Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4973535, 12 pages
https://doi.org/10.1155/2022/4973535

Research Article

Self-Attention-Based Edge Computing Model for Synthesis Image to Text through Next-Generation AI Mechanism

Hamdan Ali Alshehri,[1] **N. Junath**,[2] **Poonam Panwar**,[3] **Kirti Shukla**,[4] **Saima Ahmed Rahin**,[5] **and R. John Martin**[6]

[1] *Faculty of Computer Science and Information Technology, Jazan University, Jizan, Saudi Arabia*
[2] *Information Technology, University of Technology and Applied Science, Ibri, Oman*
[3] *Chitkara University Institute of Engineering and Technology, Chitkara University, Chandigarh, Punjab, India*
[4] *Galgotias University, Noida, India*
[5] *United International University, Dhaka, Bangladesh*
[6] *Faculty of Computer Science and Information Technology, Jazan University, Jizan, Saudi Arabia*

Correspondence should be addressed to Saima Ahmed Rahin; srahin213012@mscse.uiu.ac.bd

Received 5 April 2022; Revised 8 May 2022; Accepted 14 May 2022; Published 10 June 2022

Academic Editor: Vijay Kumar

Image synthesis based on natural language description has become a research hotspot in edge computing in artificial intelligence. With the help of generative adversarial edge computing networks, the field has made great strides in high-resolution image synthesis. However, there are still some defects in the authenticity of synthetic single-target images. For example, there will be abnormal situations such as “multiple heads” and “multiple mouths” when synthesizing bird graphics. Aiming at such problems, a text generation single-target model SA-AttnGAN based on a self-attention mechanism is proposed. SA-AttnGAN (Attentional Generative Adversarial Network) refines text features into word features and sentence features to improve the semantic alignment of text and images; in the initialization stage of AttnGAN, the self-attention mechanism is used to improve the stability of the text-generated image model; the multistage GAN network is used to superimpose, finally synthesizing high-resolution images. Experimental data show that SA-AttnGAN outperforms other comparable models in terms of Inception Score and Frechet Inception Distance; synthetic image analysis shows that this model can learn background and colour information and correctly capture bird heads and mouths. The structural information of other components is improved, and the AttnGAN model generates incorrect images such as “multiple heads” and “multiple mouths.” Furthermore, SA-AttnGAN is successfully applied to description-based clothing image synthesis with good generalization ability.

1. Introduction

Image synthesis based on text description (text to image, t2i) covers technologies such as computer vision and natural language processing and is an interdisciplinary and cross-modal comprehensive task [1]. Based on the input natural language description, the model should synthesize images consistent with the description content and have complete semantic information. This task requires the computer to understand the semantic information of the text and convert the semantic information into pixels to generate a high-resolution and high-fidelity image, which is a very challenging task. It has a wide range of application potential and can be used in computer-aided design, criminal investigation portrait generation, etc.

The rapid development of deep learning has brought significant advances in computer vision and natural language processing in theory and technology and promoted the task of text-based image synthesis to move towards high resolution, high authenticity, and high controllability. Ref. [2] used generative adversarial networks (GANs) [3] to extract sentence features of textual descriptions by using a character-level recurrent neural network, along with noise as input to a cGAN network [3]. To reduce the difficulty of high-resolution image synthesis based on GANs, Ref. [4] proposed StackGAN, which consists of two generative adversarial networks: the first stage generates low scores, and the second stage refines low-resolution images and gradually synthesizes high-resolution photos. To improve the quality of synthetic images, Xue et al. proposed StackGAN++ [5]. In addition to using multiple GANs to generate multiscale ideas, they added a regularization setting for colour consistency to the loss, which can keep the images of different scales during training. Thickness, reducing the instability of GANs training. Xu et al. introduced a global attention mechanism [6] in AttnGAN [7]. They proposed a severe attention multimodal similarity model, which used word-level and sentence-level text features as input to improve the matching between text and images.

However, the GAN-INT-CLS, StackGAN, and StackGAN++ are some methods in adversarial network that are employed to achieve high artistic graphics. They only use sentence-level features as text features and lose important synthetic image details in performing the image synthesis techniques. Level features are embedded as text, which improves semantic alignment. In addition, although the AttnGAN network uses a global attention mechanism for text images to increase the details of the generated images, it often generates birds that do not conform to natural laws, such as "two heads" and "two eyes" and other wrong photos. Generating nonsemantic bird images for AttnGAN, a GAN-based t2i network model is proposed, which uses a self-attention mechanism in the initial stage of the model better to learn important spatial and positional information in the image when synthesizing low-resolution ideas and improve the accuracy of image generation in the initial step, thereby improving the correctness of high-resolution image synthesis. When synthesizing bird graphics, there will be unusual scenarios such as "many heads" and "multiple mouths." A text synthesis single-target model SA-AttnGAN based on a self-attention strategy is established to solve such concerns. In this study, we aim at improving the semantic alignment of text and images on basis of SA-AttnGAN (attentional generative adversarial network) that is proposed to refine text features into word features and sentence features.

