AI Animation: Principle and Creative Practice of the Stable Diffusion Model

Wenjuan Chai

Contact email: okeefe@163.com

Guangzhou Academy of Fine Arts, Guangzhou, China

Abstract: Through analyzing the principle of the Stable Diffusion model, this article discusses how to use the Stable Diffusion model to draw moving surfaces in digital animation, achieve innovative screen performance effects, and form a unique artistic style, By analyzing the Stable Diffusion component; Operating principle of Diffusion; Analysis of the principles of the Stable Diffusion model such as CLIP training, exploration of various artistic forms and styles of AI animation generation, and proposed animation production methods based on the principles of the Diffusion model.

Keywords: AI animation, Stable Diffusion, Model principle, Animation creation

1. Introduction

The traditional animation production process is relatively cumbersome, requiring multiple creators to work together and meticulously complete the rendering of each frame of action. However, AI animation has affected the way humans create art. Through artificial intelligence (AI) driven animation solutions, it can simplify the animation production process and reduce costs, providing greater creative flexibility for small and medium-sized creators and projects with different needs. AI generated animation is an extension of AI generated images. From a technical perspective, animation is a sequence of multiple "pictures", that is, frames, and there are graphic, logical, and other levels of association between each frame in the sequence. Therefore, from a strict technical perspective, generating animation is more difficult than AI generating drawings. The release of Stable Diffusion is a milestone in the development of AI image generation, which is equivalent to providing the public with an available high-performance model. The generated images have higher quality, faster operation speed, and lower resource and memory requirements. The latest image generation capabilities demonstrated by the Diffusion model far exceed people's expectations. Images with stunning visual effects can be created directly from textual descriptions. What is the operating principle of the model? In order to control the type of image generated by the model, how can text be incorporated into the process for description? How does AI generate various artistic styles of animation through "text+"?

2. Introduction to Stable Diffusion

The rise of diffusion models can be seen as a major factor in achieving breakthroughs in the

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334612

field of artificial intelligence generated art.Stable Diffusion is a potential text to image diffusion model jointly developed by CompVis, Stability AI, and LAION (Large Scale Artistic Intelligence Open Network), which can generate incredible images with any text input, enabling billions of people to create amazing art in seconds. LAION is a non-profit organization with members from around the world that aims to provide large-scale machine learning models, datasets, and related code to the public. The basic dataset for Stable Diffusion is the 2b English language tag subset of LAION-5B. Due to the fact that Stable Diffusion is an open source model, many developers have modified and processed it since it was opened to the public, creating many different versions with different characteristics. For example:Stable Diffusion Demo、DreamStudio Beta、Stable Diffusion、Stable Diffusion WebUI 1.4、Pharmapsychotic Stable Diffusion、Deforum Stable Diffusion v0.3、Stable Diffusion Interpolation V2.1and so on.

3. Principle of Stable Diffusion Model

The content created by AI is generated through learning and evolution of massive database content, which is its underlying logic. This article simplifies some technical details of Stable Diffusion, leaving only its core algorithm logic. Stable Diffusion mainly includes the following four ways to generate animation: 1. generate animation based on "text" prompts as input; 2. Generating images or animations based on "video" material; 3. Generating images or animations based on "audio" material; 4. Generate animation based on "text+".

3.1 How Stable Diffusion Works

Stable Diffusion is not a single model, but a system composed of multiple components and models. From the perspective of the overall model, it includes a text encoder for capturing semantic information in text, and image generator . In the text encoder of the CLIP model, input some text strings and output a list of numbers to represent each word/token in the text, that is, convert each token into a vector. The digital information is then submitted to an image generator to generate the final pixel image (Figure 1).

