Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30639

Research Progress of GAN in Image Generation

Zhijie Zhang

College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
2022110187@email.szu.edu.cn

Abstract. Since their introduction in 2014, Generative Adversarial Networks (GANs) have profoundly advanced the field of computer vision. Research on their applications primarily centers on image generation, which is underpinned by the core concept of an adversarial game between generator and discriminator networks. This dynamic process progressively refines the model's ability to distinguish real from synthetic data and to generate highly realistic images, ultimately driving the entire system toward a Nash equilibrium. However, the image quality generated by early GAN models was subpar and the structure was prone to incoherent, particularly at high resolutions. Moreover, there were issues such as training instability and model collapse. These challenges, in turn, motivated and propelled further in-depth research into GANS. Grounded in the task paradigms of GANs for image generation, this review systematically summarized the developmental trajectory and ongoing challenges from 2014 to 2025 in critical domains including generator design, loss functions, training stabilization and cross-modal or multimodal synthesis. It also proposes two relatively new research directions: interactive editing (e.g., DragGAN) and adversarial distillation of diffusion models (e.g., Diffusion-GAN), thereby offering references for scholars conducting research on image generation. Furthermore, this paper finds that there are numerous metrics for evaluating generative models, such as FID and IS. How to establish a unified evaluation system and conduct qualitative and quantitative analysis of models is also one of the key research directions.

Keywords: Image Generation, Generative Adversarial Network, Conditional Generation, Training Stability, Adversarial Distillation

1. Introduction

Image generation, a highly active area in computer vision, holds immense potential across diverse industries. For instance, Conditional Diffusion Models (DMs) have been employed in medical imaging to generate synthetic medical samples [1]. A representative case is the work by Akrout et al., who trained conditional DMs on a skin lesion dataset to produce augmented samples for various lesion types [2]. The early development of generative models based on deep neural networks (DNNs) was relatively slow. Pioneering architectures such as Restricted Boltzmann Machines (RBMs) [3] and Deep Belief Networks (DBNs) [4] often produced samples with limited quality. However, a significant shift occurred with the introduction of Variational Autoencoders (VAEs) [5] and Autoregressive Models (AR) [6], which ushered in a new era for deep generative modeling.

DOI: 10.54254/2755-2721/2026.TJ30639

These models leverage unsupervised or self-supervised learning paradigms to effectively learn the underlying probability distributions of training data, such as images and text. While likelihood-based generative models offer reliable training with well-defined objective functions and stable image generation capabilities, they face inherent limitations: AR suffers from sequential generation constraints that severely impede processing speed and is non-parallelizable [6]. Whereas VAE, while faster, often produces perceptually blurred outputs [7]. Furthermore, since the objective revolves around approximating the overall data distribution, generated samples tend to exhibit conservative characteristics.

In 2014, Ian Goodfellow et al., drawing from game-theoretic principles, proposed the Generative Adversarial Network (GAN) [8], a framework that fundamentally reformed the paradigm of image generation. This adversarial training scheme enables implicit learning of data distributions by engaging a generator and a discriminator in a dynamic game, driving the system toward a Nash equilibrium. Notably, GANs bypass the computationally intractable probability density estimation required by traditional deep generative models based on maximum likelihood estimation [8]. This paper outlines three core task paradigms in GANs: unconditional image generation, conditional image generation, and cross-modal and multimodal synthesis. It reviews architectural advancements in generator design, loss functions and training optimization, as well as progress in cross-modal generation. Furthermore, two emerging research directions—interactive editing and adversarial distillation of diffusion models—are highlighted to guide future research focus in this domain.

2. GANs task paradigms

2.1. Unconditional image generation

Unconditional image generation constitutes the foundational paradigm for image synthesis models. In this framework, a model operates independently of any external conditional information, relying solely on a noise vector z sampled from a prior distribution (e.g., Gaussian distribution or uniform distribution). The generator (G) maps this input to a complete image, $x = G(z)$. Training follows a minimax adversarial game between G and a discriminator D: D is trained to distinguish real from generated samples, while G is optimized via an adversarial loss to produce increasingly realistic outputs until D can no longer tell them apart [8]. The content and style of the generated images are governed by the data distribution captured by the model, exhibiting inherent stochasticity. While the architecture and training workflow of unconditional models are relatively simple, they lack semantic control over the output. Representative models include the original GAN and DCGAN [8,9].