The main objective of this article lies in the following two points:

(1) Based on the AttnGAN model, this article proposes to add a self-attention module in the initial stage to improve the original model to generate bird pictures that do not conform to the norm and optimize the IS and FID index scores in the CUB [8] dataset. The actual generation effect shows that the SA-AttnGAN network model proposed in this article can generate realistic and natural bird pictures.

(2) A text-generated image clothing dataset is also produced, expanding the application field of t2i technology for other researchers and laying a data foundation.

1.1. Organization. This work is organized into various modules where Module 1 discusses the introduction followed by Section 2, which states the related work in this field. Section 3 elucidates the proposed model followed by the "Result and Analysis" section that is stated in Section 4. The last section discusses the conclusion of the work.

2. Related Work

Image synthesis for early text descriptions mainly combined retrieval and supervised learning [9]. First, the information and "imageable" text units are determined by the correlation between keywords (or key phrases) and the image; then, based on the current test conditions, the text unit retrieves the regions most likely to be related to the image content and finally optimized to an image layout to associate the textual description with the image content. However, due to limited training methods, this method can only change the characteristics of specific images and cannot synthesize ideas with entirely new content based on textual descriptions. With the deepening of research. Each image is modelled as a combined foreground and background Attribute2Image [10] method. Attribute2Image learns from given attributes to generate ideas containing different characteristics, such as gender, hair colour, and age. Although the above techniques can synthesize relatively realistic images, they are still limited by limited descriptive properties. With the development of multimodal learning, a batch of image synthesis models based on generative adversarial networks and deep convolutional decoders have emerged [11, 12]. Generative adversarial networks (GANs) proposed by Ref. [13] mainly consist of a discriminator and a generator. The generator tries to generate synthetic images and thus "trick" the discriminator; the discriminator tries to distinguish between authentic images and synthetic images. Based on such characteristics, GANs can be used in the field of image synthesis based on the text description, and the purpose of adversarial training is defined as image synthesis based on text description: through the continuous "generation" and "discrimination" of raw images and "fake images," and the relationship between image content and text description is gradually improved, and finally, the purpose of describing synthetic images based on text is achieved. However, there may be certain negatives to picture-generating technology, such as overcrowding, artificial visual mismatch difficulties, and uniformity or normalization of scanned images, but these concerns are rare.

Tang et al. pioneered deep convolution-based GANs (DC-GANs) [14] for text-image synthesis [15]. DC-GANs use a character-level recurrent neural network to extract sentence feature vectors from text descriptions and use them with noise as input to GAN [16]. StackGAN [17] focuses on improving synthetic images' quality and increasing the resolution of synthetic images from 64 × 64 to 256 × 256 through two GANs based on word feature vectors. As a further expansion, StackGAN++ [18] improved StackGAN into an end-to-end network, which reduced the instability of GANs training and increased the colour loss function and improved the colour expression of the synthesized image. It is a multistage generative adversarial network, which is based on edge computing features and mimics the model for limiting the latency incurred while processing the image synthesis of data in multigenerative network and works on a real-time processing model. Edge computing necessitates greater storage space. Because of the large amount of datasets for images synthesis process, it also poses a significant security risk and necessitates sophisticated infrastructure, which limits the usage of employing this model. Given the successful application of attention mechanisms in various fields of deep learning, AttnGAN [19] first introduced a global attention mechanism into the area of text synthesis images. It uses a text encoder to extract text feature vectors at the sentence and word levels, calculate their similarity to global image features and local image features, and improve the correlation between synthetic images and description texts through the proposed DAMSM pretraining method. *Clarity.* With the deepening of research, image synthesis based on text description has achieved unprecedented high resolution, multiobjective and controllability: HD-GAN [20] uses cascaded network results to increase the resolution to 512 × 512. For photo-realistic image creation through semantic layouts, the author has suggested a novel Edge-assisted generative adversarial network [21]. We provide us with an edge-preserving MRI image reconstruction technique on intermittent multiscale feature engineering and a (EP IMF-GAN)-generative adversarial network [22]. Despite improvements, the resolution of image sequences seems far from optimal due to largely unaddressed obstacles. Obj-GAN [23] can synthesize multitarget images with a complex layout, gradually generate from design, shape to content, and improve the model collapse problem in complex image synthesis. To solve the problem of overall composition reset caused by the change of specific text attributes (colour, target), ControlGAN [24], based on AttnGAN structure, proposed to use a channel and spatial attention mechanism to increase word-image region feature matching and perceptual loss and other constraints.

