Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33192

Applications and Development of Generative Adversarial Networks in Text-to-Image Synthesis

Yijia Fu

Faculty of Data Science, City University of Macau, Macau, China
D24090107503@cityu.edu.mo

Abstract. In text processing, visual images are usually generated from textual descriptions, and computer systems can perform similar tasks. As a core task in the field of multimodal generation, text-to-image synthesis aims to generate high-quality images that are visually realistic and semantically consistent based on natural language descriptions. Therefore, this paper reviews the application and development of Generative Adversarial Networks (GANs) in text-to-image synthesis, traces their technological evolution, analyzes key breakthroughs, core modules, and current bottlenecks, and proposes future research directions. This paper explores recent literature to outline the practical progress of GANs, covering fundamental principles, technological innovations, and module optimizations. Additionally, it conducts an in-depth analysis of technical bottlenecks such as cross-modal semantic mapping, detail generation, and training stability, and reviews optimization strategies and future research trends. The results indicate that though diffusion models have dominated in recent years in terms of generation quality, handling of complex scenes, and semantic consistency, GANs still perform better in terms of inference speed, fine-grained control, and training efficiency on large-scale datasets. However, GANs face challenges such as training instability, limited comprehension of long texts, attribute binding errors, and insufficient high-resolution detail.

Keywords: Generative Adversarial Networks, text-to-image synthesis, semantic consistency, visual realism, multimodal fusion

1. Introduction

Text-to-image (T2I) synthesis is one of the core challenges at the intersection of computer vision and natural language processing, which generates realistic and semantically consistent images based on natural language descriptions. With advances in multimodal learning, methods like Generative Adversarial Networks (GANs) and diffusion models have gradually achieved significant research results.However, the vast disparity between text and image modalities continues to pose challenges for cross-modal semantic mapping, manifesting as incorrect attribute binding, object misalignment, and insufficient understanding of complex, lengthy texts [1-3]. Moreover, errors tend to accumulate during multi-stage generation, leading to detail loss and content inconsistencies in generated images, while the inherent instability of GAN training limits generative diversity and scalability [4-6]. Although diffusion models have dominated in terms of generation quality, complex scene handling, and semantic consistency due to their iterative denoising mechanism, issues such as slow inference

© 2026 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/).

DOI: 10.54254/2755-2721/2026.GU33192

speed and weak controllability remain unresolved, leaving room for further optimization of GAN methods [7]. By combining literature review and case analysis, this paper traces the development of GANs in the T2I field, examines current research gaps, and outlines future optimization directions. It focuses on enhancing inference efficiency, semantic consistency, and generative diversity through the integration of GANs with diffusion models and large language models, aiming to advance T2I tasks toward greater intelligence and efficiency.

2. Basic concepts and techniques of generative adversarial networks

2.1. Basic principles and training process of GANs

GANs achieve data generation through the adversarial training of a generator and a discriminator [1]. The generator generates samples from random noise, mimicking the distribution of real data, while the discriminator evaluates the authenticity of the samples, distinguishing between generated and real samples. Through game-theoretic optimization, the generator and discriminator iteratively refine each other, producing samples that increasingly resemble real data. Specifically, the training objective of a GAN is to optimize the following value functions:

$$\min_{D}\max_{G} V\left(D,G\right)=\mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x)\right]+\mathbb{E}_{z\sim p_{z}(z)}\left[\log(1-D(G(z)))\right] \tag{1}$$

where G is the generator, D is the discriminator, x represents a real sample, and z is random noise.

In T2I tasks, GANs are extended to conditional generative adversarial networks (cGANs). In this case, the generator not only relies on random noise but also generates images based on textual descriptions. The discriminator not only judges whether an image is real but evaluates the semantic consistency between the image and the textual description. Therefore, in T2I tasks, GANs not only generate visually realistic images but also ensure semantic alignment between the image and the text. Figure 1 illustrates the Text-Conditional cGAN architecture, integrating text embeddings with image feature maps to guide image generation [3].

