Super Resolution Poster Generation Model Based on Adversarial Generative Network

Jinxuan Li, Xiuqin Deng

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou 510006, Guangdong, China

Abstract: ***With the continuous advancement of computer technology, image generation technology is also developing rapidly. This article proposes a poster generation method using Chinese text as input, which improves the second stage of the StackGan model to a super-resolution model (GAN_SR3). We introduce the CLIP model to map Chinese text to high-dimensional encoding, improving the matching degree between text and image. By inputting encoding and random noise in the first stage of StackGan, low resolution images are generated and then fed into the improved super-resolution model SR3 to obtain high-resolution posters. Compared with other models on the self built dataset, the experimental results showed that GAN_SR3 achieved higher Inception Score and smaller FID on the movie poster and online poster datasets.***

Keywords: Text image generation; Adversarial generative networks; Deep learning; Image Generation.

1. INTRODUCTION

With the rapid development of Internet technology, the speed of information dissemination is also showing an increasing trend. In this context, posters, as a vivid, intuitive, and attractive promotional tool, are widely used in various fields such as business and culture. Traditional poster design typically requires designers to create using an electronic drawing board in a specific application on a computer, which requires a lot of time and effort. Meanwhile, with the increasing demand for personalized services, manual production is clearly not efficient enough for design tasks that require a large or frequent production of posters. Therefore, achieving diverse design styles and improving poster production efficiency have become urgent challenges in the field of poster design.

On the other hand, traditional poster design often relies on the designer's experience and intuition, while poster generation models based on adversarial generative models rely on mathematical laws and computer technology, and are not limited by the personal experience and limitations of traditional designers. To some extent, this model can handle more complex design tasks. In addition, by utilizing technologies such as artificial intelligence to create posters, we can draw on a large amount of poster data and use deep learning techniques to continuously optimize algorithms, thereby achieving a more efficient poster generation process. This method brings more possibilities to the field of poster design and is expected to improve the efficiency and diversity of poster production. With the continuous advancement of computer technology, image generation models have received widespread attention and research, and have been applied in several fields. In financial technology, Pal et al. [1] developed an AI-based credit risk assessment system with intelligent matching mechanisms for supply chain finance. Computer vision research has seen significant progress through several innovations: Peng et al. [2] proposed a source-free domain adaptation method for human pose estimation, Pinyoanuntapong et al. [3] introduced GaitSADA for mmWave gait recognition, and Zheng et al. [4] created DiffMesh for motion-aware human mesh recovery from videos. Zhang et al. [5] extended machine learning applications to biomechanical anomaly detection in big data environments. Smart infrastructure solutions have emerged through Fang's [6] QoS-aware cloud-edge architecture for water management and Qi's [7] interpretable neural network for inventory forecasting. Healthcare applications have advanced with Wang's [8] RAGNet model for arthritis risk prediction, while Zhou et al. [9] applied LSTMs to optimize UAV path planning. Economic analytics has benefited from Yang et al.'s [10] big data-driven approach for economic cycle prediction, complemented by Tu's [11] Log2Learn system for intelligent network optimization through log analysis. Computer vision continues to evolve with Ding et al.'s [12] novel attention mechanism for clothing-changing person re-identification. Finally, in materials science, Wang et al. [13] demonstrated AI's potential in mechanical engineering through their multiscale shakedown analysis for predicting loading capacity of auxetic tubular structures.

2. RELATED WORK

2.1 Generative Adversarial Networks

Goodfellow et al. proposed Generative Adversarial Networks (GANs) in 2014. GANs, as a milestone in the field of unsupervised learning, have opened up new research avenues in the field of data generation. In the GANs model, a neural network acts as a generator (G), taking noise as input and generating the target image through a series of convolution, pooling, and other operations. Then, it is passed along with the real image to the next neural network, which acts as a discriminator (D), responsible for distinguishing which input image is the real image and which is the generated image.