3. Network Model (SA-AttnGAN)

Like AttnGAN, the SA-AttnGAN network structure is proposed in this article is divided into pretraining and a multistage generative adversarial network. The DAMSM module [25] in AttnGAN is introduced in the pretraining network. This module contains a text encoder and an image encoder to extract features and calculate the DAMSM loss as part of the generator loss function. At the same time, the generative adversarial network consists of three pairs of generators and discriminators are composed to process images of 64 × 64, 128 × 128, and 256 × 256 stages, respectively.

Like most GAN-based t2i models, the self-attention-based text generative image network (SA-AttnGAN) proposed in this article adopts a multistage high-resolution image synthesis strategy (Figure 1). The generators G_0, G_1, and G_2 synthesize images with resolutions of 64 × 64, 128 × 128, and 256 × 256, respectively.

In the G_0 stage, the conditional enhancement module F^{ca} is used first; the CA [26] module in Figure 1 is used to process the sentence feature vector $\bar{e}$ to obtain a low-dimensional text conditional vector. It is then concatenated with the noise vector $z \in \mathbb{R}100$ as an input containing multiple upsampling blocks F0, as shown in the following equation:

$$h_{0'} = F_0\left(z, F^{ca}\left(e^-\right)\right), \tag{1}$$

where h_0' represents the hidden node, which contains the image information generated in the initial stage.

Unlike AttnGAN, this article introduces a self-attention mechanism [27], which assigns additional weight information through autonomous learning between image feature maps. As a result, the final feature map contains more spatial and positional information.

As shown in Figure 2, first transform h_0' into feature spaces f and g, where W_f and W_g are perception layers, as shown in the following equations:

$$f\left(h_{0'}\right) = W_f\left(h_{0'}\right), \tag{2}$$

$$g\left(h_{0'}\right) = W_g\left(h_{0'}\right). \tag{3}$$

And calculate the weight information $\beta_{j,i}$ and the calculation formula is shown in the following formula:

$$\beta_{i,j} = \frac{\exp\left(s_{j,i}\right)}{\sum_{i=1}^{n} \exp\left(s_{j,i}\right)}, \tag{4}$$

where $s_j, i = f\left(h_{0'}\right)^T g\left(h_{0'}\right)$. $\beta_{j,i}$ represents the weight information of the i-th position when synthesizing the j-th area of the image. It learns the space and position information in the feature map through the self-supervision mechanism and assigns a greater weight value to the important detailed information in the image. It is beneficial to generate more meaningful images in the initial stage. Then, convert h_0' to the third feature space u, as shown in the following equation:

$$u\left(h_{0'}\right) = W_u h_{0'}, \tag{5}$$

where W_u is the perception layer of the feature space u, which is used to change the dimension size of the feature. Then, multiply the weight map $\beta_{j,i}$ and $u\left(h_0'\right)$ to get the image feature matrix m_j with attention mechanism, as shown in the following equation:

$$m_j = \sum_{i=1}^{N} \beta_{j,i} u\left(h_{0}{}'_{i}\right). \tag{6}$$

Finally, use conv_1 × 1 to convert the obtained image feature matrix m_{j} to the feature space v, as shown in the following equation, so that the received image feature size is the same as the input image feature size:

$$v\left(m_j\right) = W_v m_j. \tag{7}$$

Using h_0 to represent the output result of F^{sa} (i.e., the SA module in Figure 1), by using the attention mechanism, the generated image in the initial stage will contain more meaningful position and spatial information, as shown in the following equation:

