Research on Controllable Image Generation Technology for Chinese Text

Yongxia Hu [1,2] and Dong-Hyun Kim [3,*]

[1] Department of Information Engineering, Guangzhou Vocational College of Technology & Business, Guangzhou, China
[2] Department of Computer and Information Engineering, Youngsan University, Yangsan-si, Republic of Korea
[3] Department of Mechanical and Automotive Engineering, Youngsan University, Yangsan-si, Republic of Korea
Email: iamhuyongxia@gmail.com (Y. H.); dhkim@ysu.ac.kr (D.H.K.)
*Corresponding author

Abstract— **With the rapid development of artificial intelligence technology, text-generated image technology has garnered widespread attention, which shows great potential in enhancing human-computer interaction, increasing the credibility of visual content, and creating novel works of art. In this paper, an improved Generative Adversarial Network (GAN) model based on attention mechanism, Improved AttnGAN, is proposed to deal with the challenges of existing technologies in dealing with complex text input, improving image clarity and authenticity, and enhancing semantic consistency between text and image. By introducing the SimAM attention mechanism and optimizing the AttnGAN architecture, our model achieves significant improvements in both image generation quality and variety. The experimental results show that the Improved AttnGAN model is superior to StackGAN, DM-GAN, MirrorGAN, DF-GAN, AttnGAN, and other models. The Improved AttnGAN has obvious advantages in terms of image quality and realism.**

Keywords—**image generation, Chinese text, GAN, self-attention mechanism, AttnGAN, Improved AttnGAN**

I. Introduction

Text image generation is a technology that generates the corresponding image by using natural language description, which belongs to the cross field of computer vision and natural language processing. It has important theoretical significance and research value. For example, it can be used to enhance human-computer interaction, improve the credibility of visual content, and create novel works of art. The primary research challenges in text image generation involve understanding the semantic information of text, converting this semantic information into image features, and generating high-quality, diverse images [1]. At present, the research methods in the field of text-generated images are mainly based on generative antagonistic networks, which can achieve a certain degree of text-to-image mapping by using its powerful representation and generation capabilities. However, there are still many challenges and problems in the field of text-generated images, such as how to deal with complex and diverse text input, how to improve the clarity and authenticity of images, and how to strengthen the semantic consistency between text information and images [2]. Therefore, there is still a lot of research space and development potential in the field of text-generated images, which is worthy of further exploration and innovation.

In order to meet the requirement of improving the semantic consistency of GAN, this paper introduces the self-attention mechanism on the basis of the traditional GAN, and constructs the AttnGAN model. This model can focus on different angles of words and paragraphs, and generate images that are more consistent with the text. This paper makes another improvement on the basis of AttnGAN, which improves the quality of the generated image while ensuring the semantic consistency between the text and the image.

This paper makes several key contributions to the field of controllable image generation for Chinese text. Firstly, we propose an Improved AttnGAN model that integrates the SimAM attention mechanism, enhancing both image quality and semantic consistency without increasing model complexity. Secondly, our model demonstrates superior performance compared to existing state-of-the-art models, as evidenced by both qualitative and quantitative evaluations. Lastly, we provide a detailed analysis of the model's architecture and its impact on cross-modal generation tasks, offering new insights for future research.

The remainder of this paper is organized as follows. Section II provides the theoretical basis for our proposed model, including an overview of image generation technology and the attention mechanism. Section III describes the architecture of our controllable image generation model and how it integrates the SimAM attention mechanism. Section IV presents the empirical analysis, including the experimental setup, results, and discussion. Finally, Section V concludes the paper and highlights potential areas for future research.

Manuscript received March 5, 2025; revised April 8, 2025; accepted June 19, 2025; published August 19, 2025.

II. Literature Review

A. Image Generation Technology

The generative antagonistic network consists of a generator and a discriminator, wherein the generator generates a corresponding image according to a word vector or a sentence vector obtained by coding, and the discriminator is used for discriminating the authenticity of the image [3]. Continuous learning can not only improve the quality of the image generated by the generator, but also improve the ability of the discriminator to identify the true and false images, and ultimately get more realistic false images, so as to achieve the purpose of identifying the true and false images.