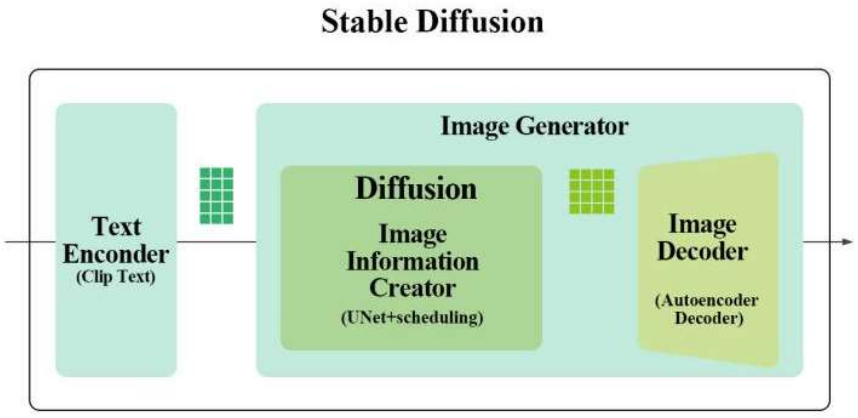

Figure 1. Stable Diffusion Components and Model System

3.1.1 Text Encoder

Section The language understanding component in the model uses the Transformer language model, which can convert input text prompts into token embedding vectors, that is, 77 token embedding vectors, with each vector containing 768 dimensions[1] . The Stable Diffusion model uses ClipText (GPT based model), and a larger language model can bring more image quality improvements than selecting a larger image generation component (Figure 1).

3.1.2 Image Generator

The interior of the image generator contains multiple components, mainly including the following two stages:

1) Image information creator

The image information creator achieves many performance gains. This component runs multiple steps to generate image information, among which steps are also parameters in the Stable Diffusion interface and library, typically defaulting to 50 or 100[2]. The image information creator runs entirely in the image information space (or hidden space), which makes it run faster than other Diffusion models that work in pixel space; Technically, this component consists of a UNet neural network and a scheduling algorithm (Figure 1).

2) Image Decoder

The image decoder draws the final image using the information matrix processed from the image information creator. The entire process can be run only once to generate the final pixel image, with dimensions of (3,512,512), i.e. (red/green/blue, width, height) (Figure 1).

3.2 Latent Diffusion Model

The potential diffusion model (LDM/Latent Diffusion Model) is a Stable Diffusion algorithm proposed by the CompVis and Runway teams in December 2021. Its diffusion process occurs in the potential space. Firstly, the image encoder compresses the image from Pixel Space to a smaller dimension of Latent Space, capturing the more essential information of the image. Then, the encoded image is spread forward and backward in the potential space[3].

The training of diffusion models can be divided into two parts:

3.2.1 Forward diffusion

Add noise to the images in the potential space and conduct a diffusion process. The forward diffusion process gradually adds Gaussian noise to the input image in a total of T steps. This process will generate a series of noise image samples$x_1, \ldots, x_T$. When$T \to \infty$, the final result will become a completely noisy image. As shown in Figure 2, a closed form formula is used to directly sample the noisy image at a specific time step t[4].

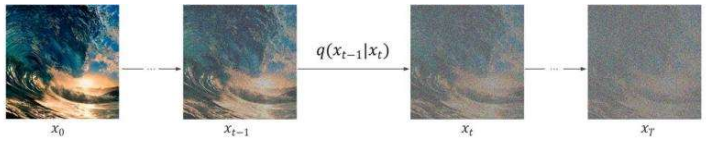

Figure 2. Forward Diffusion Process

Using a closed form formula to directly sample noisy images at a specific time step t , rather than designing an algorithm to iteratively add noise to the image [5]. The closed form sampling formula can be obtained through the reparameterization technique:

If $z \sim N(\mu, \sigma^2)$ then

$$z = \mu + \sigma\varepsilon \quad \text{where } \varepsilon \sim N(0,1)$$

Through this technique, we can represent the sampled image x_t as:

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\varepsilon_{t-1}$$

Then we can recursively expand it to obtain the closed form formula:

$$\begin{aligned}
x_t &= \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\varepsilon_{t-1} \qquad \varepsilon_{0,\cdots,}\varepsilon_{t-2,}\, \varepsilon_{t-1} \sim N(0, I) \\
&= \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\, x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\varepsilon_{t-2} + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\,\bar{\varepsilon}_{t-2} \\
&\vdots \\
&= \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1 x_0} + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\varepsilon \\
&= \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon
\end{aligned}$$

Here ε are $i.i.d.$ (independent and identically distributed) standard normal random variables. It is important to distinguish them using different symbols and subscripts because they are independent and their values may differ after sampling.

