INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

REVIEW ARTICLE

OPEN ACCESS

Received: 29-11-2022

Accepted: 17-01-2023

Published: 17-02-2023

Citation: Mehmood R, Bashir R, Giri KJ (2023) Deep Generative Models: A Review. Indian Journal of Science and Technology 16(7): 460-467. https://doi.org/10.17485/IJST/v16i7.2296

* **Corresponding author**.

rumaan.bashir@islamicuniversity.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Deep Generative Models: A Review

Rayeesa Mehmood[1], Rumaan Bashir[1]*, Kaiser J Giri[1]

1 Department of Computer Science, Islamic University of Science &Technology, Kashmir, Jammu and Kashmir, India

Abstract

Objectives: To provide insight into deep generative models and review the most prominent and efficient deep generative models, including Variational Auto-encoder (VAE) and Generative Adversarial Networks (GANs). **Methods:** We provide a comprehensive overview of VAEs and GANs along with their advantages and disadvantages. This paper also surveys the recently introduced Attention-based GANs and the most recently introduced Transformer based GANs. **Findings:** GANs have been intensively researched because of their significant advantages over VAE. Furthermore, GANs are powerful generative models that have been widely employed in a variety of fields. Though GANs have a number of advantages over VAEs, but, despite their immense popularity and success, training GANs is still difficult and has experienced a lot of setbacks. These failures include mode collapse, where the generator produces the same set of outputs for various inputs, ultimately resulting in the loss of diversity; non-convergence due to oscillatory and diverging behaviors of the generator and discriminator during the training phase; and vanishing or exploding gradients, where learning either ceases to occur or occurs very slowly. Recently, some attention-based GANs and Transformer-based GANs have also been proposed for high-fidelity image generation. **Novelty:** Unlike previous survey articles, which often focus on all DGMs and dive into their complicated aspects, this work focuses on the most prominent DGMs, VAEs, and GANs and provides a theoretical understanding of them. Furthermore, because GAN is now the most extensively used DGM being studied by the academic community, the literature on it needs to be explored more. Moreover, while numerous articles on GANs are available, none have analyzed the most recent attention-based GANs and Transformer-based GANs. So, in this study, we review the recently introduced attention-based GANs and Transformer-based GANs, the literature related to which has not been reviewed by any survey paper.

Keywords: Variational Autoencoder; Generative Adversarial Networks; Autoencoder; Transformer; Self-Attention

1 Introduction

Generative models learn and capture the inner probabilistic distribution of training samples in order to produce new samples similar to it. In other words, they model simulated observations that are drawn from probability density functions. Such models include Gaussian Mixture Model, Naive Bayes Model, Hidden Markov Model and Restricted Boltzmann Machine(1). These models employ specialized functions to approximate real distributions, enabling them to make desired interpretations and perform outstanding accomplishments. However, each conventional generative model has a specific functional form, which is challenging to develop because of its complicated expression. As a result, the attention shifted away from these shallow models, toward deep generative models.

Deep generative models are multi-layer nonlinear neural networks that are used to simulate data dependency(2). Deep generative models have sparked a considerable interest because of their ability to provide a very efficient approach to evaluate and understand unlabeled data. They do not suffer from the capacity limitation and can learn to generate high-level representations solely from data. More crucially, when back-propagation is allowed, deep generative model training becomes extremely efficient, resulting in much superior performance than shallow models. The traditional popular deep generative models are from the Boltzmann family, which include Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs). The taxonomy of the deep generative models is shown in Figure 1. Deep generative prototypes also have a number of significant limitations such as they have high computational cost during inference process, have complicated structures, and also they may not be well generalized. Moreover, these models are not able to perform well when used with bigger datasets like ImageNet. All these limitations led to a limited attraction from the researchers towards them.

