Anime Style Image Generation Based on StyleGAN3

Mingjia Deng[1*], *Mengran* Zhang[2]

[1]college Of Computer Science, Chongqing College of Mobile Communication, Chongqing, 401420, China
[2]College of Software, Henan Normal University, Xinxiang City, Henan, 453007, China

Abstract. With the rise of anime culture and the development of computer vision technology, automatically generating images with a specific comic style has become a research hotspot. The animation industry is in urgent need of high-quality and diverse comic-style images, but traditional hand-drawing methods are inefficient. This paper chooses to use the existing Style-Based Generative Adversarial Network（StyleGAN3） model because of its excellent image generation capabilities and high-resolution output. Compared with the earlier StyleGAN2, StyleGAN3 eliminates aliasing in generated images. StyleGAN3 can achieve more natural refinement and generate more realistic and stylized images. This paper adjusts the latent space vector of StyleGAN3 to achieve precise control of anime style features, and to achieve the purpose of anime-style image generation based on StyleGAN3. The output results show that the generated images not only accurately retain the key style elements of the animation, but also show a high degree of diversity and authenticity, which fully verifies the effectiveness and feasibility of the method proposed in this paper. This research result provides strong technical support for animation creation, game design, and other fields, greatly enriches image resources, and significantly improves creation efficiency. It has important application value and practical significance.

1. Introduction

The automatic generation of anime-style images has become a research hotspot in the field of computer vision. As the number of Internet users surges to 5.44 billion, animation culture is becoming more popular around the world, and digital technology is also developing rapidly [1, 2]. This technology has shown great application potential in the fields of animation creation, game design, and virtual image generation, which can significantly reduce the cost of creation and improve the efficiency and diversity of creation [3]. However, most of the anime-style image generation methods currently on the market rely on traditional hand-drawing or rule-based algorithms, which have obvious limitations in efficiency, flexibility, and realism [4].

* Corresponding author: jackdeng7@ldy.edu.rs

Looking back at previous research, many scholars have tried to use Generative Adversarial Networks (GANs) to generate image styles. For example, Karras et al. proposed introducing a style-based generator architecture in GANs, which achieved unsupervised separation of high-level attributes and random changes in generated images and significantly improved the quality of generated images [5]. However, Karras et al. also pointed out in a subsequent research article that StyleGAN has difficulty in achieving a good balance between detail fidelity and style consistency when processing images of a specific style (such as anime style) [6]. Jahanian et al. made significant progress in studying the operability of GANs [7]. They experimentally demonstrated the possibility of creating realities while changing the distribution picture through "steering" in the latent space. However, GAN models still reflect the biases of the datasets they were trained on. This means that although GANs can adapt well to standard datasets, they may not generalize well to real-world data with different characteristics or distributions. For example, they are not good at accurately expressing anime styles [7].

Specifically, this paper elaborates on the anime-style image generation method based on StyleGAN3. It includes key steps such as model training, style feature extraction, and fusion. This paper aims to expand the application scenarios of GANs in the field of image style transfer and provide new ideas and solutions for the automatic generation of anime-style images.

2. Method

2.1 Dataset

The Flickr-Faces-HQ (FFHQ) dataset is a high-quality face image dataset created for benchmarking GANs [8]. The dataset contains 70,000 high-quality PNG images with a resolution of 1024×1024, covering facial images of different ages, races, and diverse backgrounds. FFHQ not only has significant diversity in the basic features of face images but also has good coverage of accessories (such as glasses, sunglasses, hats, etc.), increasing the realism and complexity of the data.

The images were processed using dlib for automatic alignment and cropping, and only images with permissive licenses were included. A variety of automatic filters were applied during the dataset construction to optimize the image quality. Finally, Amazon Mechanical Turk was used to manually check to remove images of occasional statues, paintings, or photos.

2.2 Model selection

This article selects the StyleGAN3 pre-trained model officially released by NVIDIA for configuration T (translation equivalent value) and configuration R (translation and rotation equivalent value) [9]. The StyleGAN3 pre-trained model is an image generation model built based on GAN, which inherits and improves its predecessor models StyleGAN and StyleGAN2.