3.2.2 Reverse diffusion

Denoising an image based on some conditions to obtain a potential representation of the generated image. The denoising step can flexibly be conditional on text, image, and other forms (Text2img with text as the condition, img2img with image as the condition). By using a cross attention mechanism to enhance its denoising U-Net, the internal diffusion model is transformed into a conditional image generator.As shown in Figure 3, it is different from the forward process, $q(x_{t-1}|x_t)$ cannot be used to invert noise because it is difficult to calculate. Therefore, it is necessary to train neural networks $p_\theta(x_{t-1}|x_t)$ to approximate $q(x_{t-1}|x_t)$[6]. $p_\theta(x_{t-1}|x_t)$ It is approximately subject to normal distribution, and its mean and variance are set as follows: $\begin{cases} \mu_\theta(x_t, t) & := \tilde{\mu}_t(x_t, x_0) \\ \sum_\theta(x_t, t) & := \tilde{\beta}_t I \end{cases}$

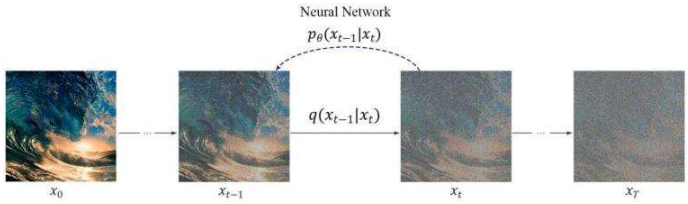

Figure 3. Reverse Diffusion Process

Loss is defined as negative logarithmic likelihood:

$$\text{Loss} = -\log\left(\rho_\theta(x_0)\right)$$

This setting is very similar to the settings in VAE. We can optimize the lower bound of the variation rather than the loss function itself. By optimizing the computable lower bound, the non treatable loss function can be optimized indirectly.

$$-\log\rho_\theta(x_0) \leq -\log\rho_\theta(x_0) + D_{KL}\left(q(x_{1:T}|x_0) \| \rho_\theta(x_{1:T}|x_0)\right)$$

$$\vdots$$

$$-\log\rho_\theta(x_0) \leq \mathbb{E}_q\left[\log\frac{q(x_{1:T}|x_0)}{\rho_\theta(x_{0:T})}\right]$$

It can be seen that forward diffusion can be calculated using a closed form formula. Reverse diffusion can be achieved using trained neural networks.

4. Creative Practice: AI Animation

AI animation is a popular direction for AI to generate works. Creators use AI painting models to generate stylized AI animations by inputting keywords related to screen effects such as themes, light, scenes, angles, atmosphere, and specific camera movement methods through "video, audio, video+text" and other means.AI animation *Rock, Paper, Scissors* uses the model Stable Diffusion to convert images frame by frame into animation styles, and turns live images into animation. AI Animation *Rock, Paper, Scissors* is produced by the Corridor team active in the field of special effects, Written&Directed by Niko Purringer and Sam Gorski. they have explored effective production methods and completed more than 7 minutes of AI animation.

4.1 Manufacturing difficulties

4.1.1 Eliminating "Noise"

The essence of both live photography and animation is continuous images. The problem is that after AI converts the style of each frame, the difference between frames is quite large, and the results of convergence are not smooth enough. If noise is frozen for a period of time, a more continuous image can be obtained, but the image will have a smear feeling and details will also appear from time to time. The solution is to reverse the operation, directly generating a

noise map from the original image frame, and then overlaying the noise map on the original video. The noise will not change too randomly, nor will it always stay in place.

