ISSN 2959-6157

Research on Image Style Transfer Based on a Generative Model

Minjie Zhu

Zhuji Hailiang Foreign Language School, Zhuji,311800, China;

Abstract:

The generation of artificial intelligence technology is rapidly developing. As a representative, Sora created in text, image rendering, and video production shows excellent ability. This article discusses the image-style transmission algorithm based on generating a confrontation network (GAN) and compares it with the Sora model. Studies have shown that GAN performs well in image style transfer, especially in maintaining the consistency of the image content and generating high-quality images. In the future, the expected artificial intelligence technology will further improve the quality and diversity of generating content by combining the diffusion model and GAN and expanding its application potential in the fields of art creation and virtual reality.

Keywords: Generative AI, Sora, Image Generation, GAN, Image style transfer.

1. Introduction

With the rapid development of technology and the exponential growth of data volume, the field of generative artificial intelligence is becoming a key player in various social fields [1]. Sora is a prime example, known for his exceptional content production skills and strong language understanding abilities. This artificial intelligence marvel has demonstrated its extraordinary capabilities in creative fields such as producing written materials, visual arts, and video content, providing powerful help for numerous applications. At the same time, AI entities like ChatGPT are also making significant contributions across a wide range of practical uses. Sora is the first text-based video generation model developed by OpenAI [2]. It can generate up to 60 seconds, according to the text description. These videos are not only realistic but also include complex scene design, vivid character expressions, and complex camera movements. According to the Soras technology report, Sora not only understands the needs of users in the prompts but also understands how to present these contents in the real world. The appearance of Sora not only completely changed the market landscape of generating AI in the video field but also showed that the arrival of AGI (general artificial intelligence) may be earlier than expected [3].

The Sora model is a high-end video generation model developed by OpenAI. This model combines the technical advantages of the diffusion model and the transformer architecture to generate the final video image by gradually eliminating noise. This generation method not only makes the scene more detailed but also gives the model the ability to learn complex and dynamic.

SORA uses the power of deep learning algorithms based on advanced transformers to explain text data so that it can distinguish the complex, far dependence in the text and more effectively grasp the text semantic content [4]. In order to create images, it uses a complex deep learning architecture, such as generating a confrontation network (GAN). These structures consist of a pair of components: generators used to make images and discriminations used to evaluate the authenticity of these images. Sora is extremely versatile for video production, capable of generating clips in a variety of styles, from landscapes to animated characters, movie sequences, and more. It has 3D coherence, ensuring that objects and elements in your video remain consistent in three dimensions, even during complex camera movements. Additionally, Sora specializes in image style transfer, leveraging the transformation architecture or GAN framework to produce impressive results in the field of image style transfer.

Scholars from various countries have conducted in-depth research on this topic. The field of image style transfer remains a prominent research area in computer graphics and image processing. Reference [5] introduces a novel concept where segments of the output image are designed to mirror the content of the corresponding segments in the input image. To solve this problem, a contrastive learning-based method is introduced to enhance the interaction between input and output. This structure aims to enhance mutual information exchange, and empirical evidence indicates that it successfully facilitates style transfer between unpaired images. This technique not only achieves directional style transformation but also enhances image quality and slashes the time required for training. An et al. introduced ArtFlow [6]. Meanwhile, Zhu et al. [7] introduced CycleGAN, a system that concurrently trains both generative and discriminative networks. The generative network is responsible for fabricating deceptive content, while the discriminative network works to revert the altered image back to its original form. Additionally, they introduced a cyclic consistency loss function to facilitate transformations between unmatched image pairs.

This article explores the algorithm of style transfer using a GAN network-based model and compares it with SORA.

2. Image Style Transfer Technology and Generative Artificial Intelligence Technology

2.1 Introduction and Classification of Image Style Transfer Techniques

At present, the methods of style transfer mainly rely on several core deep learning architectures, including convolutional neural networks (CNN), generative adversarial networks (GAN), and recently emerging diffusion models. Convolutional neural networks, with their powerful feature extraction capabilities, can effectively capture image content and style information. They are typically used for preliminary feature analysis and transformation in style transfer tasks. Generative Adversarial Networks utilize adversarial training mechanisms to generate content with the target style through competition between generators and discriminators[8] while maintaining the structure and details of the original image. As an emerging technology, diffusion modeling provides a new method for generating samples for style transfer by simulating the gradual denoising process of data. This method has shown great potential in generating high-quality and high-resolution images[9].

