



A Survey of GAN Based Image Synthesis

Jiahe Ni*

Engineering Research Center of Digital Forensics of Ministry of Education, School of Computer, Nanjing University of Information Science & Technology, Nanjing, 210044, China

*Corresponding Author: Jiahe Ni. Email: a2234503831@163.com

Received: 01 February 2023; Accepted: 06 March 2023

Abstract: Image generation is a hot topic in the academic recently, and has been applied to AI drawing, which can bring Vivid AI paintings without labor costs. In image generation, we represent the image as a random vector, assuming that the images of the natural scene obey an unknown distribution, we hope to estimate its distribution through some observation samples. Especially, with the development of GAN (Generative Adversarial Network), The generator and discriminator improve the model capability through adversarial, the quality of the generated image is also increasing. The image quality generated by the existing GAN based image generation model is so well-paint that it can be passed for genuine one. Based on the brief introduction of the concept of GAN, this paper analyzes the main ideas of image synthesis, studies the representative SOTA GAN based Image synthesis method.

Keywords: Deep learning; image synthesis; SOTA; generative adversarial network

1 Introduction

Image synthesis technology has been studied in academia for a long time. With the development of GAN (Generative Adversarial Networks), the image quality generated by the image generation model is getting higher and higher, and has reached the same quality as the real image. And as the computing power of the industrial graphics cards are also improving year by year, the image generation models are also slowly applied to real life, such as digital people, virtual people, AI painting.

However, there are still many challenges in image generation technology. One of the challenges comes from GAN itself. In the training process of the GAN, the GAN is composed of two neural networks, a generator and a discriminator. The generator tries to generate the fake samples to decepit the discriminator, and the discriminator tries to distinguish the fake samples from the generated samples. Under this, the performance of generator and discriminator is continuously improved. But in the actual training process, it is often difficult to achieve this balance because the discriminator is easier to train, so that the pattern collapses. The second challenge comes from the generation of high-resolution images. When the resolution of the generated image reaches 512×512 or 1024×1024 , the quality of the generated image will be reduced. It is difficult to keep the details of the image lively.



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Progressive training can alleviate this phenomenon, but it will bring artifacts in the training process. In addition, if the image used for training contains too much information and content, the quality of the generated image will be reduced. For example, the images used for training are from cartoon scenes, which contain a plurality of characters or more complex spatial layout.

The demand for digital virtual human and AI painting in real life has promoted the development of image generation. Researchers have proposed various schemes [1,2] to optimize and solve the above problems and generate higher quality pictures. A more stable GAN loss function [3-5] is used to avoid problems during GAN training. The quality of generated images in high-resolution image generation task can be improved through style-based generation, mapping latent variables to more smooth latent space and other methods. At present, the style-based generation model has achieved impressive results in generating image quality. However, these generation models may still have some defects, which makes them perform poorly in training large unstructured datasets.

In this paper, first, briefly introduce the GAN generation model and the loss function researches that makes the GAN training process stable. Then this paper will systematically introduce some methods of image generation quality reaching SOTA models in recent years and discusses why it can generate images with higher and higher quality.

2 GAN and its Loss Function

In this section, we will briefly discuss the relevant background of GAN and the optimization of its loss function to make the training of GAN more stable.

2.1 GAN

GAN [6] is mainly composed of two networks, the generation network G and the discriminant network D. The idea of the generation model G is to package a noise into a realistic sample, and the discriminant model D needs to judge whether the incoming sample is real or false. The two parts promote each other. With the continuous improvement of the discriminant ability of the discriminant model D to the sample, the forgery ability of the generation model G also increases.

The loss function of GAN is shown in Eq. (1):

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

G generates similar images according to real images, and D distinguishes whether the input image is a real image, or a false image generated by G. G and D are mutually reinforcing. The purpose of G is to make D feel blurred by the images generated, and D does not know whether they should be divided into realistic or fake images. The purpose of D is to accurately distinguish between realistic and fake images. Therefore, the images generated by G will become more and more true, and the discrimination ability of D will become stronger and stronger, finally reaching a balance. This Nash equilibrium ensures that all images generated by GAN can belong to the same category as the original image.

2.2 DCGAN

DCGAN [7] is a better improvement after GAN. Its main improvement is mainly in the network structure, which greatly improves the stability of GAN. The network structure based on DCGAN is widely used in related research.