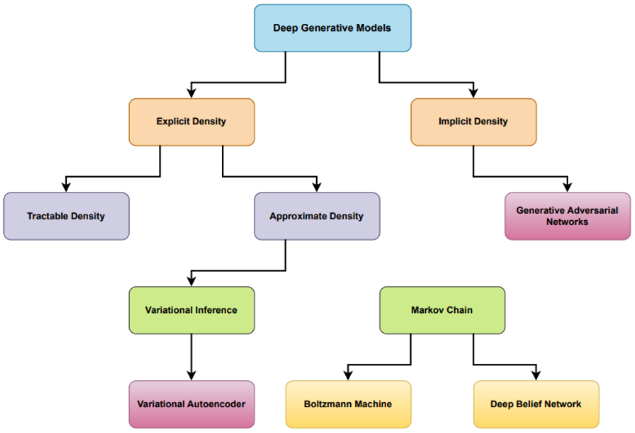

Fig 1. Taxonomy of Deep Generative Models

In recent years, two new deep generative models have evolved namely Variational Auto-encoders (VAEs)(3) and Generative Adversarial Networks (GANs)(4). Because of their prospective applications and success in the computer vision domain, GAN and VAE based models are gaining popularity. VAEs and GANs have become the potential unsupervised learning methods for complex distributions. Many survey studies on generative models are available; however they tend to focus on the specifics of their own approach rather than making comparisons with other approaches. In some of these works, the application domain takes precedence over theory. In this work, the authors have presented a review analysis of the two most prominent deep generative models from a theoretical point of view. Their pros and cons have also been discussed. Moreover, we have attempted to provide a hierarchical progress of DGM from VAE to the most recently introduced Transformer based GANs.

Recently, GANs have become a popular study topic and have been extensively investigated since 2014. With their many benefits, GANs are potent generative models that have grabbed the curiosity of the scientific community. Recently, researchers have incorporated different state of the art mechanisms in the GAN framework for better performance in generating realistic images with fine details. Recently, attention-based GANs were developed to handle long-range and global dependencies. Additionally, research has recently begun to look into the usage of transformers for generative tasks in the hopes of increasing the creation of complex images through increased expressivity. A need to compile recent studies about generative models has arisen as a result of this growing interest. None of the papers is available in literature that surveys the attention-based GANs and the Transformer based GANs. In this paper, we have reviewed these attention-based GANs and the GAN using Transformers. Besides, a comparison between VAE and GANs, challenges faced by GANs in general and transformer based GANs have also been discussed.

The paper is organized as follows:

Section 2 gives the following:

Subsection 2.1 first describes the Auto-encoder followed by the Variational Auto-encoder. It also gives the advantages and disadvantages of both Auto-encoder as well as Variational Auto-encoder.

Subsection 2.2 describes Generative Adversarial Networks along with their advantages and disadvantages.

Subsection 2.3 provides the overview of the Attention-based GANs.

Subsection 2.4 discusses the Transformer-based GANs.

Section 3 provides a comparison between VAE and GANs and the challenges faced by GAN and Transformer based GANs

Section 4 concludes the paper with a conclusion.

2 Methodology

2.1 Auto-encoder (AE)

Auto-encoders are among the simplest yet most elegant approaches to generative modeling. AE is an artificial neural network that tries to learn the compressed data so as to reconstruct the input as its output. It is an unsupervised feed forward non recurrent network that uses back-propagation to learn a sparse set of features for describing the training data. The dimensionality of the generated output vectors matches that of the input vector. Either end-to-end or layer-by-layer AE training is possible. Three components make up an AE which include:

a. Encoder: This component takes the input and encodes it into a compressed latent code representation

b. Code: It represents the compressed latent space representation that has been obtained by reducing the dimensionality of input and at the same time retaining as much as possible information in it.

c. Decoder: It takes the lower dimensional latent representation and reconstructs the original image from it with a minimum reconstruction error.

The primary objective of AE is to analyze the lower dimensional feature representations of the unlabeled data, and learn that, to generate the samples different from those that are present in the training set. Therefore, it must be restrained properly in order to prevent it from learning a trivial identity function which can result in generating simply the copy of the given input as the output. Auto-encoders are used for the compression tasks where the encoder latent space representation has low cardinality as compared to the input data with all the important features still intact.