2.3 StyleGAN3 generator

StyleGAN3 is an improvement on StyleGAN2, mainly targeting the "stickiness" phenomenon that exists in StyleGAN2 when translating or rotating. For animal images, normally the features on the face should move together. However, in StyleGAN2, the beard fails to fit tightly against the skin and moves naturally with facial movements. This situation indicates that the model may rely on the absolute coordinate system during feature generation,

and does not move well with surrounding features. Aittala et al. believe that the fundamental problem of this phenomenon is that the current generator network architecture is a convolution + nonlinear + upsampling structure, and such an architecture cannot achieve good Equivariance. Specifically, convolutional layers and upsampling layers may lose some spatial information when processing images, especially in shallow networks.The output features (coarse features) of the shallow network mainly control the presence or absence of the output features (fine features) of the deep network, but do not precisely control their positions. This means that even if the shallow features are translated or rotated, the positions of the deep features may not be adjusted accordingly, resulting in spatial inconsistency in the generated image [5].

The generator architecture of StyleGAN3 inherits the basic structure of StyleGAN2, but has made many improvements. The main improvements include the following aspects:

Fourier Features: StyleGAN3 introduces Fourier features to enhance the generator's ability to process input latent vectors. Fourier features enable the generator to better capture the translation and rotation invariance of images by mapping the latent vectors to the frequency domain.

Fourier feature mapping: The latent vector is transformed through a Fourier transform to generate frequency domain features. Frequency domain operation: Linear transformation and nonlinear operation are performed in the frequency domain to enhance the generator's control over image details.

Translation and rotation invariance: StyleGAN3 introduces translation and rotation invariance in the generator, making the generated images more continuous and smooth when translated and rotated. This is achieved by using translation and rotation invariant convolution operations in various layers of the generator. Rotation Equivariant Convolutions: The convolution layers in the generator are designed to be rotation equivariant, which means that when the input image is rotated, the output image will also rotate accordingly without distortion. Rotation-Invariant Features: By introducing rotation-invariant feature representations, the generator can better handle rotation-invariant image generation tasks.

Nonlinearity and filtering: According to the previous analysis, filtering is required after nonlinear activation, and Karras uses the order of upsampling-nonlinear activation-downsampling for processing. After analysis, it is believed that the upsampling operation after activation in StyleGAN2 and the upsampling operation of nonlinear activation in StyleGAN3 can be integrated for processing, that is, the order of upsampling is changed, and a customized CUDA kernel is used to complete the entire operation. This can save memory and speed up training by 10 times. The model processed in this way can effectively improve translation equivariance.

3. Experiment

3.1 Model training

First, configure the environment, and create and activate the StyleGAN3 Python environment. Download the Flickr-Faces-HQ dataset as 1024x1024 images and create a zip archive. Use the pre-trained network model, reference different model weights through local file names or URLs, and modify the training parameters and weight paths to generate different styles of face photos. Automatically calculate the Fréchet Inception Distance (FID) of each network pickle exported during training. Monitor the training progress through regular checks (or TensorBoard). After obtaining a certain amount of data, this article compares the effect graphs generated by StyleGAN2 and StyleGAN3.

During the model training process of this experiment, this article is based on the StyleGAN3 pre-trained model released by NVIDIA, and two different configurations are selected: T (translation-only equivalent) and R (translation and rotation equivalent).

3.2 Experimental setup steps

The baseline model architecture of this article is StyleGAN3 based on GAN, which consists of a Generator and Discriminator. The generator is responsible for generating realistic images, while the discriminator is used to distinguish between real and fake images. The two enhance each other through adversarial training to optimize the quality of generated images. StyleGAN3 follows the modular design of StyleGAN2 and adds improvements to combat aliasing. These improvements include hierarchical representation, style control, and pixel-by-pixel discriminator, which make the model more advantageous in terms of clarity and detail expression of generated images.