$$h_0 = F_{sa}\left(h_{0'}\right). \tag{8}$$

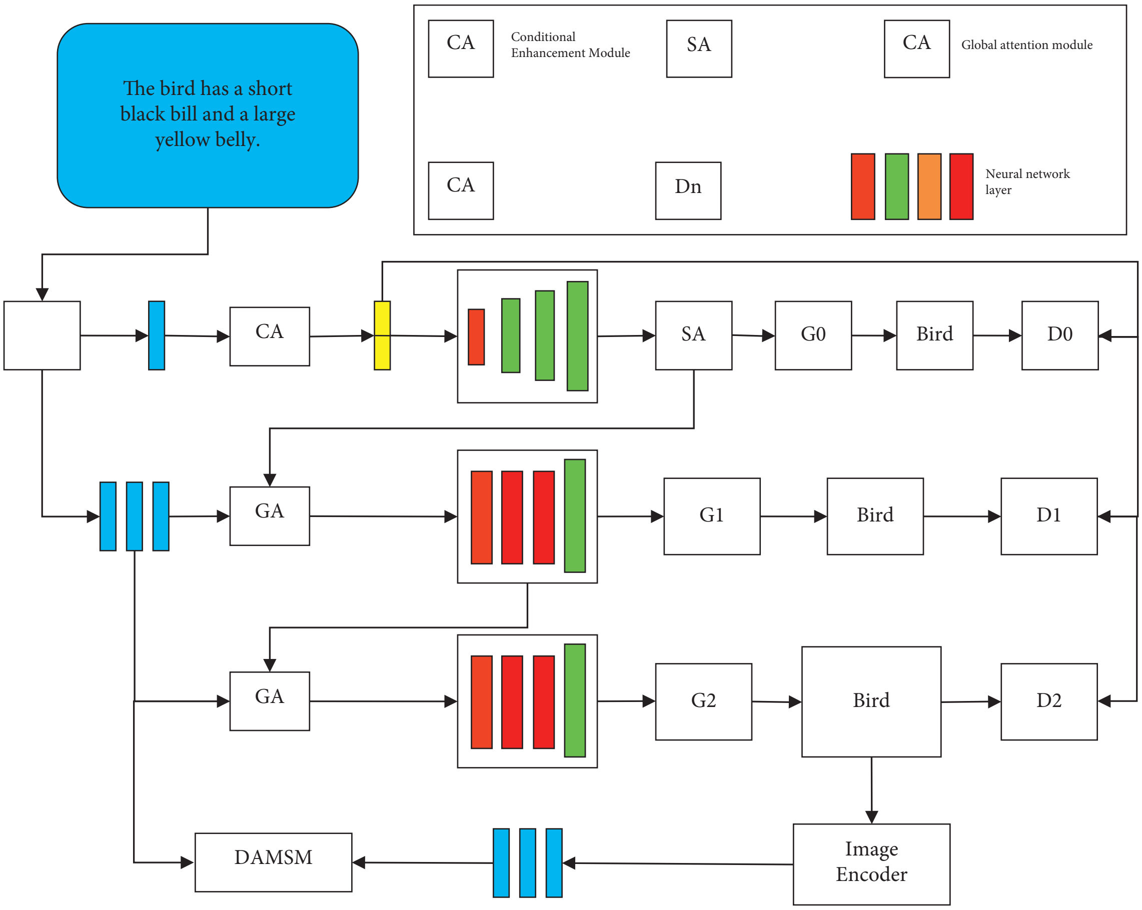

Figure 1: SA-AttnGAN network structure diagram.