Before the rise of deep learning and traditional style transfer methods, it is to analyze a specific style of image and give this style. Construct a mathematical statistical model and then apply it to the target image. Make stylistic changes to align with the constructed mathematical statistical model. Matching, but this method cannot separate the style of the image separately. Leave with the continuous development of convolutional neural network technology. The image style transfer based on degree learning has demonstrated better performance than traditional methods[10].

2.2 Introduction and Classification of Generative Artificial Intelligence Technology

As an important branch of technology, generating artificial intelligence (GAI) has gradually proved its huge potential and extensive application prospects. Generate the inherent mode and distribution of artificial intelligence to learn and simulate data so as to generate new content similar to the original data but not the same as the original data. This ability enables major progress in various fields, such as text, image generation, and voice synthesis in text, and has made significant progress. This article will explore the principles of generating artificial intelligence in detail, as well as the application and future development trends.

The core of artificial intelligence is its strong learning and power generation ability. It is usually based on deep learning technology, especially neural network models. This technology learns the inherent mode and characteristics of data by training a large amount of data. The generation of artificial intelligence models can be roughly divided into two categories: probability-based models and neural network-based models.

Probability generation models, such as hidden Markov models (HMM) and probability context syntax (PCFG) [11], are mainly dependent on probability statistics and probability graph models to characterize the data generation process. These models generate new data distributed by calculating the probability of data sequence. However, the performance of such models in processing complex data is usually poor, so it is difficult to capture the deep structure and semantic information of data [12].

In recent years, the generating models of neural networks, such as automatic encoders, mutant automatic encoders (VAE), and generating confrontation networks (GAN) and Transformers, have gradually become mainstream. These models understand the depth features of data and generate new data by constructing complex neural network structures. Among them, GAN is used to build two neural networks, generators, and discriminations for confrontation training to generate more realistic and diversified data. On the other hand, the transformer uses the self-attention mechanism [13] to capture the remote dependence in the data and generate high-quality text or images.

The generation of artificial intelligence technology is mainly divided into two types: generating confrontation networks (GANs) [14] and mutant automatic encoders (VAE). These two types of methods and goals are to generate content, but both have brought revolutionary changes to the field of artificial intelligence.

3. Principle of Image Style Transfer Based on GAN Network

3.1 Introduction to GAN Network

GANs[15] are based on the idea of adversarial training. They are essentially composed of two competing neural networks. This competitiveness helps them mimic any data distribution.

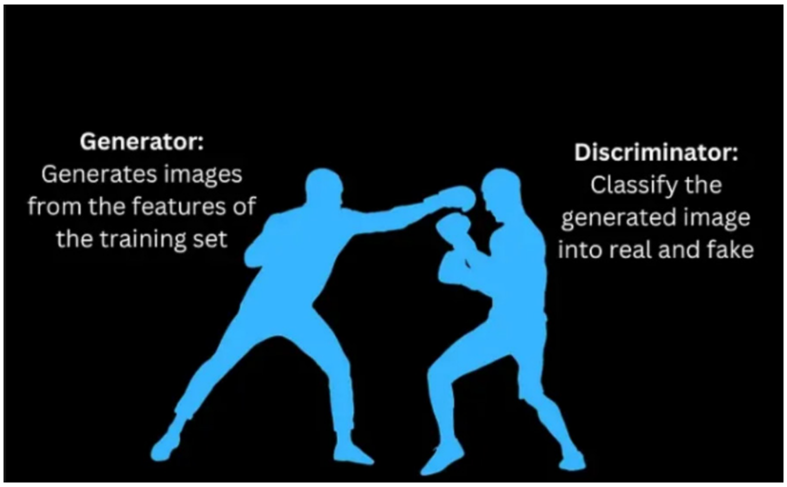
Generator:
Generates images from the features of the training set
Discriminator:
Classify the generated image into real and fake

Figure 1: GAN Network Structure Diagram[16]

As shown in Figure 1, We can imagine the GAN architecture as a battle between two boxers. In the process of conquering the game, both sides are learning each other's movements and skills. They didn't know much about their opponents at the beginning. As the game progresses, they learn and become better and better.

Generative Adversarial Networks (GANs) are an innovative neural network structure first proposed by Ian Goodfellow and his team in their paper "Generative Adversarial Networks," published in June 2014. The most remarkable ability of this network architecture lies in its potential to generate surreal images, videos, music, and text. GANs learn features from training data and create new images based on these learned patterns. For example, the image shown in Figure 1 is a result generated using GAN technology. The GAN architecture consists of two main networks:

1. Generator: Striving to transform arbitrary static into findings that seem to originate from genuine data collection[17].