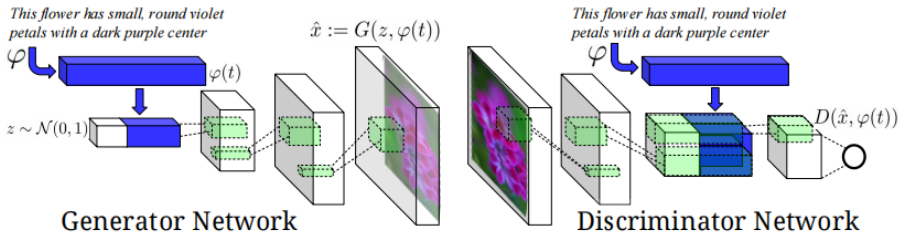


Figure 1. Architecture of the text-conditional cGAN

2.2. Major variants and technological evolution of GANs

In T2I synthesis, GAN variants tackle training instability, improve image quality and diversity, and have become essential tools. Early variants focused on fundamental stability and image quality. For instance, Wasserstein GAN enhanced training stability through the Earth-Mover distance loss and gradient penalties, while DCGAN introduced convolutional architectures to improve the quality of unsupervised image generation [2]. Furthermore, cGAN and Auxiliary Classifier GANs (ACGAN) further incorporated text conditions to generate images aligned with the descriptions.

DOI: 10.54254/2755-2721/2026.GU33192

In T2I tasks, subsequent research has further focused on semantic consistency and detail control. StackGAN uses a multi-stage stacked architecture to progressively enhance image resolution, while AttnGAN introduces an attention mechanism to strengthen word-level semantic alignment [4,5]. In addition, MirrorGAN and ControlGAN further optimize image-text alignment via re-description and word-level control, whereas DF-GAN adopts a single-stage deep fusion architecture to simplify the model and improve training efficiency [6,8,9]. More recently, the StyleGAN series (combined with CLIP), StyleGAN-T, and GigaGAN generate high-quality images at 10B parameters. Through style mapping, adaptive instance normalization, large-scale training, and efficient attention, these models boost inference speed, controllability, and performance on large datasets [10-12]."

2.3. Performance evaluation and optimization methods for GANs

In T2I tasks, GAN performance is measured using FID and CLIP Score. FID measures the distance between the feature distributions of generated images and real images, primarily evaluating visual realism and diversity. Besides, CLIP Score calculates the similarity between the embedding vectors of images and text descriptions, further evaluating semantic consistency and the alignment of object attributes and scene relationships.

As GAN models for T2I tasks have evolved, performance on the CUB bird dataset has steadily improved. Early GAN-INT-CLS set a baseline with an FID of 186 for semantic matching [3]. Based on this, AttnGAN introduced an attention mechanism, reducing FID to around 68 and achieving a CLIP Score of 0.29 [5]. Then, DF-GAN further optimized performance through single-stage fusion, hence lowering FID to approximately 45 and increasing the CLIP Score to 0.37 [6]. More recently, StyleGAN-T and GigaGAN improved FID to about 42 and 41.2, respectively, while significantly boosting inference speed and overall efficiency [10,11]. Despite these gains in visual realism and semantic consistency, current evaluation methods have limitations: FID does not directly capture semantic misalignment, and CLIP Score struggles with fine-grained details or long, complex texts. Future work should refine matching-aware discriminators to further lower FID and add a CLIP alignment loss to enhance semantic consistency, balancing visual and semantic quality.

3. Key issues and technical bottlenecks in text-to-image synthesis

3.1. Cross-modal semantic mapping and consistency challenges

Text and images differ inherently in semantic mapping: text is made of discrete symbols, whereas images convey continuous visual information. Early T2I models used global sentence vectors as conditional inputs but failed to effectively capture word-level details, thus leading to discrepancies between generated images and textual descriptions. In complex scenes, generated images often misrepresent the text [3].

To address the cross-modal semantic mapping problem, AttnGAN introduced a deep attention mechanism, enabling the generator to focus on regions in the image relevant to the text keywords, improving semantic consistency. When handling long texts and complex descriptions, issues such as attribute binding errors and object misalignment persist, particularly when multiple elements and details are involved, limiting the accuracy and consistency of the generated images. For example, bird images generated by early T2I models do not match the textual descriptions in terms of color and shape, exposing the challenges of cross-modal mapping, as shown in Figure 2 [3]. And Figure 3 shows that, although AttnGAN's attention mechanism helps, issues like background blur and shape misalignment remain at the low-resolution stage [5].