2.2. Conditional image generation

Conditional image generation establishes an extensible core paradigm for cross-modal and multimodal synthesis models. This framework incorporates not only random noise z, but also conditional information c, such as text prompts, class labels or semantic maps. The model first encodes these non-visual conditions into representation vectors, which are then fused with z. For instance, text inputs are commonly vectorized using CLIP-ViT, BERT, or LSTM. The generator G(z, c) learns to merge z with c to synthesize novel images aligned with the input conditions [10]. This paradigm significantly mitigates generation randomness and enables preliminary controllable image synthesis. However, it faces persistent alignment challenges, particularly with complex or abstract textual descriptions, and generally requires large-scale, accurately-annotated datasets. Representative models include cGAN [10], Pix2Pix [11], and StyleGAN2 [12].

DOI: 10.54254/2755-2721/2026.TJ30639

2.3. Cross-modal and multi-modal synthesis

The core idea of cross-modal synthesis is to establish a shared and aligned semantic space across different modalities. Taking text-to-image generation as an example, the framework generally comprises five key stages: (1) Encoding, (2) Alignment, (3) Conditioning Injection, (4) Generation, and (5) Evaluation & Refinement (typically assessed via CLIP-Score [13]). Specifically, an encoder (e.g., CLIP's text encoder) first maps the source modality (text prompts) into a high-dimensional semantic embedding. The model is then trained to align these textual embeddings with corresponding image embeddings (extracted by an image encoder) within a unified latent space. Subsequently, this aligned representation guides the generative process of a downstream model, such as GAN or diffusion model, to synthesize the final image.

Multimodal synthesis, as a natural extension of cross-modal synthesis, further involves parallel encoding of multiple modalities, fusion into a unified vector, and multi-conditional injection, all operating within a cohesive modality space to synthesize outputs such as images or videos. This approach demonstrates strong performance in applications including visual planning, complex scene generation, and three-dimensional generation [14].

Cross-modal synthesis or multimodal synthesis is characterized by cross-domain transfer capabilities. Nevertheless, the challenge of modality alignment remains [14], and performance depends heavily on high-quality, well-paired multimodal datasets [15]. Additionally, how to objectively and quantitatively evaluate the correspondence between generated outputs and input conditions remains an open research question. Representative models include CMCGAN [15] and CI-GAN [16].

3. The progression of GANs: from core innovations to frontier expansion

3.1. Architecture evolution and training optimization

3.1.1. The progression of generator architectures

Original GANs implicitly model data distributions through adversarial training, effectively circumventing the blurriness issues in VAE and the slow sequential generation of AR models. However, their generators, primarily composed of fully connected layers, were incapable of synthesizing high-resolution images with complex spatial structures. In 2015, Radford et al. introduced Deep Convolutional Generative Adversarial Networks (DCGAN) [9], which established key architectural principles by incorporating convolutional layers and Batch Normalization (BN), thereby laying the groundwork for stable training and improved image quality. Nonetheless, directly training high-resolution GAN models remains highly prone to instability, often leading to training collapse and structurally incoherent large-scale images.

In 2017, Karras et al. introduced Progressively Growing Generative Adversarial Networks (PGGAN) [17]. This work proposed a progressive growing strategy, starting training at a low resolution (4×4) and gradually adding layers to both the generator and discriminator as training stabilized. This approach yielded a breakthrough by stably generating high-resolution 1024×1024 facial images for the first time, representing a significant milestone in image generation quality. Despite this achievement, the model's generative process remained largely opaque, offering no fine-grained, semantic control over image attributes such as specific facial features or hairstyles.