In the original generative adversarial network, both the generator and discriminator used multi-layer perceptrons as the network framework. In order to learn the distribution R_p of the generative model from the real data X_{real}, a random noise vector z is generated by sampling from the prior random distribution, and then the random noise vector z is input into the generative model to generate X_{fake}. The objective function of the generator model optimization process is shown in (1):

$$min_G V(D,G) = E_{z-pz}\left[log\left(1 - D\left(G(z)\right)\right)\right] \tag{1}$$

The task of the discriminator is to determine the authenticity of the data generated by the generator and output a probability value indicating that the data is true. Usually, the discriminator is trained to become a binary classifier that receives two types of data: real data and fake data generated by the generator. For each input data, the discriminator generates a probability value between 0 and 1, indicating the likelihood that the input data is real. When the output of the discriminator is incorrect, its weight will be updated to improve its accuracy. By continuously optimizing weights, the discriminator gradually learns how to better distinguish between real data and fake data, thereby assisting the generator in generating more realistic virtual data. This process is one of the core mechanisms in generative adversarial networks, used to drive the generator to continuously improve the authenticity of the generated data.

The objective function (2) of the optimized discrimination model is shown in equation:

$$max_D V(D,G) = E_{x-pdata(x)}[logD(x)] + E_{z-p(z)}\left[log\left[\left(1 - D\left(G(z)\right)\right)\right]\right] \tag{2}$$

2.2 Denoising Diffusion Probabilistic Models

In 2020, Ho et al. proposed the Denoising Diffusion Probabilistic Models (DDPM) model [7] to address the instability of generative network GANs during training and the inability to generate high-resolution images at once. DDPM is divided into forward diffusion process and backward inverse diffusion process, where the forward process is the process of gradually adding Gaussian distributed noise to real image data until the data becomes completely noisy. The specific steps are as follows: first, sample x_0 from the real image distribution $\mathrm{q(x_0)}$, then randomly generate a diffusion frequency t in the interval (1,..., T), and then sample a random noise from the standard normal distribution. According to equation (3), obtain the sample x_t.

$$x_t = \sqrt{\bar{a}x_0} + \sqrt{1 - \bar{a}_t\theta} \tag{3}$$

Where $\sqrt{\bar{a}_t}$ is a hyperparameter.

The backward diffusion process is derived as follows:

$$\mathrm{x_{t-1}} = \frac{1}{\sqrt{\mathrm{a_t}}}\left(\mathrm{x_t} - \frac{1-\mathrm{a_t}}{\sqrt{1-\bar{\mathrm{a}}_\mathrm{t}}}\epsilon\theta(\mathrm{x_t},\mathrm{t})\right) \tag{4}$$

Fit the noise εθ to obtain the original image after t rounds of denoising.

3. POSTER GENERATION MODEL

After drawing inspiration from the StackGAN model [10] and DALL-E2 model [6], the GAN-SR3 model was born by improving the second stage of the StackGAN generation model. The model process is shown in Figure 1:

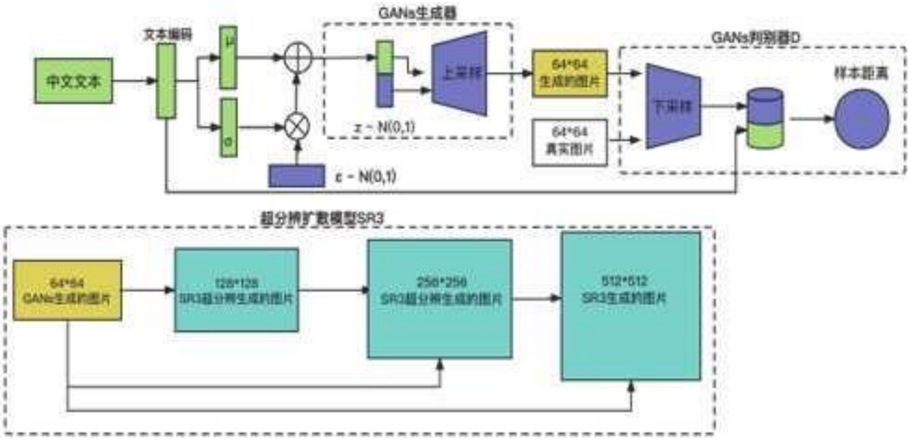

Figure1: The flow chart of GAN_SR3 model

Firstly, a pre trained Chinese CLIP model [15] was used for text embedding. By inputting prompt text into the model, a text vector was obtained$T_\theta = (T_1, \cdots, T_N)$.