DOI: 10.54254/2755-2721/2026.GU33192

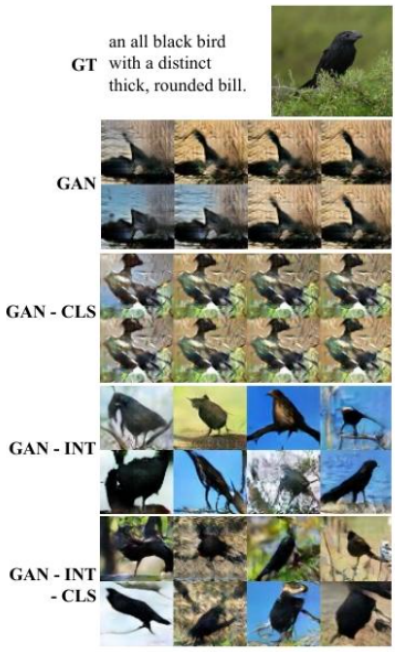


Figure 2. Deviations in bird images generated by early T2I models



Figure 3. AttnGAN-generated images and attention visualizations

By modeling object layouts, Obj-GAN cuts attribute errors and misalignment in T2I generation [14]. However, with complex text, alignment between generated images and descriptions remains imperfect, especially for detailed images. CLIP improves image-text alignment by mapping them into a shared semantic space, but on large-scale datasets, its performance still needs regularization to ensure consistency and stability [10,11].

3.2. Detail generation and content consistency issues

High-resolution image generation with consistent details has long been a challenge for GANs in T2I synthesis, particularly in multi-stage processes where errors accumulate and details degrade. To address this challenge, StackGAN proposed a two-stage generation strategy: first generating rough shapes and colors, then refining textures and details. However, error accumulation between stages during generation often reduces the quality of the final image. Figure 4 shows that images generated by StackGAN still display coarse shapes and inconsistent textures during refinement. In contrast, DM-GAN alleviates this issue using a dynamic memory module, but errors persist [13].

DOI: 10.54254/2755-2721/2026.GU33192

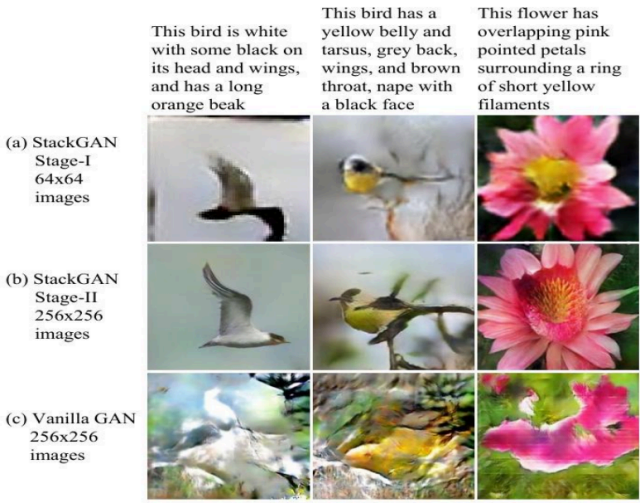


Figure 4. Comparison between single-stage GAN and multi-stage StackGAN (a-b: StackGAN: error accumulation, rough shapes; c: single-stage GAN: blurriness, lack of detail)

To address this issue, DF-GAN introduces a single-stage deep fusion architecture that avoids entanglement among multiple generators and directly produces high-resolution images through deep text-image fusion blocks. As shown in Figure 5, DF-GAN effectively mitigates error accumulation common in multi-stage methods, significantly enhancing the realism of image details and content consistency. Compared to traditional multi-stage methods, DF-GAN achieves superior performance in both detail accuracy and alignment between images and text descriptions [6].