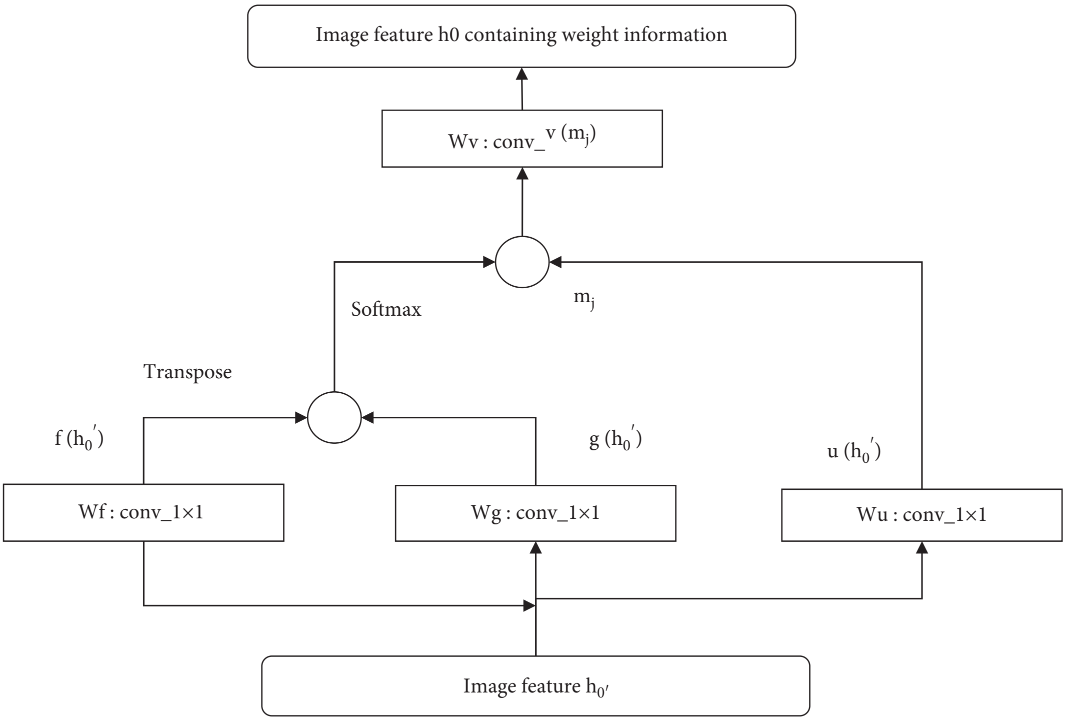

Figure 2: Self-attention mechanism (SA).

In the G_1 and G_2 stages, the hidden nodes h_i of different stages are used as input, and $\hat{x}_i$ is used to represent the generated images of different resolutions, as shown in the following equations:

$$h_i = F_i\left(h_{i-1}, F_i^{GA}(e, h_i)\right), \tag{9}$$

$$\widehat{x}_i = G_i(h_i), \quad i = 1, 2, \tag{10}$$

where F_i^{GA} is the i-th stage global attention generation module [28] and F_i is the i-th stage segments containing neural network layers such as upsampling blocks.

Among them, the generator loss function is defined as

$$\mathrm{L} = \mathrm{L}_G + \lambda \mathrm{L}_{\mathrm{DAMSM}}, \tag{11}$$

in

$$\left\{\frac{\mathrm{L}_G = \sum_{i=0}^{m-1} \mathrm{L}_{Gi}}{\mathrm{L}_{Gi} = -(1/2)\mathbb{E}_{\widehat{x}_i \sim G_i}\left[\ln D_i(\widehat{x}_i)\right] - (1/2)\mathbb{E}_{\widehat{x}_i \sim G_i}\left[\ln D_i(\widehat{x}_i, \overline{e})\right]}\right., \tag{12}$$

where $\mathrm{L}_{\mathrm{DAMSM}}$ is the loss function derived using a pretrained network, and λ is a hyper parameter that determines how much the DAMSM module affects the generator loss function [29].

As shown in Figure 1, the D_0, D_1, D_2 multidiscriminators used in this article are calculated in parallel, and the input image sizes are 64×64, 128×128, and 256×256, respectively. The discriminator D_{i} consists of two parts $D_i(D_i^1, D_i^2)$, where $i = 0,1,2$, each part contains different discriminative content, D_i^1 discriminates the authenticity of the image, and D_i^2 discriminates the semantic consistency between the image and the text. The definition is shown in the following formula:

$$\left\{\frac{\mathrm{L}_D = \sum_{i=0}^{m-1} \mathrm{L}_{D_i}}{\mathrm{L}_{D_i} = \mathrm{L}_1^{\mathrm{uncondition}} + \mathrm{L}_2^{\mathrm{condition}}}\right., \tag{13}$$

where $\mathrm{L}_1^{\mathrm{uncondition}}$ is used to identify whether the input image is real and $\mathrm{L}_2^{\mathrm{condition}}$ is used to identify whether the input image is related to text. The calculation formulas are shown in the following equations:

$$\begin{aligned} \mathrm{L}_1^{\mathrm{uncondition}} &= -\frac{1}{2}\mathbb{E}_{x_i \sim P_{\mathrm{datai}}}\left[\ln D_i(x_i)\right] \\ &\quad - \frac{1}{2}\mathbb{E}_{\widehat{x}_i \sim P_{G_i}}\left[\ln\left(1 - D_i(\widehat{x}_i)\right)\right], \end{aligned} \tag{14}$$

$$\begin{aligned} \mathrm{L}_2^{\mathrm{condition}} &= -\frac{1}{2}\mathbb{E}_{x_i \sim P_{\mathrm{datai}}}\left[\ln D_i(x_i, \overline{e})\right] \\ &\quad - \frac{1}{2}\mathbb{E}_{\widehat{x}_i \sim P_{G_i}}\left[\ln\left(1 - D_i(\widehat{x}_i, \overline{e})\right)\right]. \end{aligned} \tag{15}$$

4. Experiment and Result Analysis

This article chooses AttnGAN as a comparison model. AttnGAN uses an attention mechanism in text generation images and uses sentence-level and word-level text features as input to improve the clarity of synthesized images.