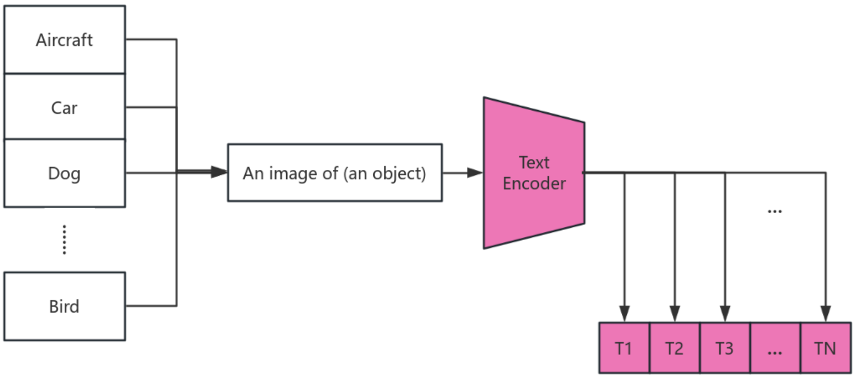

Figure2: The text encoding process of CLIP model

Subsequently, the randomly generated one-dimensional Gaussian noise is concatenated at the end of the obtained text vector and input into the generator G of the GANs model, and enters the second stage.

Generator G generates a 64 * 64 image through up sampling and convolution, and inputs it together with the real image to discriminator D. In discriminator D, the idea of the WGAN model [16] is borrowed to combine the generated simulated sample distribution P_g with the original sample distribution P_r, treating them as a set of all possible joint distributions. Then, calculate the distance between them and the expected value of the distance. Through model training, the generator optimizes towards the lower bound of the expected value to better simulate all possible joint distributions. In this process, Wasserstein distance is used to measure the distance between two samples, namely:

$$|f(_{x_1}) - f(_{x_2})| \leq {}^{k}|_{x_1} - {}_{x_2}| \tag{5}$$

Where k is the Lipschitz constant of function f (x).

According to equation (6), the distance formula for the distribution of the two can be summarized as:

$$L = K\left[D\left((\text{real}) - D\left(G(x)\right)\right)\right] \tag{6}$$

Due to the fact that generator G aims to generate results that are closer to the distribution of the original samples, the loss function of generator G can be abbreviated as:

$$G(loss) = -D\big(G(x)\big) \tag{7}$$

The task of the discriminator is to distinguish between the two, therefore the loss function is:

$$D(loss) = D\big(G(x)\big) - D(red) \tag{8}$$

Through continuous iterative training, when generator G can generate images that discriminator D cannot distinguish between real and fake, the training enters the third stage.

In the third stage, the SR3 model that achieves super-resolution through repeated refinement is adopted [17]. The SR3 model, as shown in Figure 3, applies the denoising diffusion model to conditional image generation. Through continuous iteration refinement, it predicts the required noise removal for each iteration, ultimately achieving the generation from low resolution images to high-resolution images.

The prediction of noise in the SR3 model is learned using the U-Net network [18] shown in Figure 4, where the real image and the target resolution image obtained by bicubic interpolation upsampling are cascaded in the channel dimension as inputs to the U-Net.