3.3 Performance evaluation

This article uses FID as the main evaluation indicator of image generation quality. FID is a commonly used metric in GAN, which is used to evaluate the similarity between the distribution of generated images and the distribution of real images. The lower the FID value, the smaller the gap between the generated image and the real image in visual and statistical characteristics, that is, the higher the generation quality. During the training process, the FID value of each network snapshot is automatically calculated by default, so that the improvement of generation quality can be tracked in real-time. In addition, it is recommended to check the changes in the FID value regularly or use TensorBoard to visualize the progress of training. By observing the fluctuations in the FID value, you can judge the stability and convergence of the training, to better adjust the training parameters.

$$d^2 = \left|\left|mu_1 - mu_2\right|\right|^2 + Tr(C_1 + C_2 - 2 * sqrt(C_1 * C_2)) \quad (1)$$

In the given context, mu_1 and mu_2 represent the mean vectors of the real data and the generated model, respectively, and C1 and C2 represent their covariance matrices. The trace of the matrix is indicated by "Tr", and the 2-norm of the matrix is indicated by "||*||".

The StyleGAN2 generator consists of a mapping network that converts latent code into an intermediate space w and a composite network G that starts with the constant Z_0 and applies convolution, nonlinearity, upsampling, and noise layers to produce the output image Zn. The latent code w modulates the convolutional kernel in G. StyleGAN2 also uses skip connection and regularization techniques.

The goal is to equate each layer of g to a continuous signal transformation so that finer details are transformed with coarser features. To measure the degree of equal dispersion, the peak signal in decibels (dB) vs. noise to compare the images obtained by converting the inputs and outputs of the network.

$$EQ - T = 10 \cdot \log_{10}\left(\frac{I_{max}^2}{E_w} \sim w, x \sim X^2, p \sim V, c \sim C\left[\left(g\left(t_{x[z_0]}; w\right)_{c(p)} - t_{x[g(z_0;w)]_c}(p)\right)^2\right]\right) \quad (2)$$

Each image set corresponds to a unique random selection of latent vectors w sampled at a particular integer pixel position p in the overlapping effective region V. The spatial transformation operator tx applies a 2D shift using the integer offset x sampled from the

distribution X2. In addition, a similar metric EQR is defined for rotation, and the rotation angle is randomly selected from a uniform distribution between 0° and 360°.

4 Result

As shown in Figures 1, 2, a female anime portrait and a male anime portrait were obtained by training the FFHQ dataset.

Fig. 1. Female style anime portrait (Photo/Picture credit: Original)

Fig. 2. Male-style anime portrait (Photo/Picture credit: Original)

Algorithm comparison: The results of StyleGAN2, StyleGAN3-T, and StyleGAN3-R are given in Table 1. The results show that StyleGAN3 proposed by the author Tero Karras et al. is still very competitive compared to StyleGAN2 under the FID evaluation standard, while the flexible layer specification StyleGAN3-T and the rotation-isomorphic StyleGAN3-R perform similarly on FID, both showing good translation invariance [5, 10].

Table 1. Data results of StyleGAN2, StyleGAN3-T and StyleGAN3-R are displayed

Generator	FID↓	EQ-T	EQ-R
StyleGAN2	5.14	-	-
StyleGAN3-T	4.62	63.01	13.12
StyleGAN3-R	4.50	66.65	40.48

Utilizing StyleGAN2 alongside Karras et al.'s StyleGAN3 variants, namely StyleGAN3-T and StyleGAN3-R, has maintained a competitive FID score comparable to StyleGAN2. Both StyleGAN3-T and StyleGAN3-R exhibit excellent translation equivalence in terms of FID, with StyleGAN3-R uniquely offering rotation equivalence as anticipated. When applied to the FFHQ dataset at 1024x1024 resolution, these generators possess parameter counts of 30.0M, 22.3M, and 15.8M respectively, and their training durations are 1106, 1576 (a 42% increase), and 2248 hours.