2. Discriminator: Striving to ascertain whether a particular instance originates from the genuine dataset or is an imitation produced by the generator.

The primary objective of a discriminator is to ascertain the authenticity of an image, distinguishing between genuine and counterfeit images. Essentially, it boils down to a conventional supervised categorization task, enabling the utilization of conventional network architectures designed for classification. On the other hand, a generator's mission is to produce fraudulent data that mirrors the characteristics of the genuine data distribution as closely as possible. GANs primarily emulate the interaction between generators and discriminators by employing neural network structures.

The architecture features a series of cascaded convolutional layers, culminating in a fully connected layer equipped with a sigmoid activator. The sigmoid activator is employed due to the nature of the task being a binary classification challenge, where the objective of the network is to generate probability forecasts within the range of 0 to 1. In this context, a value of 0 signifies that the output image from the generator is deemed counterfeit, whereas a value of 1 indicates authenticity[18].

The inception of Convolutional Neural Networks (CNNs) dates back to 1998 when the LeCunLeNet-5 [19] was introduced, aiming to tackle the challenge of recognizing handwritten digits, thus marking the genesis of CNN technology. This network introduced a foundational structure that consists of a convolutional layer, a pooling layer, and a fully connected layer, which has since been adopted as the standard CNN architecture. Within the realm of CNNs, it is imperative to utilize various convolutional filters to process the input image (each filter passes through the image entirely to derive features, creating a feature map, with several filters generating multiple maps to increase convolutional depth), facilitating the extraction of diverse image characteristics (employment of multiple filters for convolution). Additionally, the architecture necessitates the stacking of multiple convolutional layers to perform deeper convolutions, thereby capturing intricate features within the image (achieving deep convolution).

The seminal GAN publication from 2014[19] featured the employment of a multi-layer perceptron (MLP) architecture for crafting both the generator and discriminator networks. Nevertheless, subsequent research has demonstrated that incorporating convolutional layers can bolster the discriminative power of the discriminators, subsequently refining the generator's precision and the entire model's performance. These enhanced GANs are referred to as DCGANs (Deep et al.).

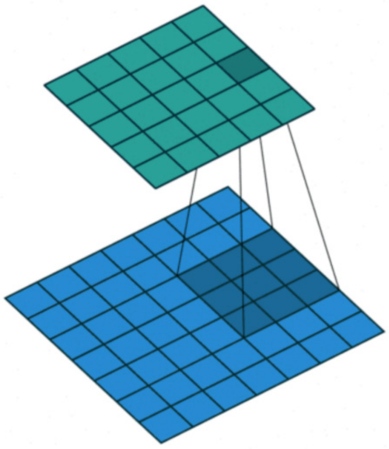

Figure 2: Convolutional Structure Diagram

3.2 Principle of Image Style Transfer Based on GAN Network

After the emergence of the confrontation network, the use of GANS for style transfer has become a hot topic in the research field.In 2017 [20], P.Silla et al. The work image of Silla et al. Posted by image translation with conditional EversarialNetworks [21] proposed that a method can generate corresponding style images according to given conditions. To reduce. The loss between the generated image and the original image is calculated by the use condition for generating the confrontation network (CGAN) loss function.

Incorporating the L1 penalty, the comprehensive workflow of the CycleGAN model is depicted in Figure 3. This workflow consists of a pair of generators, denoted as G and F, each fulfilling a distinct representation. If we consider the transformation from the X space to the Y space to be currently in effect, generator G is employed to synthesize images in the Y space using X space images as input. Conversely, generator F is tasked with creating images in the X space from Y space images. Concurrently, the CycleGAN architecture features a duo of discriminators, identified as DY and DX, which are responsible for verifying the legitimacy of the produced images in both the X and Y domains. As such, CycleGAN can be viewed as a merged entity of two separate GAN structures.