Fig. 1 shows the generator structure of DCGAN. The generator receives a random noise z , and then generates an image $G(z)$ by up sampling. The up sampling mainly adopts the deconvolution algorithm. G receives a 100 dimensions random noise z , which is converted into a $4 \times 4 \times 1024$ feature map through Project and reshape, and then generates an image with a size of $64 \times 64 \times 3$ through multiple deconvolution layers.

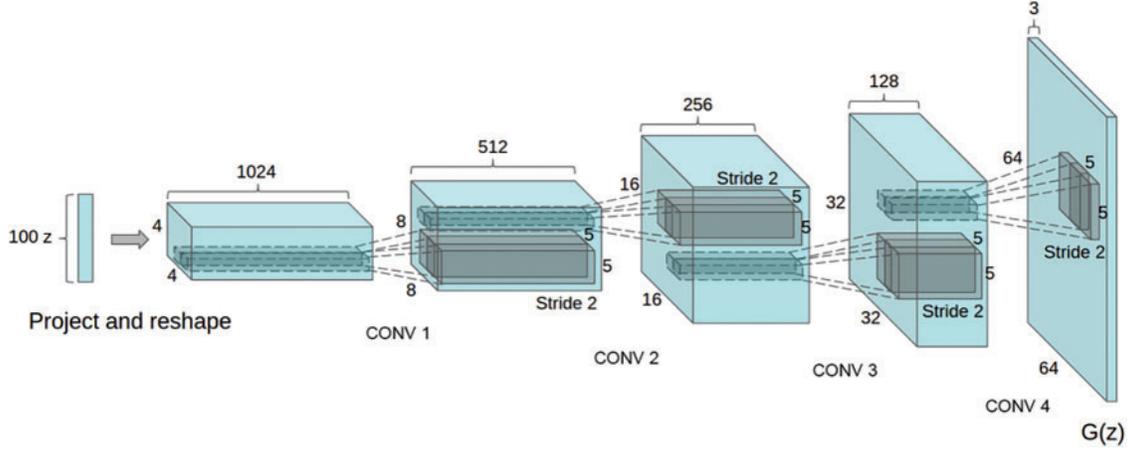


Figure 1: The generator structure of DCGAN

2.3 WGAN

Since the advent of GAN, there have been problems such as training difficulties, the loss of generators and discriminators cannot indicate the training process, and the lack of diversity of generated samples. Since then, many papers have been trying to solve the problem, but the effect is not satisfactory. WGAN [8] solved the following defects of GAN. WGAN completely solves the problem of unstable training of GAN, and it is no longer necessary to be careful to balance the training level of generator and discriminator. WGAN basically solves the collapse mode problem of mode collapse and ensures the diversity of generated samples.

WGAN redefines the loss function of GAN using Wasserstein distance [9], while the loss of the original GAN is based on JS divergence. The advantage of Wasserstein distance over JS divergence is that even if the two distributions do not overlap, Wasserstein distance can still reflect their distance. JS divergence is abrupt, either maximum or minimum, and Wasserstein distance is smooth. If the gradient descent method is used to optimize parameters, the former two cannot provide gradient at all, but Wasserstein distance can. Similarly, in high-dimensional space, if two distributions do not overlap or the overlapping part can be ignored, JS divergence can neither reflect the distance nor provide gradient, but Wasserstein can provide meaningful gradient.

The gradient of the weight which model using WGAN loss function is truncated with a constant value K to meet the required Laplace continuity condition.

The loss function of WGAN are shown as Eqs. (2) and (3).

$$L_G = -\mathbb{E}_{x \sim P_g} [f_w(x)] \quad (2)$$

$$L_D = \mathbb{E}_{x \sim P_g} [f_w(x)] - \mathbb{E}_{x \sim P_r} [f_w(x)] \quad (3)$$

JS divergence can neither reflect distance nor provide gradient, but Wasserstein can provide meaningful gradient. Different from DCGAN, WGAN mainly improves GAN from the perspective of loss function, and the improved WGAN can achieve good performance even on the full link layer.

2.4 WGAN-GP

WGAN-GP [10] is an improved version after WGAN, mainly improving the conditions of continuity limitation.

WGAN truncates the model weights, so that most of the weights are concentrated on the value of $-K$ and K , which makes the deep neural network unable to fully its strong fitting ability. Moreover, the forced clipping weight is easy to cause the gradient to disappear or explode. The reason for gradient disappearance and gradient explosion lies in the selection of shear range. If the selection is too small, the gradient will disappear. If it is set a little larger, the gradient will become a little larger every time it passes through a layer of network, and gradient explosion will occur after a layer of network. To solve this problem and find an appropriate way to meet lipschitz continuity condition, WGAN-GP uses gradient penalty to meet this continuity condition.