In 2018, Karras et al. introduced A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN) [18], which revolutionized the generator design. The architecture

DOI: 10.54254/2755-2721/2026.TJ30639

employs a mapping network to transform the input noise z into an intermediate latent code w, which is then fed into each generator layer via Adaptive Instance Normalization (AdaIN). This mechanism enables control over image attributes at different scales. Ideally, each dimension in this latent space would control an independent semantic factor, thereby enhancing the diversity and controllability of the generated images and mitigating mode collapse [19]. However, the disentanglement in StyleGAN's latent space is typically only partial. Complex nonlinear couplings between attributes persist, often leading to attribute entanglement and visual artifacts during editing operations. For example, unintentionally altering head pose when attempting to modify a smile, or causing facial contour distortions when changing the hairstyle [19].

3.1.2. The optimization of loss functions

In the original GAN formulation, an optimal discriminator causes the generator's loss function to be equivalent to minimizing the Jensen-Shannon (JS) divergence between the real and generated distributions [8]. A paramount limitation arises when these distributions have disjoint supports, leading to vanishing gradients and unstable training. To address these limitations, Martin Arjovsky et al. introduced the Wasserstein GAN (WGAN) [20], which replaces the JS divergence with the Wasserstein distance. This metric provides stable gradients even for non-overlapping distributions, thereby significantly mitigating mode collapse and enhancing training stability. However, the initial WGAN enforced the requisite Lipschitz continuity via a simplistic weight clipping strategy, which often introduced optimization difficulties.

To address the limitations of weight clipping, there are two main directions. One prominent line of work is the Wasserstein GAN with Gradient Penalty (WGAN-GP) proposed by Gulrajani et al. [21]. It replaces weight clipping with a gradient penalty that enforces the discriminator's gradient norm to approximate unity at interpolated points between real and generated data, thereby satisfying the Lipschitz constraint. This method effectively improves training stability and produces higher-quality images while retaining the benefits of the WGAN framework. However, this comes at the cost of increased computational overhead, as the gradient penalty requires additional forward and backward passes for the interpolated points.

An alternative approach, Spectral Normalization (SN), was proposed by Miyato et al. [22]. This technique mathematically enforces a Lipschitz constraint by normalizing the spectral norm (the largest singular value) of each weight matrix in the discriminator, effectively bounding the Lipschitz constant of each layer to 1 and controlling model complexity. While SN achieves higher computational efficiency and is more suitable for rapid training, its stringent hard constraint can potentially over-limit the discriminator's representational capacity.

3.2. Cross-modal generation and fusion

The pioneering Generative Adversarial Network - INTerpolation - CLSifier (GAN-INT-CLS) framework, introduced by Reed et al. in 2016 [15], addressed the fundamental challenge of text-to-image synthesis. Building upon the Conditional Generative Adversarial Network (cGAN) architecture, it integrated a matching-aware discriminator (GAN-CLS) and manifold interpolation learning (GAN-INT). The model utilized a pretrained Convolutional Neural Network - Recurrent Neural Network (CNN-RNN) network to encode text descriptions into vectors, which were then injected into both the generator and discriminator. However, a primary limitation was its generation of low-resolution images lacking fine detail.

DOI: 10.54254/2755-2721/2026.TJ30639

To address the challenges, Zhang et al. (2017) proposed Stacked Generative Adversarial Networks (StackGAN) [23]. This model employs a two-stage generation architecture: the Stage-I GAN generates low-resolution sketches from textual descriptions, while the Stage-II GAN refines the image through inpainting and detail enhancement, conditioned jointly on the initial sketch and Gaussian-noise-embedded text encodings, ultimately producing 256×256 resolution images. StackGAN pioneered high-definition text-to-image synthesis; however, its control mechanism remains implicit, typically encoding the entire sentence into a single global variable. This approach results in limited granularity in controlling object layouts within complex scenes, leading to coarse-grained generation.