Regarding adaptable layer configurations, StyleGAN3-T has seen significant improvements in equivariant quality, yet certain visual imperfections persist. It has been observed that the filter attenuation methods employed by Karras et al. (as specified in configuration G) are insufficient for the lowest resolution layers, which have high-frequency content nearing their bandwidth limits, necessitating more intense attenuation to prevent aliasing. To attain rotation-equivariance, StyleGAN3-R incorporates two adjustments: firstly, it swaps out 3x3 convolutions for 1x1 convolutions across all layers and doubles the quantity of feature maps to uphold the model's capacity. In this configuration, the transfer of information between pixels is facilitated exclusively through upsampling and downsampling processes. Secondly, sinc-based downsampling filters are replaced with radially symmetric jinc-based filters, which are meticulously designed using the Kaiser technique, with an exception for two vital sampling layers where maintaining the underlying non-radial spectral characteristics of the training data is essential. These modifications, collectively referred to as configuration R, significantly improved EQ-R without detrimental effects on the FID score, even though the number of trainable parameters per layer was reduced by 56%.

5 Conclusion

This paper proposes an image generation task based on StyleGAN3, which aims to efficiently and naturally output new anime-style face images. Experiments are also conducted to explore the application potential of StyleGAN3 in generating images in a specified comic style. The experimental results show that StyleGAN3 can efficiently and accurately capture the unique features of the anime style, and the generated images are visually highly similar to real anime works, which verifies the effectiveness and practicality of this method.

The anime-style image generation technology based on StyleGAN3 has greatly enriched the diversity of image creation, making the technology applicable to most specific application scenarios. For example, StyleGAN3 provides strong support in the fields of anime creation, game design, and virtual image generation. Through this technology, users can quickly generate anime images with personalized characteristics to meet their needs in entertainment, social interaction, etc., thereby greatly improving user experience and creation efficiency.

However, this paper has some limitations. According to previous studies, the model still faces certain challenges in generating certain complex details. For example, there is still room for improvement in the fine depiction of facial expressions of anime characters and the richness of clothing textures. In addition, the training time and computing resource consumption of the model are also one of the key issues that need to be solved at present. These problems limit the widespread promotion of StyleGAN3 in practical applications.

In future research work, we plan to continue to optimize the StyleGAN3 model, improve its ability to generate complex details and work to reduce computing resource consumption. In the future, we will explore more efficient training methods and resource optimization strategies to improve the operating efficiency of the model while maintaining the quality of generation. At the same time, we will also keep an eye on the latest progress in anime-style image generation technology, combine the latest research results in deep learning, computer vision, and other fields, and continuously expand the application scenarios and boundaries of

this technology. Promote the development of anime-style image generation technology based on StyleGAN3 to a higher level.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. International Telecommunication Union (ITU), The World in Stats: Internet Users (Latest available year).
2. J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer International Publishing, 694–711 (2016).
3. T. Park, M. Y. Liu, T. C. Wang, et al., Semantic image synthesis with spatially-adaptive normalization, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2337–2346 (2019).
4. P. Isola, J. Y. Zhu, T. Zhou, et al., Image-to-image translation with conditional adversarial networks, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134 (2017).
5. T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410 (2019).
6. T. Karras, S. Laine, M. Aittala, et al., Analyzing and improving the image quality of stylegan, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8110–8119 (2020).
7. A. Ahanian, L. Chai, P. Isola, On the "steerability" of generative adversarial networks, arXiv preprint arXiv:1907.07171 (2019).
8. T. Karras, M. Aittala, S. Laine, et al., Alias-free generative adversarial networks, Advances in Neural Information Processing Systems 34, 852–863 (2021).
9. Y. M. Liao, Y. F. Huang, Deep Learning-Based Application of Image Style Transfer, Mathematical Problems in Engineering 2022, 1693892 (2022).
10. A. Sauer, T. Karras, S. Laine, et al., Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis, In International Conference on Machine Learning, PMLR, 30105–30118 (2023).