The generator takes a low-resolution image as input, processes the blurry image in accordance with the data distribution of the real image, and finally output the generated sample. The discriminator is a binary classifier, which distinguishes the input generated sample from the real image. The final recognition result is represented by a probability value between 0 and 1. The larger the value is, the closer the generated sample is to the real image, and the smaller the value is, the greater the difference between the generated sample and the real image is [4]. The result is fed back to the generator for iterative training. At the beginning of the training process, the generator and the discriminator will update their own parameters to minimize the value of the loss function, and continue to optimize the iteration, so that the network structure reaches a balanced state, and the final output value reaches 0.5. At this time, the discriminator has been unable to distinguish the generated sample from the real image, and the generator also achieves the optimal effect. The structure diagram of the generation countermeasure network GAN is shown in Fig. 1.

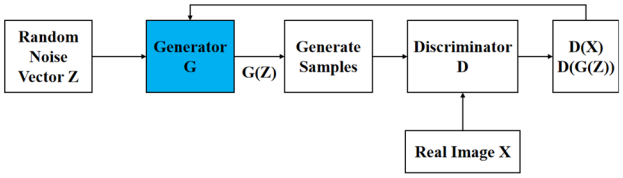

Fig. 1. Generated countermeasure network structure diagram.

The function of the generator is to minimize the objective function. The smaller the probability value of the objective function is, the stronger the ability of the generator to disguise the generated samples as real data is. After the alternate training of the two models, the discriminator loses the ability to distinguish the authenticity of the generated sample and the real image, and the discriminator and the generator network module are optimal when they reach the Nash equilibrium state. GAN is used in many fields, such as face generation, image generation, image inpainting and so on, and it is most widely used in the field of image generation. StackGAN is the first two-stage model and the most typical GAN model. AttnGAN is a new attention-generating network developed on the basis of StackGAN [5].

The AttnGAN model architecture consists of multiple generator and discriminator networks, each taking the previous network's hidden state and the word-context vectors as inputs in image generation. The generator in AttnGAN utilizes an attention mechanism to draw different sub-regions of the image by focusing on words that are most relevant to the sub-region being drawn. Specifically, the generator first encodes the natural language description into a global sentence vector and generates a low-resolution image in the first stage. In subsequent stages, it utilizes the image vector in each sub-region to query word vectors by using an attention layer to form a word-context vector, which is then combined with the regional image vector to generate new image features.

We have referenced several state-of-the-art models in the field of text-to-image synthesis, including: StackGAN, which introduced a two-stage generative adversarial network for high-resolution image synthesis. AttnGAN, which incorporates attention mechanisms to improve the fine-grained generation of images from text descriptions. DM-GAN, which focuses on improving the diversity and quality of generated images. MirrorGAN, which explores symmetric structures in image generation. And DF-GAN, which enhances the feature fusion in the generation process.

B. Attention Mechanism

SimAM is an attention mechanism that based on local self-similarity of feature maps. It dynamically adjusts the weight of each pixel by calculating the similarity between each pixel and its surrounding pixels in the feature map, so as to enhance the important features and suppress the irrelevant features [6]. The innovation of SimAM lies in its parameter-free feature, which enables the model to achieve excellent performance while maintaining low complexity. The SimAM attention mechanism has been applied to several computer vision tasks, such as image classification, object detection, and image segmentation, and has achieved impressive results. The full 3D weights for attention is shown in Fig. 2.

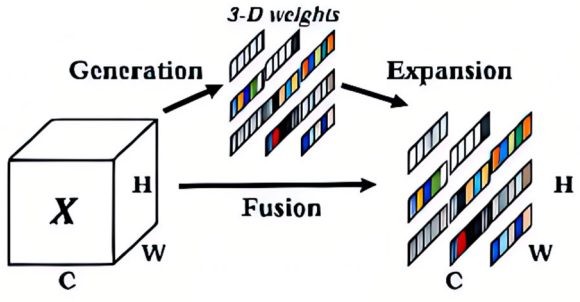

Fig. 2. Full 3D weights for attention.

In this section, we introduce the theoretical basis of our proposed model. We utilize the Generative Adversarial Network (GAN) as the core framework. GAN comprises a generator, G, and a discriminator, D. The generator, G, is designed to produce realistic images, whereas the discriminator, D, evaluates the authenticity of these images. We represent the input text as T, the word embeddings as W, and the generated image as I.

