (ICML), 2023: 30105–30118.

[27] Kang Minguk, Jeong Jiseob, Lee Jong Chul, et al. Scaling up GANs for text-to-image synthesis. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 10124–10134.

[28] Lin Sheng, Wang Anlan, Yang Xiaoqing. SDXL-Lightning: Progressive adversarial diffusion distillation. arXiv preprint, arXiv:2402.13929, 2024.

[29] Huang Norman, Gokaslan Aaron, Kuleshov Volodymyr, Prabhakaran Vinodkumar. The GAN is dead; long live the GAN! A modern GAN baseline. Advances in Neural Information Processing Systems, 2024, 37: 44177–44215.

[30] Tian Kun, Jiang Yizhe, Yuan Zhe, Peng Bin, Wang Liwei. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in Neural Information Processing Systems, 2024, 37: 84839–84865.

[31] Wu Chenfei, Ji Yong, Zhang Shuhuai, et al. Qwen-image technical report. arXiv preprint, arXiv:2508.02324, 2025.