A Bi-LSTM [30] with a layer number of 1 is adopted in the text encoder, the word embedding size is 300, and the dimension of text features is 256. The inception-v3 [31] network is used in the image encoder to extract image features. The global image feature dimension is 2 048, the local image region feature contains 768 channels, and each channel dimension is 289, similar to the AttnGAN [32] network. The parameters of each track are the same. Adam [33] is used as the optimizer in the training phase, and the learning rate is set to 0.0002. In the network loss function, the hyperparameter λ is set to 5. Batch_size is set to 10.

4.1. Dataset. In this article, the CUB dataset is selected for training the model. CUB is a public dataset in the t2i field produced by the University of Cambridge [34], which contains more than 10,000 pictures of more than 200 species of birds. Among them, 8855 photos are used for training, and 2933 are used for testing. Ten text descriptions accompany each image. Its report covers more than ten bird's heads, beak, breast, and crest attributes.

4.2. Evaluation Parameter. To ensure the comparability of experimental results, this article selects Inception Score [35] (IS) and Frechet Inception Distance (FID) [36] for comparison. IS indicator is especially proposed by StackGAN for CUB with a complete set of evaluation algorithms (https://github.com/hanzhanggit/StackGAN inception-model) and is widely used in other t2i works. The greater the IS index, the finer the produced image, the greater the variety, and the improved the model reliability. Another frequently utilized assessment index is FID. It computes the real samples and creates the higher dimensional space difference across them. Based on the AttnGAN model, this article proposes to add a self-attention module in the initial stage to improve the original model to generate bird pictures that do not conform to the norm and optimize the IS and FID index scores in the CUB dataset [37, 38].

The algorithm principle is as follows:

$$\mathrm{IS} = \exp\left(E_x D_{KL} t\left(p(y|x) \mid p(y)\right)\right), \tag{16}$$

where x represents the generated sample and y represents the label predicted by the algorithm, and by calculating the Kullback–Leibler divergence of $p(y|x)$ and $p(y)$ distribution, the larger the value, the better the model generation result. The higher the IS index, the clearer the generated image, the higher the diversity, and the better the model stability. FID is another commonly used evaluation index. It calculates the actual samples and generates the distance between them in the feature space. The algorithm principle is as follows:

$$\mathrm{FID} = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{tr}\left(\sum_x + \sum_g 2\left(\sum_x \sum_g\right)^{1/2}\right), \tag{17}$$

where μ_r represents the mean of the actual image features, μ_g represents the mean of the generated image features, Σ_x represents the covariance matrix of the primary image features, $\sum_g$ represents the covariance matrix of the

Table 1: Comparison of methods on the CUB dataset.

Method	IS	FID
GAN-INT-CLS [2]	2.88 ± 0.04	68.79
GAWWN [16]	3.62 ± 0.07	—
StackGAN [5]	3.70 ± 0.04	51.89
StackGAN-v2 [6]	3.84 ± 0.06	30.30
HDGAN [10]	4.15 ± 0.05	—
AttnGAN [7]	4.36 ± 0.03	15.38
Proposed approach	4.52 ± 0.03	14.25

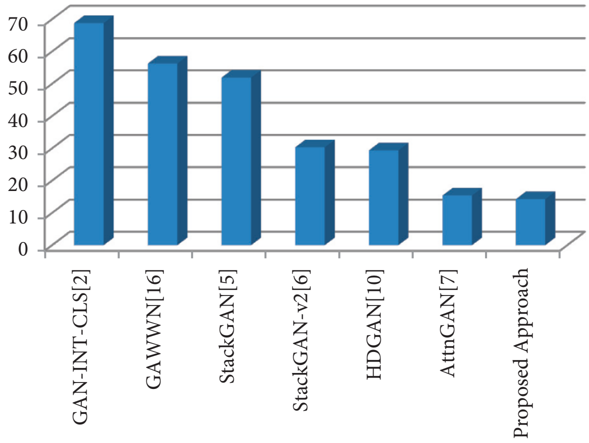

Figure 3: Comparison of methods on the CUB dataset for FID.