In order to better extract the image feature information, it is necessary to add an importance weight to each convolutional neural network, that is, to assign a larger weight to the neurons with spatial inhibition effect, which can be achieved by measuring the linear separability between neurons. Therefore, the neuron energy function can be obtained.

$$e_{t(w_t,b,y,x_i)} = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \quad (1)$$

where M is the number of energy functions; Minimizing the energy function is equivalent to training the linear separability between neuron t and other neurons in the same channel. A regularization term is added to the energy function and reduces to:

$$e_{t(w_t,b,y,x_i)} = \frac{1}{M-1}\sum_{i=1}^{M-1}[-1 - (w_t x_i + b_t)]^2 + [1 - (w_t t + b_t)]^2 + \lambda w_t^2 \quad (2)$$

The lower the energy, the greater the difference between the neuron t and the peripheral neuron, and the higher the importance, the enhancement processing of the input features can be expressed as:

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right)\Theta X \quad (3)$$

E represents the energy function; X denotes the input feature layer. The energy function E can implement the attention mechanism once it has passed through the Sigmoid activation function.

III. Materials and Methods

This section details the construction of Controllable Image Generation Model.

A. Architecture of Controllable Image Generation Model

The integration of the SimAM attention mechanism into the AttnGAN model offers a novel approach to improve both image quality and semantic consistency in a computationally efficient manner.

To significantly enhance the alignment between image generation and text semantics, this paper makes an innovative improvement on the basis of AttnGAN model, and skillfully embeds SimAM attention mechanism in its self-attention mechanism [7]. This design aims to efficiently and accurately extract the spatial features of the feature map through the SimAM mechanism, without introducing additional model parameters, thus maintaining the model's simplicity and computational efficiency. The network embedding the SimAM attention mechanism is shown in Fig. 3.

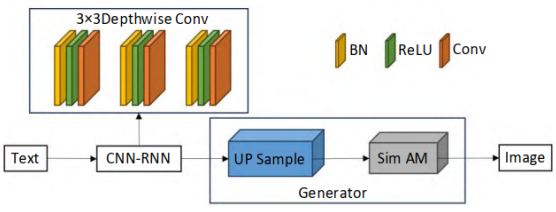

Fig. 3. Networks Embedding the SimAM Attention Mechanism.

In this paper, the SimAM module is integrated into the structure of the generator, which enables the generated images to be more focused on the key and important information in the input text [8]. In this way, the generated image content is not only closer to the text description, but also more vivid and accurate in detail. At the same time, thanks to the organic combination of self-attention mechanism and SimAM, the consistency between text and image has been greatly improved, which further enhances the expressiveness and robustness of the model in cross-modal generation tasks. This improvement not only provides new ideas and methods for the field of image generation, but also lays a solid foundation for the further development of related research in the future.

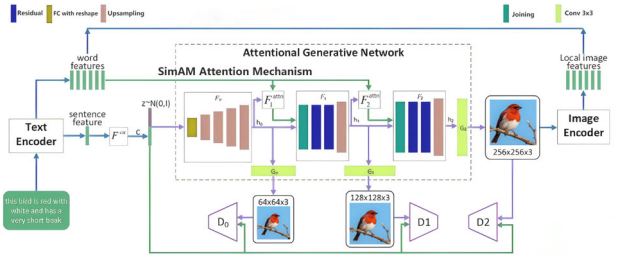

Fig. 4. Block diagram of the Improved AttnGAN model.

Fig. 4 illustrates the detailed block diagram of the proposed Improved AttnGAN model. The architecture comprises multiple generator and discriminator networks, each taking the previous network's hidden state and the word-context vectors as inputs. The generator utilizes an attention mechanism to focus on different sub-regions of the image by querying the most relevant word vectors. This multi-stage refinement process allows for higher resolution and more detailed image generation. The discriminator evaluates the generated images by distinguishing between real and fake samples, ensuring the authenticity and quality of the generated images. The block diagram clearly shows how the SimAM attention mechanism is integrated into the generator to enhance feature extraction and improve semantic consistency between text and images.