To achieve finer-grained control, Xu et al. introduced the Attention Generative Adversarial Network (AttnGAN) [24], a milestone model in the field of text-to-image synthesis. It was the first framework to incorporate the attention mechanism into GANs, achieving word-level alignment between text and image. A key innovation is the introduction of the Deep Attentional Multimodal Similarity Model (DAMSM), a dedicated loss module that segments the entire image into small regions and encodes them into feature vectors, while the text description is encoded into sentence-level and word-level embeddings. These three components are projected into a shared semantic space, where the similarity between each image region feature and all word features is computed. Although AttnGAN substantially improves the realism of image details [25], the attention mechanism tends to focus excessively on local semantic correlations, which may lead to insufficient control over global structure and result in distorted object shapes or disproportional compositions.

In 2021, Zhang et al. from Google Research proposed the Cross-Modal Contrastive Generative Adversarial Network (XMC-GAN) [26]. The model, which introduces a novel lose function, the Cross-Modal Contrastive Loss (CMCL), is trained with a novel objective that combines inter-modal (image-text) and intra-modal (image-image) contrastive losses, pulling positive sample pairs together and pushing negative samples apart in a shared feature space to enforce stronger global semantic alignment. While this mechanism significantly enhances the model's understanding and constraint of global semantics in images, the need for substantial negative samples in contrastive learning imposes a considerable computational burden.

Subsequently, Kang et al. introduced the Giga-scale Generative Adversarial Network (GigaGAN) [27], which achieved a fundamental breakthrough in generator architecture. This model employs a staged, modular generator and progressive training to stabilize billion-parameter-scale learning. GigaGAN innovatively replaces localized spatial noise injection (e.g., per-pixel noise as in StyleGAN2) with a global style-based modulation mechanism. This design mitigates unnatural local texturing and yields more diverse images that are closely aligned with text semantics. While retaining a single-step generation paradigm and leveraging modular design, GigaGAN achieves inference hundreds of times faster than diffusion models and supports near-lossless super-resolution upscaling.

3.3. Research frontiers

3.3.1. Interactive editing

Proposed in 2023 through a collaboration between the Max Planck Institute, MIT, and other institutions, Drag Generative Adversarial Network (DragGAN) [28] marks a new frontier in GAN research: interactive image generation and editing. Its core mechanism operates as follows: when a user drags a point p on an image to a target t, the model first locates the corresponding feature point f_p within the generator's feature maps. It then iteratively refines the latent code w via

DOI: 10.54254/2755-2721/2026.TJ30639

backpropagation using a Motion Supervision loss. This loss minimizes the L2 Norm between the target t and the feature point f_p along with its local neighborhood, precisely driving the feature point toward its target over iterations. A key insight is that the movement of a single point in feature space corresponds to a continuous and natural deformation of the entire object (e.g., an eye) in the image space. Furthermore, the generator automatically infers plausible Three-dimensional geometry and texture, effectively avoiding the artifacts and fragmentation common in pixel-level editing. DragGAN thus enables precise manipulation of poses and expressions in portraits and animal figures, with promising applications in areas like product design.

3.3.2. Adversarial distillation of diffusion models

Adversarial distillation effectively combines the paradigmatic strengths of diffusion models and GANs which distills knowledge from a powerful, pre-trained diffusion model (the teacher model) into a generator (the student model) capable of single or few-step generation.

In this setup, the teacher model remains fixed, offering a stable learning target, while the student model typically adopts a U-Net-based generator architecture. Representative works include Diffusion-Generative Adversarial Network (Diffusion-GAN) [29], which formulates single-step denoising as an adversarial task, and the Adversarial Diffusion Distillation (ADD) framework [30], which produced the SDXL-Turbo model for high-quality, single-step image synthesis. By integrating the superior generation quality and training stability of diffusion models with the fast inference speed of GANs, the adversarial distillation paradigm constitutes a notable milestone in generative AI, marking a paradigm shift from model competition to complementary model fusion [31].

4. Conclusion

This paper provides a systematic review of the evolutionary trajectory and cutting-edge advancements of Generative Adversarial Networks (GANs) in the field of image generation. It encompasses the paradigm shifts from unconditional and conditional generation to cross-modal synthesis, while offering an in-depth analysis of the GAN family's development—spanning foundational architectures, training stabilization, and multimodal integration.