The loss function of WGAN-GP are shown as Eq. (4)

$$L = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} \left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1 \right)^2 \right] \quad (4)$$

3 GAN Based Image Synthesis Methods

Image synthesis refers to AI learning to generate images that do not exist in the real world. GAN based methods can generate images with better quality. Specifically, the input of GAN is random noise. After a series of deconvolution and up-sampling operations, the random noises are mapped to the corresponding image in the image domain. Image generation can be used for virtual human, digital human [11], etc. to protect privacy. And it can be used in the AI painting field to reduce the cost of painters in the game and multimedia industries. With the development of GAN, gradual generation and style based generation, image generation technology ushers in new development opportunities, which are easier to realize and have better quality.

According to the proposed order, this paper will introduce the image generation methods that have reached SOTA in recent years.

3.1 PGGAN

If directly generate a large resolution image as 1024×1024 , and establish a mapping network G from the latent code, it must be difficult to work. Because during the generation process, discriminator D can easily identify the fake image generated by G , and G is difficult to train. Therefore, PGGAN [12] proposed a progressive growing for training.

The training starts with a generator and discriminator with a low spatial resolution of 4×4 pixels. With the improvement of training, layers are gradually added to the generator and discriminator networks, thus increasing the spatial resolution of the generated pictures. All existing layers remain trainable through processes. This method makes the synthesis stable in high resolution and accelerates the training speed.

In addition, GAN tends to capture the diversity of only one subset of the training set. PGGAN uses minibatch discrimination to solve this problem. They calculate feature statistics in the whole small batch, to encourage the generated and trained minibatch images to have similar statistics. To solve this

problem, PGGAN add a minibatch layer at the back of the discriminator to learn a large tensor, which projects the input into a set of statistical information.

3.2 StyleGAN

StyleGAN [13] first focused on PGGAN's generator network. It found that a potential benefit of progressive layers is that if used properly, they can control different visual features of images. The lower the layer and resolution, the rougher the features it affects. StyleGAN adds many additional modules based on PGGAN generator.

Fig. 2 shows the structure of styleGAN first improvement of StyleGAN is to add a Mapping Network consisting of eight full connection layers to the input of the Generator, and the output of the Mapping Network is the same size as the input layer. The goal of adding Mapping Network is to encode the input vector into an intermediate vector, and the intermediate vector will be subsequently transferred to the generation network to get 18 control vectors, so that different elements of the control vector can control different visual features. The second improvement of StyleGAN is to transform the intermediate vector after feature unwrapping into a style control vector, to participate in the generation process of the influence generator. The third improvement of StyleGAN is to delete the 4×4 input layer and replace it with a constant value. This can reduce the probability of generating some abnormal pictures due to improper initial input values. Another advantage is that it can help reduce feature entanglement, and it is easier for the network to learn when only latent variables are used and do not depend on the entanglement input vector. And StyleGAN uses AdaIN [14–17] to convert images to any style.

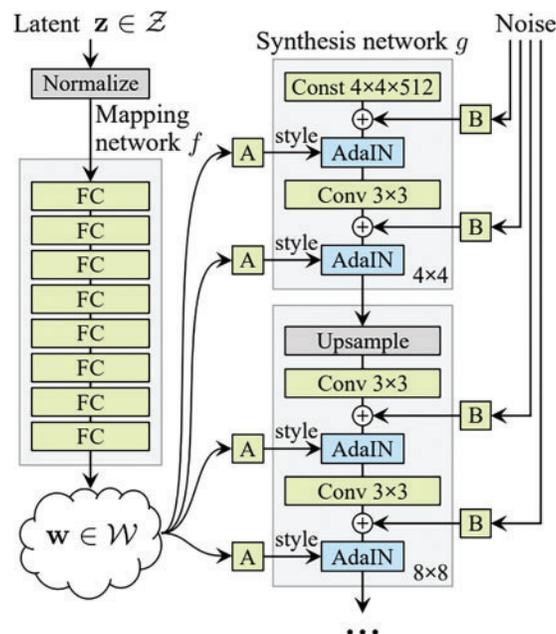


Figure 2: The structure of StyleGAN

3.3 StyleGAN v2

Although StyleGAN v1 can produce pictures of excellent quality, sometimes the resulting pictures contain water droplets and artifacts. StyleGAN v2 [18] is improved based on v1. It focuses on the

artifact problem and can generate better image data. It has improved v1 in terms of style mixing strategy, progressive growth generation method, interpolation method, etc.