B. Association Modeling between Chinese Text and Images

We encode Chinese text via the network and extract its feature vector. This process includes the preprocessing steps such as word segmentation, part-of-speech tagging, and semantic role tagging, and deep feature extraction of text using convolutional neural network and recurrent neural networks [9]. The extracted feature vector not only contains the surface information of the text, but also contains the deep semantic relationship. The Improved AttnGAN model is used to combine the text feature vector with the image generation process. In this process, deep convolution is used to process each input channel independently to generate its own feature map. Pointwise convolution is used to integrate these feature maps to generate the final output feature map.

The AttnGAN model pioneered the development of an attention mechanism dedicated to the generator, which has the ability to transform diverse word vectors and sentence

vectors into realistic images. It can accurately capture and reflect the delicate information within text description [10]. However, this algorithm exhibits high complexity during implementation and necessitates a substantial number of parameters, posing challenges for the deployment and practical application of the model. To address this issue, while preserving the original accuracy and efficiency of the model, this paper proposes a reduction in the attention generation network part of the model. The aim is to effectively decrease the model's size and enhance its practicality. In this paper, the standard 3×3 convolution in the AttnGAN model is replaced with a combination of deep convolution and pointwise convolution. In the traditional convolution operation, the core of the convolution kernel directly acts on all the input channels, which leads to a high computational overhead. In contrast, deep convolution divides this computation into two stages. Firstly, each input channel is convoluted independently to generate its own feature map. Secondly, the point-by-point convolution is used to integrate these feature maps, and the final output feature map is generated through the point-to-point convolution operation. This decomposition strategy not only significantly reduces the number of model parameters, but also makes the model more efficient use of computing resources by decomposing the convolution process. Consequently, it achieves optimization and lightweight of the model structure while maintaining the model's performance.

IV. Result and Discussion

A. Experimental Set

This experiment is based on Ubuntu 20.04 system, equipped with GeForce RTX3090 (24G) GPU, and utilizes the Python 3.9 programming language along with the PyTorch deep learning framework, version 11.3 Cuda. The optimizer chosen is the Adam optimizer with parameters set as follows: the learning rates for both the generator and discriminator are specified, and the number of epochs is set to 1000. The model employs a batch size of 32 and includes a hyperparameter within the loss function.

In the field of text image generation, common datasets include the CUB-200 (CUB) dataset, the Oxford-102 dataset, and others such as the extended version of the CUB-200 dataset called Caltech-UCSD Birds-200-2011 (CUB-200-2011). The CUB dataset, which focuses on the fine-grained classification of birds, contains 200 different species with approximately 60 images per species, totaling 11,788 images. These images were captured in natural settings. A notable feature of the CUB dataset is its detailed annotation information, which includes a bounding box, a segmentation mask, and 15 attribute annotations for each image. This data is invaluable for studying tasks such as bird identification, fine-grained classification, and image segmentation. The Oxford-102 dataset, also known as the 102 Category Flower Dataset, is a dataset dedicated to flower identification. It comprises 102 different categories of flowers, with each category containing between 40 to 258 images, summing up to 7129 images in total. The Oxford-102 dataset also provides detailed annotation information, including bounding boxes and category labels for each image. This dataset is ideal for flower classification and recognition tasks, and is also suitable for image retrieval and fine-grained classification research. CUB bird data set and Oxford-102 flower data set are selected as experimental data sets in this paper.

B. Experimental Design

During the data preprocessing stage, we begin by standardizing images from both datasets to ensure uniformity in image size, facilitating further model processing. This involves resizing all images to 256×256 pixels. For Chinese text descriptions, we employ advanced natural language processing techniques for word segmentation, part-of-speech tagging, and semantic role labeling. These preprocessing steps are crucial for extracting the text's feature vector, establishing a solid foundation for subsequent model training. We also vectorize the text data using the Word2Vec model to generate word embeddings, which are then utilized as text feature vectors for input into the model. To enhance the model's generalization ability, we augment the image data through rotation, scaling, cropping, and color transformation.

During the model training phase, we feed preprocessed images and text into the Improved AttnGAN model for end-to-end training. Throughout this process, the generator and discriminator are updated in turn to minimize the adversarial loss function. We utilize the Adam optimizer with a learning rate of 0.0002, a batch size of 32, and conduct 1000 rounds of training. Throughout the training, we continuously monitor the loss of both generators and discriminators to ensure the model neither overfits nor underfits. Moreover, we implement an early stop strategy, which terminates training prematurely if the performance on the validation set does not show significant improvement over 10 consecutive epochs.