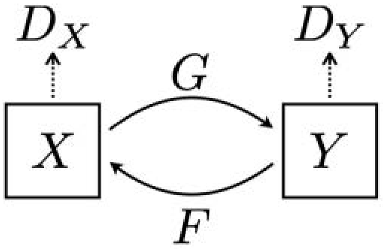

Figure 3: CycleGAN model

4. Experimental Principles and Analysis of Results

4.1 Experimental Parameters and Principles

This article uses the CycleGAN model to achieve style transfer of images; the experimental environment in this article is based on the Ubuntu 20.04 operating system and utilizes a parallel unified computing architecture. CUDA is a parallel computing architecture and development platform designed specifically for NVIDIA's graphics processing units (GPUs). It allows developers to leverage the powerful computing power of GPUs to write parallel code in C, C++, or other languages. CUDA has been widely applied in many fields, such as machine learning, graphics, scientific computing, etc. Its purpose is to use GPUs as high-performance devices. Parallel computing resources improve computational efficiency and accelerate application programs. PyTorch is an open-source machine learning framework based on Python, mainly used for deep learning and computer vision. The design of PyTorch incorporates the advantages of frameworks such as Torch while also incorporating many features of the Python ecosystem, making it easy to use. A significant feature of PyTorch is the dynamic computation of graphs, which means we can define and change the model structure at runtime, providing greater flexibility for experimentation and development. PyTorch has a wide range of applications in the fields of machine learning and deep learning, with good scalability and easy-to-use APIs.This article uses CUDA version 11.9, pytorch version 11.5, and Python version 3.7.

In the realm of generative adversarial networks, the generator's primary objective is to craft images that are so convincing that they can deceive the discriminator. Conversely, the discriminator's mission is to accurately classify whether an image originates from a generation or is a genuine, natural photograph. The generator and discriminator are optimized together because the images generated by this generator can be "indistinguishable from real" and are closer to real images. IS evaluates the quality of generated images. When measuring, two aspects will be considered: one is whether the generated image is "clear," which does not refer to high resolution. But it refers to the clear category of the image. On the other hand, the generated images have diversity; that is, images of each category try to have the same quantity as much as possible.

The dataset used in this article comes from images collected from the internet provided by Stanford, including 6288 landscape images and 1073 Monet paintings, both of which are 256 × 256 pixel RGB images

4.2 Results of experimental

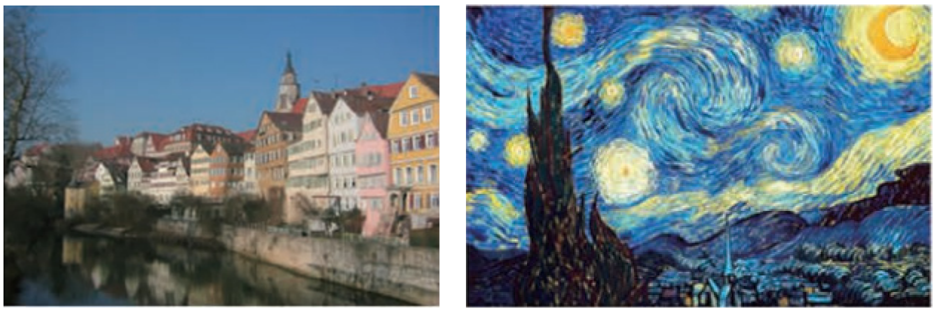

Figure 4: Original figure

The following is the style conversion of images using the GAN network. The original image and the style image that needs to be converted are shown in Figure 4. The generated result graph is shown in Figure 5; it effectively transformed the style and applied the style of Van Gogh's starry sky works to the original images.

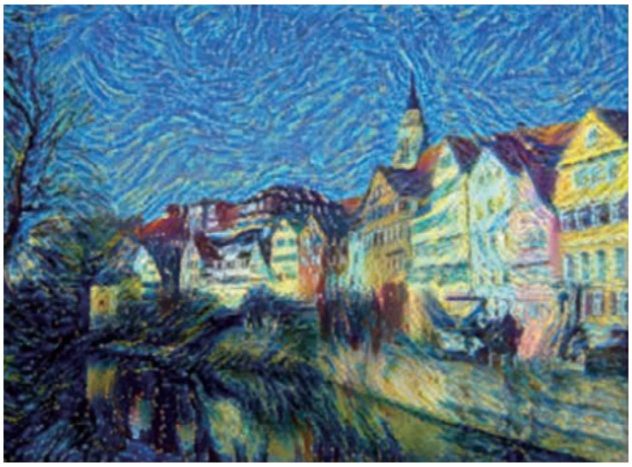

Figure 5: Image converted based on GAN network

Using Sora's feature to generate a photo and inputting the photo shown in Figure 4 before generating it, the generated result is shown in Figure 4. Both can achieve good style conversion, but we used an image style conversion based on the GAN model, which can achieve image style conversion more easily.