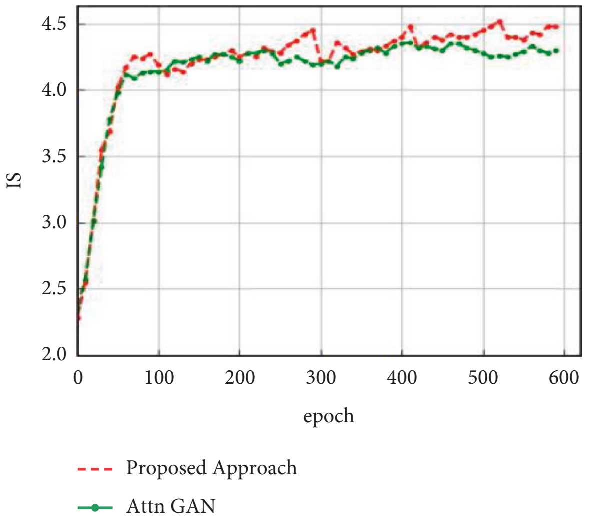

Figure 4: Changes of IS indicators under different epochs.

developed image features, and tr represents the trace of the matrix. The lower the FID value, the better the quality and diversity of the image.

4.3. Result Analysis. The SA-AttnGAN model proposed in this article is trained for 600 epochs on an RTX TITAN V graphics card with 11 GB of video memory. About 30 000 test set photos are generated for index comparison. The results are shown in Table 1 and Figures 3–5. As shown in Table 1, compared with many other representative methods, the model in this article has the highest IS index value, achieving a score of 4.52 ± 0.03, FID.

Table 2: Comparison of changes in different hyperparameter indicators.

Method	IS	FID
SA-AttnGAN, $\lambda = 0.1$	4.05 ± 0.05	30.25
SA-AttnGAN, $\lambda = 1$	4.34 ± 0.02	22.60
SA-AttnGAN, $\lambda = 5$	4.52 ± 0.03	14.25
SA-AttnGAN, $\lambda = 10$	4.24 ± 0.05	25.56

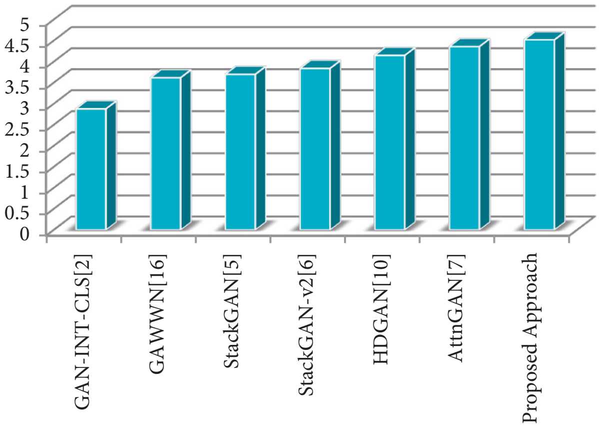

Figure 5: Comparison of methods on the CUB dataset for IS.

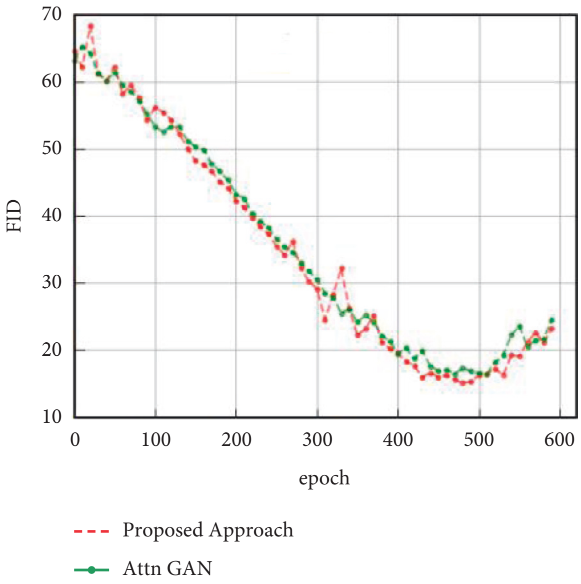

Figure 6: Changes in FID indicators at different epochs.

The indicator was the lowest, with a score of 14.25. Compared with AttnGAN, the IS index is improved by 0.16, and the FID index is reduced by 0.13.

Figure 4 shows the IS index changes for 600 epochs. The abscissa is the number of epoch iterations, and the ordinate is the IS index value. After 450 generations, the index value of our method is better than that of AttnGAN.