In this paper, the Inception Score (IS) and Frechet Inception Distance (FID) are utilized to quantify the experimental results. In the experiment, the Inception Score (IS) and Frechet Inception Distance (FID) are utilized to quantify the experimental results. For all the models, 10000 generated images are used for calculating these metrics. The IS is one of the key indicators for assessing the performance of a text-to-image generation model, capable of objectively evaluating the quality and diversity of the generated images. IS is primarily calculated using the pre-trained Inceptionv3 model to evaluate the Kullback-Leibler (KL) divergence between the marginal distribution and the conditional distribution. The calculation formula is as follows:

$$I = exp\left(E_x D_{KL}\left(p(y|x)||p(y)\right)\right) \tag{4}$$

where E represents the expectation and DKL signifies the KL divergence between the marginal and conditional distributions. There is a positive correlation between the IS score and the KL divergence. A higher IS score indicates a correspondingly higher KL divergence value.

FID is another important index to evaluate the performance of the model in the field of text-generated images. It aims to measure the similarity between two data sets, especially the similarity between the original image data distribution and the generated image data distribution. FID also uses the pre-trained Inception model to encode the features of the image, and then calculates the Frechet distance between them. The calculation formula is as follows:

$$FID = ||\mu_{x1} - \mu_{x2}||^2 + tr\left(\sum x_1 + \sum x_2 - 2(\sum x_1 \sum x_2)^{\frac{1}{2}}\right) \quad (5)$$

where: x_1 and x_2 are real image and generated image, μ_{x1} and μ_{x2} are that mean value of the respective eigenvectors, $\sum x_1 \; and \; \sum x_2$ are the covariance matrix of the eigenvectors. The lower the FID value, the closer the distribution of the real image and the generated image, the more similar the two images are, and the higher the quality of the generated image is.

We compare the performance of this model with the existing text-to-image generation methods, and analyze its advantages and disadvantages. Finally, we conduct a parameter sensitivity analysis to study the impact of different parameter settings on the performance of the model, including learning rate, batch size, and the number of training rounds.

C. Experimental Results and Analysis

In this paper, StackGAN, DM-GAN, MirrorGAN, DF-GAN and AttnGAN are selected as the comparison models. The comparison results of the Improved AttnGAN model with the other five text-to-image generation methods are shown in Table I and Table II. Table I shows the comparison results of Inception Score (IS) and Table II shows the comparison results of Frechet Inception Distance (FID).

TABLE I. COMPARISON RESULTS OF IS

Model	IS (CUB)	IS (Oxford-102)	Difference of IS (CUB)	Difference of IS Oxford-102)
StackGAN	3.7	3.2	1.32	1.11
DM-GAN	4.75	3.91	0.27	0.4
MirrorGAN	4.56	3.84	0.46	0.47
DF-GAN	4.76	3.8	0.26	0.51
AttnGAN	4.98	3.97	0.04	0.34
Improved AttnGAN	5.02	4.31	0	0

As evidenced by Table I, the IS of Improved AttnGAN is significantly higher than that of the other five algorithms. Higher IS values indicate better diversity and quality of the generated images. The IS value of Improved AttnGAN is 5.02, which is significantly higher than that of StackGAN (3.70) and DM-GAN (4.75), indicating that Improved AttnGAN has obvious advantages in the diversity and quality of the generated images. Compared with the best performing AttnGAN, the IS value of Improved AttnGAN is also slightly improved, from 4.98 to 5.02, which further proves the effectiveness of the improved model.