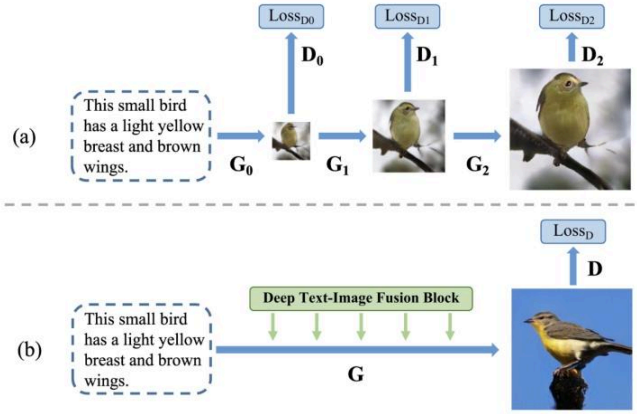


Figure 5. Comparison between multi-stage generators and DF-GAN (a: multi-stage generator, error accumulation, blurry images; b: DF-GAN, avoids error accumulation, generates high-resolution images)

3.3. Bottlenecks in training stability and generative diversity

Training stability and generative diversity in GANs for T2I tasks have long been major bottlenecks. Though early and recent optimization methods have achieved some breakthroughs, they still suffer from mode collapse and training instability under large-scale training and complex conditions.

Early optimization strategies partially alleviated these issues. For example, StackGAN enhances generative diversity and prevents mode collapse by adding noise to text embeddings to smooth the conditional manifold [4]. Meanwhile, F-GAN reduces oscillations and improves training stability by

DOI: 10.54254/2755-2721/2026.GU33192

applying a matching-perception gradient penalty [6]. With the emergence of large-scale models such as StyleGAN-T and GigaGAN, which have achieved stable training of models with hundreds of billions of parameters on massive datasets like LAION-2B, the landmark breakthrough lies in architectural innovations [4,10]. Specifically, StyleGAN-T gradually constructs feature hierarchies via enhanced text alignment mechanisms, multi-scale incremental training, and adaptive instance normalization, avoiding gradient explosion or vanishing gradients while preserving the transmission of conditional information, which enhances convergence stability. Besides, GigaGAN introduces a filter bank and linear combination mechanism to expand the generator's feature capacity without increasing parametric complexity. Combined with cross-attention, alternating self-attention layers, and a multi-scale upsampling architecture, GigaGAN effectively suppresses pattern collapse and generates richer visual variations. Through efficient feature interaction, GigaGAN reduces training oscillations and maintains high-diversity outputs within an inference time of 0.13 seconds per image. These methods expand representation (multi-scale, filters) and strengthen feedback (attention, gradient penalties) to stabilize training and avoid collapse. These methods still need large datasets, and for long texts or rare prompts, the balance between diversity and stability is limited.

4. Optimization strategies for generative adversarial networks in T2I synthesis

4.1. Optimization strategies based on task requirements

To improve T2I generation, optimization strategies target fine-grained control, deep modal fusion, efficient inference, and the trade-off between quality and speed. For fine-grained control, AttnGAN introduces an attention mechanism that allows the generator to dynamically adjust the focus area based on keywords in the text. For instance, when the text specifies a "red beak," the model directs attention to the corresponding region, thereby enhancing semantic alignment accuracy. This ensures that the generated images better meet expectations, particularly when processing complex text. To strengthen the deep integration of text and image information, DF-GAN designs a dedicated fusion module. By performing deep-level text-image fusion, it avoids information loss and greatly boots cross-modal consistency [6,10]. In terms of inference efficiency, GigaGAN employs a multi-stage progressive upsampling architecture combined with self-attention and cross-attention mechanisms, reducing the inference time for a single image to 0.13 s, thereby markedly improving efficiency. CPGAN enhances the matching accuracy between text and images by parsing text structures while maintaining latent space editing capabilities, balancing speed and controllability [15]. StyleGAN-T enhances the text alignment mechanism within the StyleGAN framework, boosting inference speed to 10 frames per second and outperforming diffusion models at equivalent speeds on performance metrics such as FID [8]. These optimization measures improve GAN performance in text-to-image synthesis tasks through semantic refinement, modal fusion, and inference efficiency.