2.1.1 Drawbacks of AE

1. Auto-encoders can only compress data that is identical to what they were trained on
2. When compared to the original inputs, the decompressed outputs will be deteriorated
3. It has no evident generative interpretation because it only does reconstruction

2.2 Variational Auto-encoder (VAE)

VAEs have become one of the common ways to learn difficult distributions without supervision. The Helmholtz Machine has influenced them. VAEs are deep generative latent variable models that model complicated data distributions by transforming simple distributions over a latent space. The VAEs' goal is to understand the training data distribution for synthesizing new data by sampling from it. The center of attraction in VAEs is that they are made on top of traditional function approximators and use stochastic gradient descent for their training. From generative modeling to semi-supervised learning to representation learning, the system offers a broad spectrum of applications. Moreover, due to the ease of their training technique, they have been employed for a broad range of downstream activities. They have been used for analyzing images, for generating images, for motion prediction domain, for Zero Shot Learning, and for generalized zero and few shot learning.

VAE is a type of auto-encoder that is based on Variational Bayes inference. It improves on a traditional AE by including a Bayesian module which learns the parameters that represent the data's probability distribution. The problem with auto-encoders is that, while the encoder has learned the input distribution and can encode it into a latent representation or codes, this latent representation is unknown. As a result, in the testing phase, the only choice for generating fresh samples is to employ random codes, which leads to unsatisfactory results. Variational Auto-encoders were created to address this problem. So, VAE is a Bayesian model that generates parameters that describe the probability distribution of the data by learning the compressed representation of AE. This method produces a latent space that is both continuous and structured. It then takes a sample from this distribution to create new output samples. The encoder, which is also referred to as recognition model and the decoder which is also referred to as generating model are two connected but independently parameterized models in the VAE. The architecture of VAE is given in Figure 2.

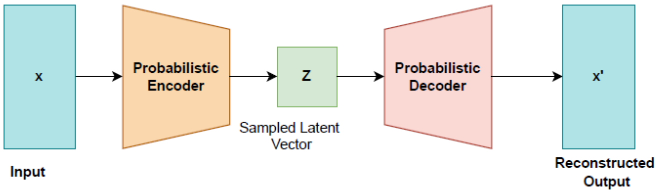

Fig 2. Architecture of Variational Auto-encoder

2.2.1 Advantages of VAE

1. By incorporating the variational lower bound, it eliminates difficult computation of marginal likelihood probability
2. It uses the reparameterization approach to circumvent the difficult Markov chain sampling procedure of latent variables
3. It regulates the latent representation vector distribution, thereby combining VAEs with representation learning to improve the downstream tasks
4. VAEs can produce meaningful data without any supervision due to their ability of learning the smooth latent representations of the input

2.2.2 Drawbacks of VAE

1. The images generated by VAEs are often unspecific, blurred and dispersed samples
2. VAE models are unable to simulate complicated, large-scale image datasets properly
3. During training, it is necessary to assess the relationship of different parameter in the network with the final output loss with the help of backpropagation. As a result, there is a need for some extra attention for the sampling process.

2.3 Generative Adversarial Network (GAN)

GAN is another emerging family of deep generative models. They are a powerful framework to learn models capable of generating natural images. Generative adversarial networks are based on a min-max game. The GAN architecture consists of two networks that train together. One is referred to as a generator and the other as the discriminator as shown in Figure 3. The generator's major goal is to produce data as close as possible to the potential distribution of the real data as feasible, whereas the discriminator's objective is to reliably identify which sample is from real data and which is from fake data. The likelihood or probability of a given sample being an actual sample is determined by the discriminator. A greater probability value suggests that the input belongs to the actual data. The sample is a false sample if the value is close to zero. The generation of an optimal solution is indicated by a probability value approaching 0.5, where the discriminator fails to understand whether the sample it received as input data is from the actual dataset or the generated dataset. At this point, we have a generator model that has learned the real-data distribution, and the optimal state is said to have been attained.