TABLE II. COMPARISON OF FID

Model	FID (CUB)	FID (Oxford-102)	Difference of FID(CUB)	Difference of FID (Oxford-102)
StackGAN	51.89	55.28	40.6	24.47
DM-GAN	16.09	43.92	4.8	13.11
MirrorGAN	18.34	43.29	7.05	12.48
DF-GAN	14.81	42.61	3.52	11.8
AttnGAN	12.15	37.82	0.86	7.01
Improved AttnGAN	11.29	30.81	0	0

As can be seen from Table II, the FID of Improved AttnGAN is significantly lower than that of the other five algorithms. The lower the FID value, the closer the distribution of the generated image and the real image is, and the higher the image quality is. Among all the comparison models, Improved AttnGAN has the lowest FID value of only 11.29, which is much lower than 51.89 of StackGAN and 14.81 of DF-GAN, showing the advantage of Improved AttnGAN in the realism of the generated images. Compared to AttnGAN, the FID value was further reduced from 12.15 to 11.29. This improvement, though small, means that in the field of image generation, even a small improvement means a significant improvement in the quality of the generated image. Combining the two indicators of IS and FID, Improved AttnGAN shows its superiority in terms of image quality, diversity and realism.

Fig. 5 presents a qualitative comparison with StackGAN, DM-GAN, MirrorGAN, DF-GAN, AttnGAN, and our Improved AttnGAN on the CUB-200, Oxford-102, and CUB-200-2011 datasets. DM-GAN and MirrorGAN demonstrate improvements in color and composition, yet still have room for improvement. DF-GAN and AttnGAN excel more in detail representation. Nevertheless, our Improved AttnGAN model outperforms others in terms of accuracy of details, color, and semantic consistency, vividly depicting the characteristics of birds and flowers, as well as the text-described scene.

We generated Fig. 6 using the model proposed in this paper based on the following prompt. A surrealist minimalist scene with traditional Chinese ink wash painting style. In the center of the image is a solitary figure dressed in red, walking on sandy ground. The figure is surrounded by several rocks, each encircled by concentric patterns reminiscent of ripples in water. (帮我生成一张图片：超现实主义的极简主义场景，带有中国传统水墨画风格。画面中央是一个身着红色衣服的孤独人物，走在沙地上。人物周围环绕着几块山石，每块岩石都被同心圆图案环绕，让人联想到水中的涟漪。背景柔和中性，纹理细腻，突出了深色岩石与人物，比例「9:16」). The background is soft and neutral in tone, with fine textures that highlight the dark rocks and the figure.

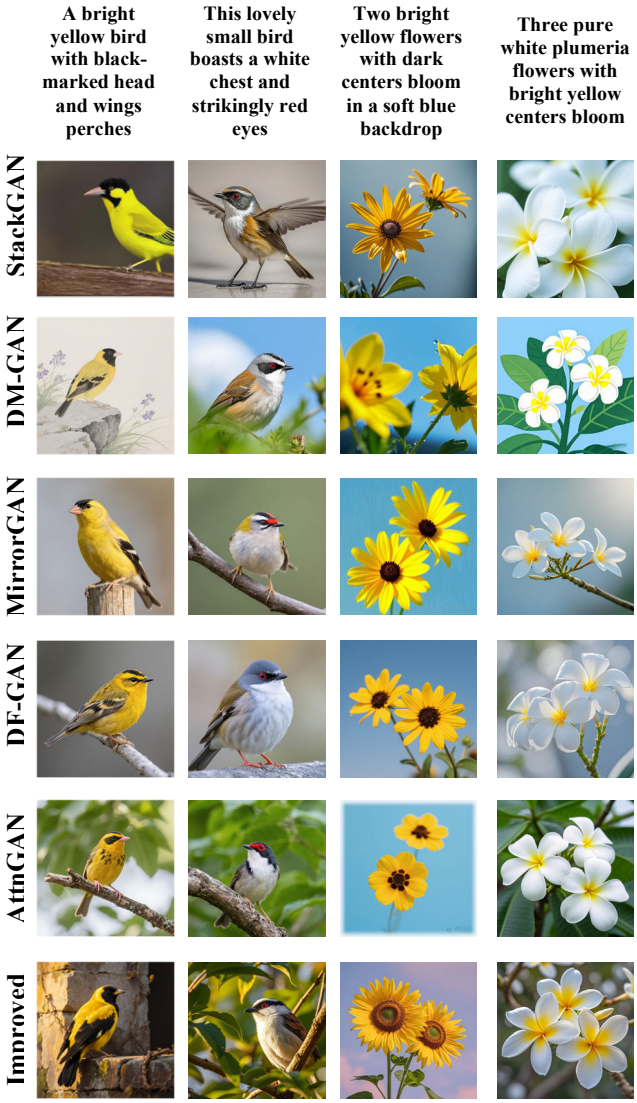

Fig. 5. Qualitative comparison with StackGAN, DM-GAN, MirrorGAN, DF-GAN, AttnGAN, and our Improved AttnGAN on CUB-200, Oxford-102 and CUB-200-2011 datasets.