Fig. 3 shows the artifacts in the image generated by StyleGAN v1. In v1, AdaIN is used to control the influence of the source vector on the generation of images. Similar to the BN layer, it aims to scale and shift the output results of the middle layer of the network to increase the learning effect of the network and avoid the disappearance of the gradient. The appearance of AdaIN has freed the style migration task from being limited to one style or requiring the length optimization process. The style and content can be combined only through normalization statistics. The generator overcomes the problem of information loss by snaking information in these layers, but this brings about the problem of watermark artifacts, which the discriminator cannot distinguish. To solve the problem of artifacts, v2 reconstructs AdaIN into Weight Demodulation, and the processing flow is shown as Eqs. (5) and (6):

$$w'_{ijk} = s_i \cdot w_{ijk} \quad (5)$$

$$w''_{ijk} = w'_{ijk} / \sqrt{\sum_{i,k} w'_{ijk} \cdot w'_{ijk} + \epsilon} \quad (6)$$

where s_i are the i th latent code, w_{ijk} are the original weight of convolution kernel, w'_{ijk} are the weight scaled by the latent code, w''_{ijk} are the weight after normalized.



Figure 3: Artifacts in the image generated by StyleGAN

In addition, unlike the strategy of the StyleGAN v1 using progressive growth, v2 began to seek other designs to make the network deeper and training more stable. For the Resnet structure, network deepening is realized through skip connection. So StyleGAN2 adopts a residual block like ResNet [19]. Use bilinear filtering to sample the previous layer up-sample and down-sample and try to learn the residual value of the next layer. Inspired by MSG-GAN [20], StyleGAN2 designed a new architecture to make use of multiple scale information generated by images. They use jump connection to map low resolution features to the final generated image.

3.4 StyleGAN v3

In the process of GAN synthesis, some features depend on absolute pixel coordinates, which will cause the details to appear stuck to the image coordinates rather than the surface of the object which generated. StyleGAN v3 fundamentally solves the problem of adhesion between image coordinates and features of StyleGAN v2. StyleGAN v3 realizes the rotate-invariant and image shift-invariant [21,22], and greatly improves the quality of image synthesis.

Fig. 4 shows the “texture sticking” in StyleGAN v2 and StyleGAN v3 has solved this problem. The first part of StyleGAN v3 improvement is mainly aimed at the overall network structure. StyleGAN v3 replaces constant input with Fourier features [23,24], because Fourier features can easily impose translation and rotation operations. At the same time, StyleGAN v3 deleted the operation of adding

tiny noises layer by layer in StyleGAN. In the complete experiment, it was found that this operation would seriously damage the equivariability.

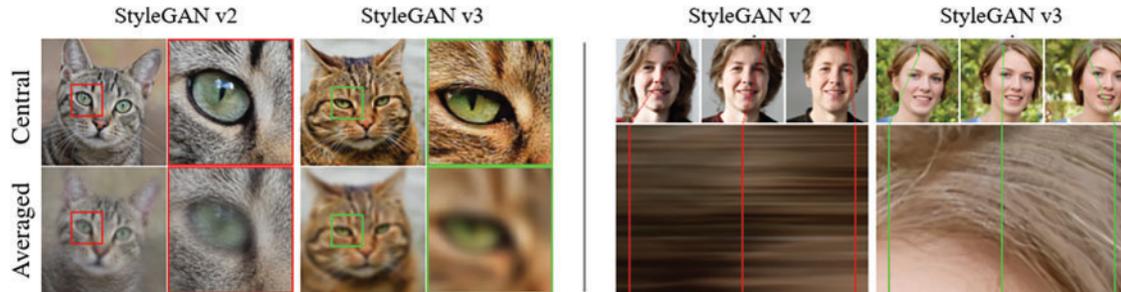


Figure 4: “Texture sticking” in StyleGAN v2 and StyleGAN v3 has solved this problem

In addition, StyleGAN v3 believes that, for the GAN model, the high-frequency information in the shallow layer may be meaningless and can be eliminated to further improve the anti aliasing ability. Therefore, in the actual training, we can consider designing a lower cut-off frequency for other network structures except the network layer that executes the highest resolution generation to improve the anti aliasing ability. At the same time, if you simply apply the same cut-off frequency to each network layer, the effect may not be very good, because for a shallower layer, its frequency component will be more concentrated near the cut-off frequency. Therefore, use an adjustable cut-off frequency strategy to gradually increase the cut-off frequency as the layer deepens. This operation not only improves the translation invariance, but also improves the image quality.