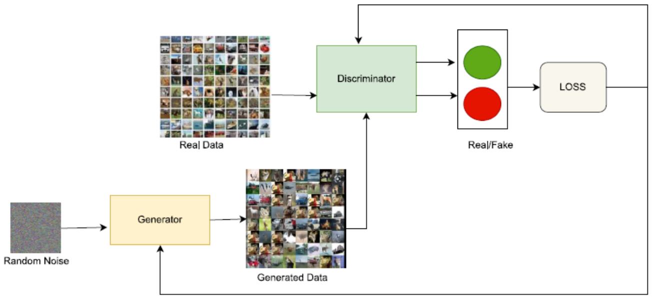

Fig 3. Architecture of Generative Adversarial Network

GANs are gaining a lot of attraction, and the demand to employ them in a variety of fields is expanding. GANs have been effectively utilized to resolve a range of problems[5]. Despite their success, GANs suffer from a number of challenges. Random initialization, architectural decisions, and hyper-parameter settings all affect how GANs solve the adversarial optimization issue. Moreover, GANs are infamous for being difficult to train and evaluate. It has a number of difficulties, training issues including non-convergence and mode collapse[6]. DCGAN, WGAN, WGAN-GP, and LSGAN are among the improved GAN versions that have been introduced to address these concerns.

2.3.1 Advantages of GANs

1. GAN's generator is a basic feed-forward network that generates data in batches rather than pixel by pixel like autoregressive models do. As a result, GAN can create samples in parallel, resulting in a significant increase in sampling speed, and this attribute allows GAN to be employed in a wider range of real-world applications.
2. GAN is a form of non-parametric production-based modeling method that does not require prior approximation training data distributions.
3. Unlike VAE, it is not necessary to approximate likelihood by inserting a lower bound. Instead, GAN is built to solve an adversarial game between its two neural networks, with a Nash equilibrium corresponding to determining the true data distribution.
4. By directly exploiting global information, GAN operates on the entire image and generates samples in less time.
5. GAN excels at capturing the image's high-frequency components. It has been demonstrated that it produces better and sharper results than other generative models, particularly VAE.

2.3.2 Disadvantages of GAN

1. GANs are more parameter efficient, but they have a hard time modeling discrete data
2. Because GAN is a novel generative model, it lacks adequate measures for comparing the performance of the models, their accuracy and the quality of the samples synthesized by them.
3. They have concerns with mode collapse and non-convergence during training

2.4 Attention-based GANs

The GANs for image generation generally use the convolutional layers in their architecture. The convolutional procedure processes the information in a local neighborhood and has poor scaling properties. All this makes the CNN-model difficult to capture long range dependency and to understand the global structure of objects. To deal with these long range and global dependencies, the models make use of the attention mechanisms. Attention mechanism captures important information from a large amount of information, and emphasizes on the most critical features in feature learning. It is now an essential component of convincing sequence modeling and transduction models in a variety of applications. It is being extensively employed in the areas of natural language processing and computer vision. Mejjati et al[7] used an attention mechanism with unsupervised image-to-image translation and were able to significantly enhance the image quality. Zhang et al.[8] introduced self-attention mechanism into convolutional GAN and proposed Self-Attention Generative Adversarial Networks (SAGAN). Self-attention, also known as intra-attention, relates several locations in a single sequence, thereby computing the response at a given location in a sequence. The generator of SAGAN has the ability to learn fine details and enhance the quality of the images generated. Yu et al[9] used Attention GANs for aerial scene classification. Tang et al[10] proposed Multi-Channel Attention Selection Generative Adversarial Network for image to image translation. Xiang et al[11] applied Attention mechanism to GANs for Semi-supervised Image Classification. Torrado et al.[12] used attention mechanisms in conjunction with conditional GANs in order to generate video games and proposed Conditional Embedding Self-Attention Generative Adversarial Network (CESAGAN). Dual Attention GAN (DAGAN) was proposed by Tang et al[13] to synthesize photorealistic and semantically consistent images. Qi et al[14] proposed Attention-Guided GANs for Data Augmentation on Medical Images. Jeha et al[15] used self-attention with progressive GANs and proposed Progressive Self-Attention GANs (PSA-GAN) in order to generate long time series samples with high quality. Schulze et al[16] employed multiple attention models and proposed Combined Attention Generative Adversarial Networks (CAGAN) for generating images from text.