4.2. Architecture- and training-based optimization methods

Improvements in architectural design and training methods have mainly enhanced GANs' image quality, semantic consistency, training stability, and inference speed. For example, in architectural design, CLIP is used as the text encoder, and together with an alignment loss, it resolves semantic ambiguity and text-image misalignment issues [9,15]. The use of multi-scale discriminators and generators allows the model to capture both global structure and local details, thereby improving image quality. GigaGAN, by incorporating a filter bank and linear combination methods, expands the generator's feature capacity while enhancing computational efficiency [10]. The alternating use

DOI: 10.54254/2755-2721/2026.GU33192

of cross-attention and self-attention mechanisms strengthens feature interactions both within and across modalities, thereby improving the semantic consistency of images. During training, DF-GAN stabilizes the training process via conditional augmentation and matching-aware gradient penalties, increasing the diversity of generated samples and effectively reducing the risk of mode collapse [6]. Meanwhile, GigaGAN's multi-scale training method gradually increases resolution, ensuring the quality of the final images [10]. Besides, DF-GAN introduces a Target-Aware discriminator, which focuses on the core targets in text descriptions, enhances the discriminator's sensitivity to semantic matching, and further optimizes the generation results [6]. These architectural and training method optimizations complement each other, improving GAN performance in terms of generation quality, semantic consistency, and inference speed, and even matching the performance of diffusion models in certain scenarios.

5. Conclusion

This paper examines the applications and advancements of GANs in text-to-image synthesis tasks, detailing their technological evolution, key modules, major challenges, and optimization strategies, while evaluating their performance improvements. These findings show that GANs are advancing rapidly in this field. While diffusion models currently lead in generation quality and handling complex scenarios, GANs retain distinct advantages in inference speed, fine-grained controllability, and efficiency in large-scale training. Yet, GANs still struggle with long texts, attribute accuracy, and ultra-high-resolution details, including limited coverage and a lack of experimental validation. Future research could explore hybrid GAN-diffusion architectures, improve large language model text encodings, and pursue large-scale multimodal pre-training to better balance quality, speed, and controllability, thereby advancing practical applications.

References

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative Adversarial Nets. arXiv: 1406.2661.
[2] Kang, M., Zhu, J. Y., Zhang, R., et al. (2023). Scaling up GANs for Text-to-Image Synthesis. arXiv: 2303.05511.
[3] Liu, M., Ma, Y., Yang, Z., et al. (2024). LLM4GEN: Leveraging Semantic Representation of LLMs for Text-to-Image Generation. arXiv: 2407.00537.
[4] Li, B., Qi, X., Lukasiewicz, T., et al. (2019). Obj-GAN: Object-driven Text-to-Image Synthesis via Adversarial Learning. arXiv: 1905.08356.
[5] Li, W., Zhang, P., Zhang, L., et al. (2019). ControlGAN: Controllable Text-to-Image Generation with Control over Word Importance. arXiv: 1910.09388.
[6] Patashnik, O., Wu, Z., Shechtman, E., et al. (2021). StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arXiv: 2103.17249.
[7] Sauer, A., Karras, T., Laine, S., et al. (2023). StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. arXiv: 2301.09515.
[8] Tao, M., Tang, H., Wu, F., et al. (2022). DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. arXiv: 2008.05865.
[9] Xu, T., Zhang, P., Huang, Q., et al. (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. arXiv: 1711.10485.
[10] Qiao, T., Zhang, L., Zhang, J., et al. (2019). MirrorGAN: Learning Text-to-Image Generation by Redescription. arXiv: 1903.05854.
[11] Reed, S., Akata, Z., Yan, X., et al. (2016). Generative Adversarial Text to Image Synthesis. arXiv: 1605.05396.
[12] Zhang, H., Xu, T., Li, H., et al. (2017). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. arXiv: 1612.03242.
[13] Zhu, M., Pan, P., Chen, W., et al. (2019). DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. arXiv: 1904.01310.

DOI: 10.54254/2755-2721/2026.GU33192

[14] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv: 1511.06434.
[15] Liang, J., Pei, W., & Lu, F. (2020). CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis. ECCV 2020.