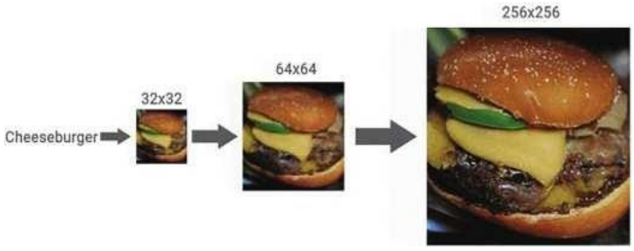

Figure3: TheprocessofSR3modelimagesuper-resolution

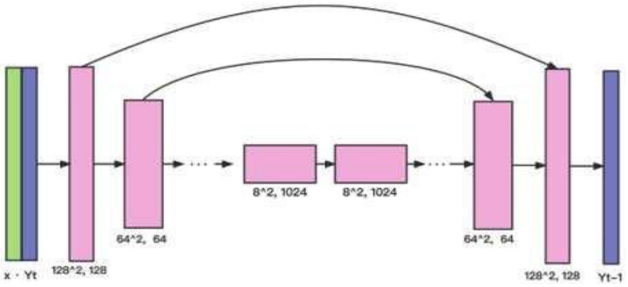

Figure4: Structure diagram of the U_net model

The training steps of the SR3 model are shown in Algorithm 1:

Algorithm 1: Trains denoising model f_θ

1) Input: Input/output images on the dataset $\mathrm{D} = \{\mathrm{x}_\mathrm{i}, \mathrm{y}_\mathrm{i}\}_\mathrm{i}^\mathrm{N}$, hyperparameter γ, where x is the low resolution image and y is the high-resolution image
2) Sampling from the image distribution p (x, y) to obtain image pairs (x, y)
3) Sampling noise from standard Gaussian distribution
4) Fit using gradient descent method
5) $\theta_\theta \left\| \mathrm{f}_\theta\left(\mathrm{x}, \sqrt{\gamma}\mathcal{y}_0 + \sqrt{1-\gamma}\epsilon, \gamma\right) \right\|_\mathrm{p}^\mathrm{p}$
6) Repeat steps 2-4 until convergence.

As shown in pseudocode algorithm 2 for iterative denoising:

Algorithm 2: Pseudo code for denoising in the T-th iteration

1) Sampling y_T from Gaussian distribution

2) t=T
3) $z \sim N(0,1)$ if $t > 1$, else $z = 0$
4) $y_{t-1} = \frac{1}{\sqrt{n}}\left(y_t - \frac{1-a_t}{\sqrt{1-y_t}} f_\theta(x, y_t, \gamma_t) + \sqrt{1-a_t^z}\right)$
5) t=t-1
6) Until t=1
7) Output y_0

4. EXPERIMENT

4.1 Dataset

During the training phase, two different types of training datasets are selected, namely movie posters and online posters. These datasets were obtained through web crawlers from design websites, social media platforms, and design work databases. The movie poster dataset includes a total of 39515 movie poster images. Each movie poster image is accompanied by 10 different descriptive texts that cover various information and features about the movie poster. In addition, the online poster dataset contains a total of 11788 poster images. Similar to movie posters, each online poster image also comes with 10 different descriptive texts used to describe the content and features of different types of posters.
The distribution of movie poster data quantity is shown in Figure 5:

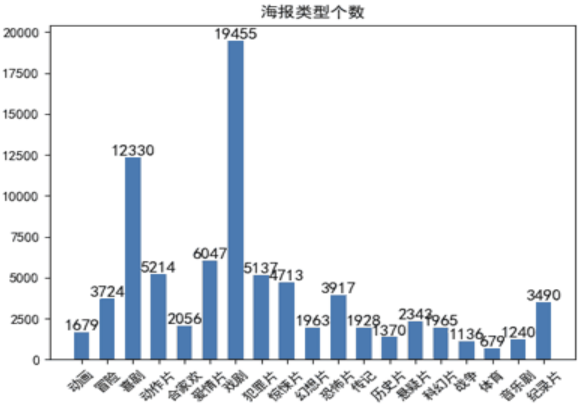

Figure5: Distribution of the number of movie posters

In the online poster dataset, poster types include academic posters, concert posters, art exhibition posters, sports event posters, speech posters, academic conference posters, tourism promotion posters, hotel promotion posters, restaurant menu posters, fitness activity posters, charity activity posters, technology exhibition posters, campus recruitment posters, wedding invitation posters, birthday party posters, children's activity posters, fashion show posters, theater performance posters, community activity posters, health and wellness posters, etc. These poster types cover the needs of different fields and occasions, including art and culture, entertainment, business, community, and personal activities. Each type of poster has its own unique design style and content requirements, allowing the poster generation model to learn more about poster design styles and knowledge, enhancing the model's generalization ability.