2.5 Transformer Based GANs

Transformers have demonstrated outstanding performance in natural language processing and then in computer vision tasks. Transformer is the first transduction model that solely relies on self-attention to compute representations of its input and output. A number of studies have successfully used the architecture in the image domain. As a result of this success of applying transformers in the vision domain, led to further exploration of replacing the commonly-used CNN backbone with Transformers in GAN for image synthesis. In an effort to facilitate the creation of complex images, the academic community has begun to investigate the use of transformers for generative tasks. Jiang et al[17] succeeded in synthesizing images with 256 × 256 resolution utilizing solely pure transformer-based architectures and a GAN fully free of convolutions. Lee et al[18] presented ViTGAN, which uses Vision Transformers (ViTs) in GANs, and proposed key strategies for assuring training stability and enhancing convergence. STrans GAN, which employs Transformers in GAN, was proposed by Xu et al[19]. It employs STrans-G, a convolutional neural network (CNN)-free generator that provides competitive results in both unconditional and conditional image production. STrans-D, a Transformer-based discriminator, also dramatically narrows the gap with CNN-based discriminators. Zhao et al[20] developed HiT, a new Transformer-based decoder/generator in GANs for generating high-resolution images. HiT separates the generative process into low-resolution and high-resolution stages, respectively, with a focus on feature decoding and pixel-level generating. It does this using a hierarchical Transformer structure. The suggested technology scales easily to produce high definition images having resolution of 1024 × 1024. However, in the high resolution stages, it reduces to MLPs, thus lacking the ability to synthesize images of high fidelity that are comparable to the Convolution based counts parts. To generate images of high resolution that are comparable to leading convolution based architectures, Zhang et al[21] presented a transformer-based GAN that makes use of Swin transformers and is scalable to high resolutions. Instead of using local attention that struggles to maintain a balance between computing efficiency and modeling capability, authors used dual attention, which simultaneously uses the context of the local and the shifting windows, resulting in higher generation quality. Hudson et al[22] proposed GANsformer, which makes use of a bipartite structure to capture long range interactions across the image. Using bipartite attention instead of self-attention, it maintains computational efficiency. Park et al[23] proposed Transformer-GAN that uses style vectors for synthesizing images. The authors also addressed the computation problem of Transformers and showed superior performance on both high resolution images as well as low resolution images. Table 1 gives the timeline as well as the application area of different Attention GANs and Transformer GANs.

Table 1. Different Attention-based GANs and Transformer based GANs with their application area

Authors	Year	Model	Application
Mejjati et al[7]	2018	Attention-guided GAN	Image-to-Image Translation
Zhang et al.[8]	2019	SAGAN	Image Generation
Yu et al[9]	2019	Attention GANs	Aerial Scene Classification
Tang et al[10]	2020	Selection GAN	Image-to-Image Translation
Xiang et al[11]	2020	Attention-Based Semi-supervised Network	Image Classification
Torrado et al[12]	2020	CESAGAN	Video game level Generation.
Tang et al[13]	2020	DAGAN	Semantic Image Synthesis
Qi et al[14]	2020	SAG-GAN	Data Augmentation
Jeha et al[15]	2021	PSA-GAN	Synthetic Time Series
Schulze et al[16]	2021	CAGAN	Text to image generation
Jiang et al[17]	2021	Transgan	Image Generation
Lee et al[18]	2021	Vitgan	Image Generation
Xu et al[19]	2021	STrans GAN	Image Generation
Zhao et al[20]	2021	HiT-GAN	High Resolution Image Generation
Zhang et al[21]	2021	StyleSwin	High-resolution Image Generation
Hudson et al[22]	2021	GANsformer	Image Generation
Park et al[23]	2021	Styleformer	Image Generation