To achieve rotational invariance, the cores of convolution and down-sampling operations need to be radial symmetric. For convolution, the convolution kernel of 3×3 is replaced by 1×1 , but this will reduce the network capacity. Therefore, the number of features output by each convolution layer is doubled while changing the convolution core. For the down sampling operation, the first order Bessel function of the first kind with radial symmetry is used as the filter to meet the down-sampling conditions. Similarly, the improvement of the rotation invariance has not been applied to the network layer generated with the highest resolution.

3.5 StyleGAN-xl

StyleGAN-xl [25] is the first model to demonstrate 1024×1024 resolution image synthesis on ImageNet scale. Experiments show that even the latest StyleGAN3 can not be well extended to ImageNet. The training will become unstable at high resolution. StyleGAN sets a benchmark method for modeling the generation of image quality and controllability. However, StyleGAN does not perform well on large unstructured datasets such as ImageNet. It is more suitable for generating face data.

The researchers first modified the StyleGAN generator and its regularization loss, and adjusted the potential space to adapt to the Projected GAN and class condition settings. Then it discusses the gradual growth again to improve the training speed and performance. Next, the characteristic network used for the training of Projected GAN is studied to find a very suitable configuration. Finally, classifier guidance is proposed so that GAN can provide category information through a pre trained classifier. In this way, it is possible to train a much larger model than before, while requiring less computation than the existing technology. StyleGAN-xl is three times larger than the standard StyleGAN v3 in depth and parameter count.

Fig. 5 shows the training procedure of StyleGAN-xl. StyleGAN-xl sends the potential code z and category label c to the pretrained embedding and mapping network G to generate style code w , which is used to modulate the convolution of synthesis network G . In the training process, gradually increase the number of layers to double the output resolution of each stage of incremental growth. Only the latest layer is trained, while keeping the other layers fixed. G only trained in the initial 16×16 stages, and remained fixed in the higher resolution stage. The combined image is magnified when it is less than 224×224 , and it passes through a CNN, a (ViT) Vision Transformer [26], and their respective feature blending blocks (CCM + CSM). At higher resolutions, CNN receives unmodified images, while ViT receives reduced frequency input to maintain low memory requirements, but still uses its global feedback. Finally, eight independent discriminators are applied to the obtained multi-scale feature map. The image is also sent to the classifier CLF for classifier guidance.

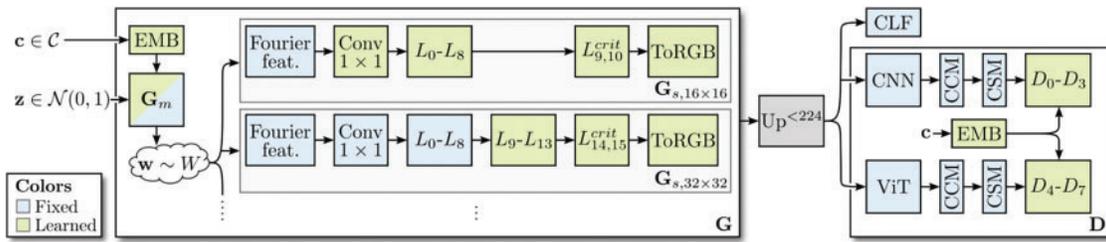


Figure 5: Training procedure of StyleGAN-xl

4 Challenges and Prospects

At present, although the image quality generated by image synthesis models is already high, it has not been widely used in the creation of painting scenes such as games, movies and so on. This is a long-term and continuous process. The existing image synthesis technology still has some limitations.

The effect of processing complex target datasets is poor. When a dataset image has an unfixed number of targets in an image, the quality of the generated image will be reduced. Existing generation models tend to focus on controlling the generation of a single target in the generation process. When there are multiple targets, the latent variables in the generation process are difficult to record the relationship between different targets. Improving the quality of the generated image on the complex target dataset is still the focus and difficulty of the research.

The effect is poor when generating images with complex details. This phenomenon mainly occurs when there are too many details in the training pictures. For example, the dataset is a full body portrait of a character, including the fingers, toes, and other details of the character. The current image synthesis models have obvious shortcomings in generating small area details on large resolution images.