4.2. Evaluation Criteria

For text generation image models, it is not only necessary to generate clear images, but also to maintain the diversity of image elements. Therefore, in order to evaluate the quality of images, internationally recognized evaluation indicators Inception Score (IS) and Frechet Inception Distance (FID) are used. Among them, IS is usually used as an indicator to evaluate diversity, while FID is used as an indicator to evaluate the quality of generated images.

The corresponding calculation formula for IS is:

$$IS(G) = \exp\left(E_{x-p_g} D_{KL}(p(y|x) \| p(y))\right) \tag{9}$$

Among them, $x \sim P_g$ represents the sample x generated by the generative model, and y represents the output category of the generated image in the initial v3 classification model. When the KL divergence between the edge distribution p(y) and the conditional distribution p(y|x) is large, the samples generated by the model have diversity.

The calculation formula for FID is shown in equation (10):

$$\mathrm{FID}(\mathrm{P},\mathrm{G}) = \left\|\mu_p - \mu_g\right\|^2 + \mathrm{T}\gamma\left(\Sigma_p + \Sigma_g - 2\left(\Sigma_p\Sigma_g\right)^{\frac{1}{2}}\right) \tag{10}$$

Among them, P is the real image, and G is the generated image.

5. ANALYSIS OF EXPERIMENTAL RESULTS

After sufficient training of GAN-SR3 with StackGan [12] and DDPM on the same dataset, testing was conducted.

Take the same text "Generate an image with Zongzi elements and Loong Boat Festival words" as input, input GAN_SR3 model, StackGan model and DDPM model respectively, and get the image generation results as shown in Figure 6.

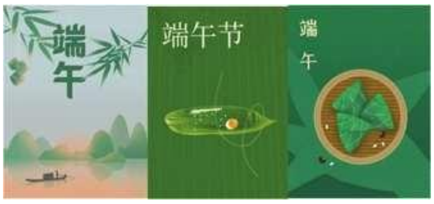

Figure 6: The images generated by DDPM, StackGan, GAN_SR3

From Figure 6, it can be seen that the images generated by the GAN-SR3 model are richer in content, closer to textual descriptions of items, and have higher resolution compared to StackGan [11]. This is because the second stage SR3 model uses three diffusion models to make the generated content match the text information more closely, with higher confidence, more detailed information, and higher resolution.

At the same time, the patterns generated by GAN_SR3 are more in line with the requirements of the text than those generated by the DDPM model, and can also generate the required text more accurately. The reason is that the understanding of text in the images generated by GAN_SR3 comes from the training set. During training, text prompts often contain relevant text titles and position information, while DDPM generates text titles and positions more randomly and uncontrollably, resulting in its inability to efficiently generate the poster images we need.

In addition, by inputting similar text prompts, GAN_SR3 is guided to generate relevant images. For example, input the text "Generate an image with the imagery of a young man yearning for a distant place and the words' May Fourth Youth Day '" to generate the image shown in Figure 7.