4.1.2 Maintain the style of painting

After removing "noise", when the image is continuously played, it still encounters the problem of "flickering", and the style of each frame is not uniform. The solution is to use the Stable Diffusion training style model, and use the Nitrosocke Diffusion model, which is trained based on various styles, to adjust the image to a specific style. In order to further solve the defect of mixing various facial features and wrinkle details, a single actor's model will be specially trained (unified green screen background+costumes) to improve the continuity between frames. If it's not coherent enough, use Deflicker's special effects plug-in .

4.2 Production process

4.2.1 Video capture

The early stage screenwriting, filming, and performance arrangement of animation are created by artists. Two artists were responsible for filming and lighting, and the recording was completed by a recording engineer.

4.2.2 Training Actors

According to the above model principle, AI "transfers" the captured image frame by frame based on the actual shot. Use a Stable Diffusion training model that can transform actors and imitation styles into counterparts. The screen style comes from the Japanese animation masterpiece "Vampire Hunter D". The creator has taken hundreds of screenshots of different characters/objects and different scenes for training.

4.2.3 Generating Images

By adding keywords, excluding keywords, and adjusting the detail values back and forth, the creator generates 24 frames per second of "animation" from live footage in batches. There will be 1-2 frames with slight jumps. The creator's solution is to put the lens into the visual effects synthesis software Fusion, add the Deflicker plug-in, and set it to the Fluoro light mode. After feeling smooth enough with the naked eye, pluck the character out of the green screen, reducing the frame rate by half to 12, making it more animated.

4.2.4 Processing background

Use the church scene in the illusory engine to fine-tune the rendering and lighting according to the desired effect. Find a scene suitable for use as a background and capture 4 church images. Place the image into Stable Diffusion, enter keywords, and further adjust the image style.

4.2.5 Synthesis

Combine the characters in the foreground with the background to add a close-up effect. Paste four church background images together and move horizontally at high speed to increase the blurring effect; Pull closer to the image with the window when you need a close-up; Add the light rays plug-in to simulate the lighting effect. Convert the 3D model (candlestick) into an animation style and place it in the foreground of the camera to increase the blurring effect,

giving the lens a more realistic sense of camera rotation. Finally, add one or two hand-painted effect lines at key locations .

5. Conclusion

Currently, artificial intelligence and machine learning cannot create art independently, but they are helping people better understand how to create art. Facts have proven that AI is a good assistant for artists and an effective tool for making artistic works easier for everyone to access. For practitioners in creative industries such as art, film and television, advertising, architecture, fashion, and new media, AI painting is becoming a powerful assistant that can assist in daily work, improve output efficiency, and reduce time and economic costs. AI can also inspire creativity, expand new animation content, and innovate interactive forms. Currently, text-based graphics technology can expand image generation to animation generation through simple technology combinations, such as step-by-step diffusion generation, but the effect is not satisfactory. In summary, AI generates animation in the same way as generating images, and is still in the forefront of exploration. Currently, the Stable Diffusion model needs to be further improved. Firstly, it cannot complete high pixel animated images, and the image quality still needs to be improved; Secondly, shots and scenes are not complex, and role interaction, shot scheduling, and design are relatively simple.

References

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan,S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep laguage understanding. arXiv preprint arXiv:2205.11487, 2022.

[2]A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748-8763.

[3]R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.

[4]Sergios Karagiannakos,Nikolas Adaloglou, “How diffusion models work: The math from scratch”, AI Summer, 29-Sep-2022.

[5]Dhariwal, Prafulla, and Alexander Nichol. “Diffusion models beat gans on image synthesis.” Advances in Neural Information Processing Systems 34 (2021),pp.8780–8794.

[6]Ho, J. Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, pp.6840–6851.