Originating from game-theoretic principles, GANs have evolved into a diverse family of models that have profoundly shaped the methodological perspectives in image generation. Numerous practical models, such as DragGAN for interactive editing, have emerged as a result. Moreover, the highly efficient single-step generation capability of GAN-based frameworks facilitates the advancement of hybrid models, including Diffusion-GAN.

However, the absence of rigorous theoretical guarantees for Nash equilibrium existence and convergence remains a fundamental challenge in GANs, leading to persistent issues including mode collapse and training instability. Meanwhile, the absence of standardized evaluation protocols is another significant obstacle—existing metrics such as FID, IS, and Precision/Recall each capture different aspects of generative performance without offering a unified perspective. A critical direction for future work lies in establishing a comprehensive evaluation framework that effectively integrates both qualitative assessment and quantitative analysis.

References

[1] Alimisis P, Mademlis I, Radoglou-Grammatikis P, et al. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions [J]. Artificial Intelligence Review,

DOI: 10.54254/2755-2721/2026.TJ30639

2025, 58(4): 112.
[2] Akrout M, Gyepesi B, Holló P, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images [C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland, 2023: 99-109.
[3] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. Neural computation, 2002, 14(8): 1771-1800.
[4] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18(7): 1527-1554.
[5] Kingma D P, Welling M. Auto-encoding variational bayes [J]. arXiv preprint arXiv: 1312.6114, 2013.
[6] Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks [C]//International conference on machine learning. PMLR, 2016: 1747-1756.
[7] Zhao S, Song J, Ermon S. Towards deeper understanding of variational autoencoding models [J]. arXiv preprint arXiv: 1702.08658, 2017.
[8] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [J]. Advances in neural information processing systems, 2014, 27.
[9] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. arXiv preprint arXiv: 1511.06434, 2015.
[10] Mirza M, Osindero S. Conditional generative adversarial nets [J]. arXiv preprint arXiv: 1411.1784, 2014.
[11] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.
[12] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.
[13] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International conference on machine learning. PmLR, 2021: 8748-8763.
[14] Balaji Y, Nah S, Huang X, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers [J]. arXiv preprint arXiv: 2211.01324, 2022.
[15] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis [C]//International conference on machine learning. Pmlr, 2016: 1060-1069.
[16] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans [C]//International conference on machine learning. PMLR, 2017: 2642-2651.
[17] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation [J]. arXiv preprint arXiv: 1710.10196, 2017.
[18] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.
[19] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.
[20] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks [C]//International conference on machine learning. PMLR, 2017: 214-223.
[21] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans [J]. Advances in neural information processing systems, 2017, 30.
[22] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks [J]. arXiv preprint arXiv: 1802.05957, 2018.
[23] Zhang H, Xu T, Li H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks [C]//Proceedings of the IEEE international conference on computer vision. 2017: 5907-5915.
[24] Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1316-1324.
[25] Hu Y, Kim D H. Research on Controllable Image Generation Technology for Chinese Text [J]. Journal of Image and Graphics, 2025, 13(4).
[26] Zhang H, Koh J Y, Baldridge J, et al. Cross-modal contrastive learning for text-to-image generation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 833-842.
[27] Kang M, Zhu J Y, Zhang R, et al. Scaling up gans for text-to-image synthesis [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10124-10134.
[28] Pan X, Tewari A, Leimkühler T, et al. Drag your gan: Interactive point-based manipulation on the generative image manifold [C]//ACM SIGGRAPH 2023 conference proceedings. 2023: 1-11.

DOI: 10.54254/2755-2721/2026.TJ30639

[29] Xiao Z, Kreis K, Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs [J]. arXiv preprint arXiv: 2112.07804, 2021.
[30] Sauer A, Lorenz D, Blattmann A, et al. Adversarial diffusion distillation [C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 87-103.
[31] Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications [J]. ACM computing surveys, 2023, 56(4): 1-39.