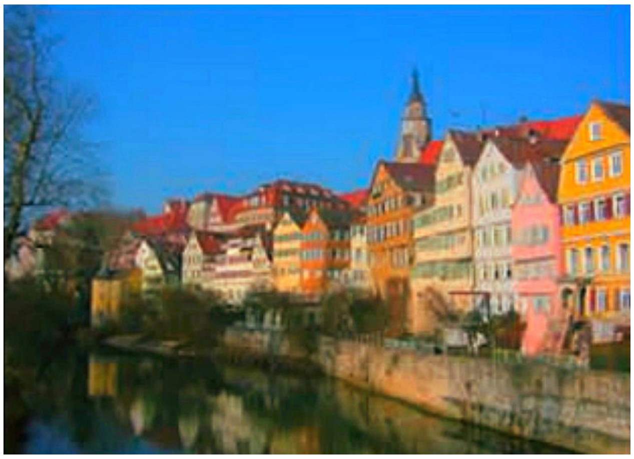

Figure 6: Image converted based on Sora

5. Conclusion and Future Development Trends

As technology advances at a breakneck pace and data expands exponentially, the domain of generative AI is becoming increasingly pivotal across diverse societal sectors. Emblematic of this trend, the Sora algorithm has showcased remarkable proficiency in areas like textual composition, visual rendering, and video generation, owing to its superior content generation skills and robust linguistic understanding. This piece delves into the Generative Adversarial Network (GANs)-based image style transfer mechanism and conducts a comparative analysis with the Sora algorithm. Research has shown that GAN models exhibit significant advantages in image style transfer, particularly in maintaining consistency of image content and generating high-quality images. Meanwhile, the Sora model's ability to generate videos also provides new possibilities for image style transfer. In the future, generative artificial intelligence technology will continue to develop towards greater efficiency and intelligence. With the continuous advancement of deep learning technology, generative models will further improve their performance in multimodal content generation, such as images, text, and videos. Especially in the field of image style transfer, the combination of GAN models and emerging diffusion models may bring even more remarkable results. The gradual denoising process of diffusion models and the adversarial training mechanism of GANs are expected to further improve the quality and diversity of generated content. Future research can focus on how to better integrate the advantages of different models, such as GAN and Transformer, as well as develop hybrid models with stronger generalization and generation abilities to enhance Sora's understanding and image generation capabilities.

References

[1] ZHU J-Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] // Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223 – 2232.

[2] ZHENG C, CHAM T-J, CAI J. The spatially-correlative loss for various image translation tasks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16407 – 16417.

[3] LEE J. Diffusion rendering of black ink paintings using new paper and ink models[J].Computers & Graphics, 2001, 25(2) : 295 – 308.

[4] ZHANG H,XUT,LIHS, et al.StackGAN: text to photorcalistic image synthesis with stacked generative adversarial networks[R]. Arxiv Preprint Arxiv:1612.03242, 2016.

[5] LEONG,ALEXANDER E,MATTHIAS B. A neural algorithmof artistic style[R]. Arxiv Preprint Arxiv:1508.06576, 2015b.

[6] HUANGG,CHEN DL,LlT H,et al,Multiscale dense net-works for resource efficient immage classification[C].ICLR,2018.

[7] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014

[8] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C] // NIPS. 2014.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[10] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11) : 2278 – 2324.

[11] XUE A. End-to-end chinese landscape painting creation using generative adversarial networks[C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021 : 3863 – 3871.

[12] ZHANG F, GAO H, LAI Y. Detail-preserving CycleGAN-AdaIN framework for image-to-ink painting translation[J]. IEEE Access, 2020, 8 : 132002 – 132011.

[13] YU J, LUO G, PENG Q. Image-based synthesis of Chinese landscape painting[J].Journal of Computer Science and Technology, 2003, 18(1) : 22 – 28.

[14] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C] // International conference on machine learning. 2017 : 214 – 223.

[15] ARJOVSKY M, BOTTOU L. Towards Principled Methods for Training Generative Adversarial =Networks[J]. arXiv preprint arXiv:1701.04862, 2017.

[16] 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.

[17] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two timescale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems(NIPS), 2017, 30.

[18] LIU Y. Improved generative adversarial network and its application in image oil painting style transfer[J]. Image and Vision Computing, 2021, 105 : 104087.

[19] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 1986(6) : 679 – 698.

[20] YAN M, WANG J, SHEN Y, et al. A non-photorealistic rendering method based on Chinese ink and wash painting style for 3D mountain models[J]. Heritage Science, 2022, 10(1) : 1 – 15.

[21] XIE S, TU Z. Holistically-nested edge detection[C] // Proceedings of the IEEE International Conference on Computer Vision. 2015 : 1395 – 1403.