3 Discussion

VAE and GANs are specifically designed to approximate data distribution in visual datasets. Both of these deep generative models have attracted a lot of attention in recent times since they seek to understand the statistical distribution of real-world data and generate high-quality natural images. VAE is a probabilistic graphical model (PGM) using Bayesian inference as its foundation. Using PGMs, it effectively and efficiently deduces under uncertainty. The primary goal of VAE is latent modeling; hence it models the probability distribution of latent data in order to generate new samples. VAEs are taught to learn both an encoder and a decoder, which allows them to reconstruct data with amazing accuracy. The variational inference problem is solved by using generative neural network and recognition neural network for maximizing the minimized data likelihood. VAEs can be utilized for both supervised and unsupervised learning tasks simultaneously because of this shared framework in the same model. VAE is regarded as one of the best density models in terms of the likelihood criterion, in addition to providing a potential ground for the development of new models. However, VAE's variational technique creates a deterministic bias in optimizing the lower limit of log-likelihood instead of likelihood itself, resulting in blurry image synthesis. In contrast, GAN uses an adversarial training process to steadily enhance the quality of the generated data, resulting in realistic and colorful images that are difficult to differentiate from real photographs.

Although GANs have significant advantages over VAEs, they, like any other technology, confront a variety of difficulties. Mode collapse and training process instability are two issues that are commonly associated with the training process of GANs. In addition, the evaluation method for comparing the results generated by GANs is other challenging area of GANs. Furthermore, because GAN is often a CNN-based model, capturing long-range dependencies and understanding the global structure of an object is difficult. This led to the introduction of Attention-based GANs which cope with these long-range and global dependencies. Attention based GAN have been applied in different applications and have been able to produce better results than simple GANs. Moreover, recently, studies have started to investigate the use of transformers for generative tasks in the hopes of improving the development of complex images through enhanced expressivity. However, Transformer deployment in the generative adversarial network (GAN) architecture is still an open and difficult issue. The quadratic complexity of self-attention operation makes attention-based and Transformer based models challenging to be used with GANs for generating high-resolution images. Additionally, these blended GAN models are presently being mostly studied for image generating tasks; their full potential for use in other application areas is yet to be fully explored.

4 Conclusion

In nearly all applications, deep learning-based models produce outcomes that are state-of-the-art. It can be observed that the deep learning community considers generative models, particularly those that incorporate adversarial training, to be the most interesting area. VAEs and GANs have been the most popular and promising generative models. Both of these models are used to generate new data samples. Compared to VAEs, GANs have grabbed the scientific community's interest due to their amazing capacity to model complex real-world distributions. GANs have proven to be exceptionally effective at synthesizing a range of datasets, especially natural images. As a result, new techniques are being explored in combination with GANs to get better performance. This paper first describes the Auto-encoders, Variational Auto-encoders and GANs along with their advantages and disadvantages. The results from the GANs have been such that recent studies have come up with the embedding of attention mechanisms and Transformers in the GANs. The literature related to these Attention Based GAN and Transformer based GANs have been reviewed in this paper. As a consequence of this analysis, it can be concluded that there is a growing interest in generative models based on GAN due to their prospective applications and capacity for many application areas. The goal of this work is to give readers a better grasp of the present level of research in this topic, as well as to give researchers in related fields new insights and inspiration.

References

1) Alotaibi A. Deep Generative Adversarial Networks for Image-to-Image Translation: A Review. *Symmetry*. 2020;12(10):1705. Available from: https://doi.org/10.3390/sym12101705.
2) Ruthotto L, Haber E. An introduction to deep generative modeling. *GAMM-Mitteilungen*. 2021;44(2):202100008–202100008.
3) Kingma DP, Welling M. 2013. Available from: https://doi.org/10.48550/arXiv.1312.6114.
4) Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Communications of the ACM*. 2020;63(11):139–144. Available from: https://doi.org/10.1145/3422622.
5) Aggarwal A, Mittal M, Battineni G. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*. 2021;1(1):100004. Available from: https://doi.org/10.1016/j.jjimei.2020.100004.
6) Jabbar A, Li X, Omar B. A Survey on Generative Adversarial Networks: Variants, Applications, and Training. *ACM Computing Surveys*. 2022;54(8):1–49. Available from: https://doi.org/10.1145/3463475.
7) Mejjati A, Richardt Y, Tompkin C, Cosker J, Kim D, I K. Unsupervised attention-guided image-to-image translation. 2018. Available from: https://doi.org/10.48550/arXiv.1806.02311.
8) Jiang Y, S C, Z W. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. 2021. Available from: https://doi.org/10.48550/arXiv.2102.07074.