Better customization. Using the image synthesis model to obtain the desired image requires that the generation process of the model can be customized. The use of latent variables in style-based synthesis provides customization to some extent. However, the actual image structure layout corresponding to the latent variable needs certain algorithms to map similar images back to the latent space, which makes it not intuitive to customize images using the latent variable.

In the future, the image generation technology still needs to be constantly improved and polished to meet various needs in practical applications. In view of the above limitations, the research on image

generation should, based on existing work, mine more methods that adapt to complex datasets and more customizable in line with human understanding.

5 Conclusion

This paper combs and summarizes the existing image synthesis technology based on GAN, especially the image synthesis technology on style-based. The goal, purpose, method, and limitation of existing image synthesis schemes are systematically summarized. To solve the problem of poor quality of high-resolution image generation, the existing image synthesis methods adopt the method of style modulation and progressive generation to improve the quality. To solve the problem of poor performance in training large unstructured data sets, existing image synthesis methods use CNN and ViT jointly to improve the image quality on such data sets.

In the future, image generation technology will continue to examine the defects of existing methods and develop a generation model that can generate higher quality images, so that this technology can gradually be used in real painting. Image generation technology has been developing and improving in the research.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford *et al.*, “Improved techniques for training gans,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 2234–2242, 2016.
- [2] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” arXiv preprint arXiv:1701.04862, 2017.
- [3] J. Zhao, M. Mathieu and Y. LeCun, “Energy-based generative adversarial network,” arXiv preprint arXiv:1609.03126, 2016.
- [4] D. Berthelot, T. Schumm and L. Metz, “Began: Boundary equilibrium generative adversarial networks,” arXiv preprint arXiv:1703.10717, 2017.
- [5] G. J. Qi, “Loss-sensitive generative adversarial networks on lipschitz densities,” *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1118–1140, 2020.
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [7] A. Radford, L. Metz and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.
- [8] M. Arjovsky, S. Chintala and L. Bottou, “Wasserstein generative adversarial networks,” in *Int. Conf. on Machine Learning*, pp. 4401–4410, 2019.
- [9] G. Montavon, K. R. Müller and M. Cuturi, “Wasserstein training of restricted boltzmann machines,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 3718–3726, 2016.
- [10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, “Improved training of wasserstein gans,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 5767–5777, 2017.
- [11] P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsonan, D. Novy *et al.*, “AI-generated characters for supporting personalized learning and well-being,” *Nat. Mach. Intell.*, vol. 3, no. 12, pp. 1013–1022, 2021.
- [12] T. Karras, T. Aila, S. Laine and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017.

- [13] T. Karras, S. Laine and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 4401–4410, 2019.
- [14] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, VEC, ITA, pp. 1501–1510, 2017.
- [15] V. Dumoulin, J. Shlens and M. Kudlur, “A learned representation for artistic style,” arXiv preprint arXiv:1610.07629, 2016.
- [16] G. Ghiasi, H. Lee, M. Kudlur, D. Manjunath, S. Vincent *et al.*, “Exploring the structure of a real-time, arbitrary neural artistic stylization network,” arXiv preprint arXiv:1705.06830, 2017.
- [17] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. Vries *et al.*, “Feature-wise transformations,” *Distill*, vol. 3, no. 7, pp. e11, 2018.
- [18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen *et al.*, “Analyzing and improving the image quality of stylegan,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8110–8119, 2020.
- [19] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, LV, USA, pp. 770–778, 2016.
- [20] A. Karnewar and O. Wang, “Msg-gan: Multi-scale gradients for generative adversarial networks,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 7799–7808, 2020.
- [21] A. Chaman and I. Dokmanic, “Truly shift-invariant convolutional neural networks,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Online, pp. 3773–3783, 2021.
- [22] R. Zhang, “Making convolutional networks shift-invariant again,” in *Int. Conf. on Machine Learning. PMLR*, California, CA, USA, pp. 7324–7334, 2019.
- [23] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan *et al.*, “Fourier features let networks learn high frequency functions in low dimensional domains,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.
- [24] R. Xu, X. Wang, K. Chen, B. Zhou, C. C. Loy *et al.*, “Positional encoding as spatial inductive bias in gans,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Online, pp. 13569–13578, 2021.
- [25] A. Sauer, K. Schwarz and A. Geiger, “Stylegan-xl: Scaling stylegan to large diverse datasets,” in *ACM SIGGRAPH 2022 Conf. Proc.*, Vancouver, VAN, CA, pp. 1–10, 2022.
- [26] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang and A. Dosovitskiy, “Do vision transformers see like convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.