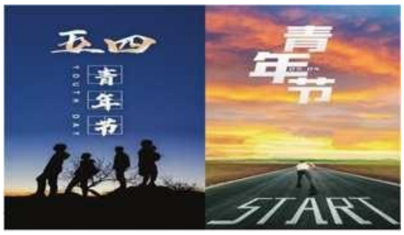

Figure7: Image of Youth Day generated by GAN_SR3

Finally, IS and FID metrics were used on the test dataset to measure the quality and diversity of the images generated by the model, and the inference time was used to compare the time required for the model to generate images. By generating 1000 images on different datasets and inputting them all into the Inception v3 classification network, the results are shown in Table 1:

Table 1: Experimental results of different models

Data set	**Movie Poster**			**Online poster**		
	IS	FID	Inf.time	IS	FID	Inf.time
StackGan	3.27	51.07	12.1s	3.74	49.58	11.4s
DDPM	3.45	52.67	8.7s	4.10	44.31	8.1s
GAN_SR3	3.70	42.52	8.6s	4.05	39.97	7.9s

6. CONCLUSION

Image synthesis technology is an important research area in the field of computer vision, with enormous potential for development in people's daily lives. Among them, generating images based on text content is a popular problem with wide practical value, which can be applied to various fields such as poster design, film production, animation production, book cover design, etc. In this article, we propose a new method that applies CLIP to text embedding models to obtain higher quality text features. Meanwhile, in order to improve the quality of generated images, a phased image generation strategy was adopted.

In the first stage, low resolution images are generated using textual information. Then, in the second stage, using the image generated in the first stage as input, the resolution of the image is gradually increased through three DDPMs. In order to prevent the loss of underlying features during the encoding process of the second stage input image, skip links are introduced between each DDPM. This simultaneously enhances the fusion of feature images at multiple scales, further improving the quality of generated images.

FUNDING PROJECTS

National Natural Science Foundation of China (12401136), Guangdong Province Graduate Education Innovation Plan Project 2024 (2024SFKC030).

REFERENCES

[1] Pal, P. et al. 2025. AI-Based Credit Risk Assessment and Intelligent Matching Mechanism in Supply Chain Finance. Journal of Theory and Practice in Economics and Management. 2, 3 (May 2025), 1–9. DOI:https://doi.org/10.5281/zenodo.15368771

[2] Peng, Qucheng, Ce Zheng, and Chen Chen. "Source-free domain adaptive human pose estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[3] Pinyoanuntapong, Ekkasit, et al. "Gaitsada: Self-aligned domain adaptation for mmwave gait recognition." 2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS). IEEE, 2023.

[4] Zheng, Ce, et al. "Diffmesh: A motion-aware diffusion framework for human mesh recovery from videos." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

[5] Zhang, Shengyuan, et al. "Research on machine learning-based anomaly detection techniques in biomechanical big data environments." Molecular & Cellular Biomechanics 22.3 (2025): 669-669.

[6] Fang, Z. (2025). Adaptive QoS‑Aware Cloud‑Edge Collaborative Architecture for Real‑Time Smart Water Service Management.

[7] Qi, R. (2025). Interpretable Slow-Moving Inventory Forecasting: A Hybrid Neural Network Approach with Interactive Visualization.

[8] Wang, Y. (2025). RAGNet: Transformer-GNN-Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction.

[9] Zhou, Dianyi, et al. "Research on LSTM-driven UAV path planning." Fourth International Conference on Advanced Algorithms and Neural Networks (AANN 2024). Vol. 13416. SPIE, 2024.

[10] Yang, W., Zhang, B., & Wang, J. (2025). Research on AI Economic Cycle Prediction Method Based on Big Data.

[11] Tu, T. (2025). Log2Learn: Intelligent Log Analysis for Real-Time Network Optimization.

[12] Ding, Y., Wang, X., Yuan, H., Qu, M., & Jian, X. (2025). Decoupling feature-driven and multimodal fusion attention for clothing-changing person re-identification. Artificial Intelligence Review, 58(8), 1-26.

[13] Wang, Lizhe, et al. "Loading capacity prediction of the auxetic tubular lattice structures by multiscale shakedown analysis." Composite Structures 314 (2023): 116938.

Author Profile

Jinxuan Li (1997-), male, master's student, mainly researching machine learning and data mining.

Xiuqin Deng (1966-), female, undergraduate, professor, mainly researching data mining and machine learning.