9) Yu Y, Li X, Liu F. Attention GANs: Unsupervised Deep Feature Learning for Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2020;58(1):519–531. Available from: https://doi.org/10.1109/TGRS.2019.2937830.

10) Tang H, Xu D, Yan Y, Corso JJ, Torr PH, Sebe N. Multi-channel attention selection gans for guided image-to-image translation. 2020. Available from: https://doi.org/10.48550/arXiv.2002.01048.

11) Xiang X, Yu Z, Lv N, Kong X, Saddik AE. Attention-Based Generative Adversarial Network for Semi-supervised Image Classification. *Neural Processing Letters*. 2020;51(2):1527–1540. Available from: https://doi.org/10.1007/s11063-019-10158-x.

12) Torrado RR, Khalifa A, Green MC, Justesen N, Risi S, Togelius J. Bootstrapping Conditional GANs for Video Game Level Generation. *2020 IEEE Conference on Games (CoG)*. 2020;p. 41–48. Available from: https://doi.org/10.1109/CoG47356.2020.9231576.

13) Tang H, Bai S, Sebe N. Dual Attention GANs for Semantic Image Synthesis. *Proceedings of the 28th ACM International Conference on Multimedia*. 2020;p. 1994–2002. Available from: https://doi.org/10.1145/3394171.3416270.

14) Qi C, Chen J, Xu G, Xu Z, Lukasiewicz T, Liu Y. Sag-gan: Semi-supervised attention-guided gans for data augmentation on medical images. 2020. Available from: https://doi.org/10.48550/arXiv.2011.07534.

15) Jeha P, Bohlke-Schneider M, Mercado P, Kapoor S, Nirwan RS, Flunkert V, et al. Progressive Self Attention GANs for Synthetic Time Series. *International Conference on Learning Representations*. 2021. Available from: https://doi.org/10.48550/arXiv.2108.00981.

16) Schulze H, Yaman D, Waibel A, Cagan. Text-To-Image Generation with Combined Attention Generative Adversarial Networks. In: DAGM German Conference on Pattern Recognition. Springer. 2021;p. 392–404. Available from: https://doi.org/10.1007/978-3-030-92659-5.

17) Jiang Y, Chang S, Wang Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*. 2021;34:14745–14758. Available from: https://proceedings.neurips.cc/paper/2021/hash/7c220a2091c26a7f5e9f1cfb099511e3-Abstract.html.

18) Lee K, Chang H, Jiang L, Zhang H, Tu Z, Liu C, et al. Vitgan: Training gans with vision transformers. 2009. Available from: https://doi.org/10.48550/arXiv.2107.04589.

19) Xu R, Xu X, Chen K, Zhou B, Loy CC. Stransgan: An empirical study on transformer in gans. 2021. Available from: https://doi.org/10.48550/arXiv.2110.13107.

20) Zhao L, Zhang Z, Chen T, Metaxas D, Zhang H. Improved transformer for high-resolution gans. *Advances in Neural Information Processing Systems*. 2021;34:18367–18380. Available from: https://proceedings.neurips.cc/paper/2021/hash/98dce83da57b0395e163467c9dae521b-Abstract.html.

21) Zhang B, Gu S, Zhang B, Bao J, Chen D, Wen F, et al. Styleswin: Transformer-based gan for high-resolution image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* . 2022;p. 11304–11314. Available from: https://doi.org/10.48550/arXiv.2112.10762.

22) Hudson DA, Zitnick L. Generative adversarial transformers. *International Conference on Machine Learning*. 2021;p. 4487–4499. Available from: https://doi.org/10.48550/arXiv.2103.01209.

23) Park J, Styleformer KY. Transformer based Generative Adversarial Networks with Style Vector. 2021. Available from: https://doi.org/10.48550/arXiv.2106.07023.