(a) (b) (c)

(d) (e) (f)

Fig. 6. Surrealist minimalist scene generated by StackGAN (a), DM-GAN (b), MirrorGAN (c), DF-GAN (d), AttnGAN (e), and our Improved AttnGAN (f).

Fig. 6 has been inserted to illustrate the qualitative comparison of generated images from various models. The figure showcases sample images produced by StackGAN, DM-GAN, MirrorGAN, DF-GAN, AttnGAN, and our Improved AttnGAN. Through visual comparison, readers can clearly discern the differences in image quality, realism, and semantic consistency among the models, emphasizing the benefits of our Improved AttnGAN.

V. Conclusion

This paper presents an Improved AttnGAN model for controllable image generation from Chinese text. Our model integrates the SimAM attention mechanism, which enhances both image quality and semantic consistency without increasing model complexity. Through extensive experiments, we demonstrated that our model outperforms existing state-of-the-art methods in terms of image quality and realism.

In conclusion, our work makes several significant contributions to the field of controllable image generation. Firstly, the integration of the SimAM attention mechanism provides a novel approach to improving semantic consistency between text and images. Secondly, our model achieves superior performance with a simpler and more efficient architecture. Lastly, our detailed analysis of the model's architecture and its impact on cross-modal generation tasks offers valuable insights for future research.

Despite the promising results, there are several directions for future research. One potential area is the exploration of more advanced embedding techniques for Chinese text to better capture semantic nuances. Another direction is the application of attention mechanisms to the discriminator to further enhance model performance. Additionally, expanding the training dataset with more diverse and high-quality images and text descriptions could improve the model's generalization ability. Finally, exploring other attention variants or hybrid attention mechanisms could lead to further improvements in image generation quality.

Our approach has some limitations. The model's reliance on pre-trained embeddings may limit semantic depth. Future work could explore advanced embedding techniques. The current attention mechanism could be extended to the discriminator for further performance gains. Additionally, our model's performance is highly dependent on the training dataset, suggesting that larger and more diverse datasets could improve results.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Yongxia Hu identified the research question and designed the study's methodology.Yongxia Hu and Dong-Hyun Kim analyzed the data.Yongxia Hu conducted experiments, and contributed to the interpretation of the findings, and wrote the paper. All authors had approved the final version.

REFERENCES

[1] H. Zhang, H. Song, S. Li *et al.*, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–37, Sept. 2023.

[2] S. Zhao, D. Chen, Y. C. Chen *et al.*, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] M. Ding, Z. Yang, W. Hong *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19822–19835, 2021.

[4] Z. Sha, Z. Li, N. Yu *et al.*, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proc. 2023 ACM SIGSAC Conf. on Computer and Communications Security*, 2023, pp. 3418–3432.

[5] H. Lu, R. Yang, Z. Deng *et al.*, "Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 17, no. 1s, pp. 1–18, 2021.

[6] X. Lv, "Chinese description generation of dual attention images based on multi-modal fusion," in *Proc. J. Phys.: Conf. Ser.*, vol. 1735, no. 1, 2021, 012004.

[7] Q. Lyu, N. Zhao, Y. Yang *et al.*, "A diffusion probabilistic model for traditional Chinese landscape painting super-resolution," *Heritage Science*, vol. 12, no. 1, 2024.

[8] M. Yalin, L. Li, J. Yichun *et al.*, "Research on denoising method of Chinese ancient character image based on Chinese character writing standard model," *Sci. Rep.*, vol. 12, no. 1, 19795, 2022.

[9] M. Ding, W. Zheng, W. Hong *et al.*, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16890–16902, 2022.

[10] Y. Pan, L. Wang, S. Duan *et al.*, "Chinese image caption of Inceptionv4 and double-layer GRUs based on attention mechanism," in *Proc. J. Phys.: Conf. Ser.*, vol. 1861, no. 1, 2021, 012044.