Figure 6 shows the change in the FID index for 600 epochs. The abscissa is the number of epoch iterations, and

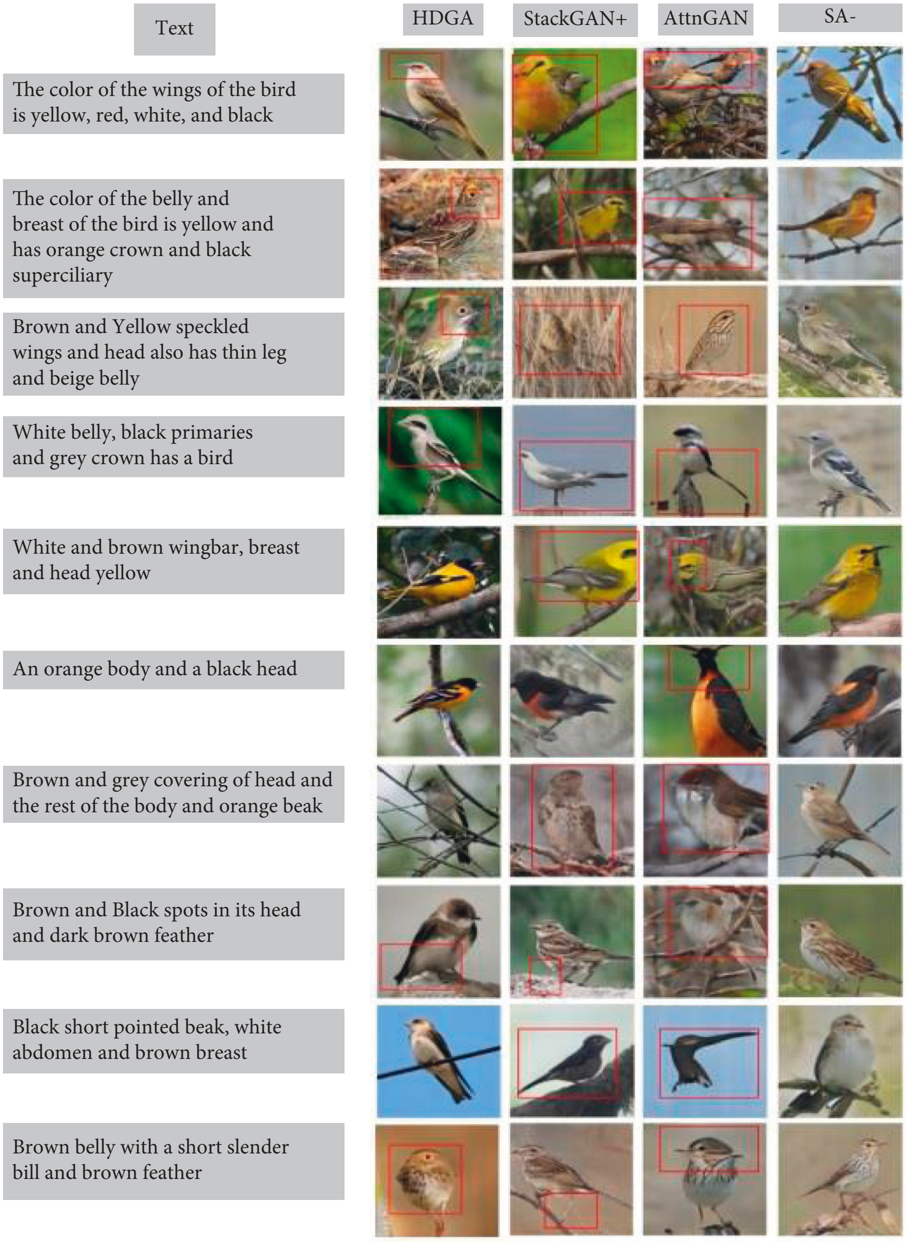
Text
HDGA
StackGAN+
AttnGAN
SA-
The color of the wings of the bird is yellow, red, white, and black
The color of the belly and breast of the bird is yellow and has orange crown and black superciliary
Brown and Yellow speckled wings and head also has thin leg and beige belly
White belly, black primaries and grey crown has a bird
White and brown wingbar, breast and head yellow
An orange body and a black head
Brown and grey covering of head and the rest of the body and orange beak
Brown and Black spots in its head and dark brown feather
Black short pointed beak, white abdomen and brown breast
Brown belly with a short slender bill and brown feather

Figure 7: Part of the testing effect.

the ordinate is the FID index value. After 380 generations, the FID value of the method in this article is better than that of AttnGAN.

The above chart shows that this article uses the self-attention mechanism in the initial stage to generate the weight mask map by learning the feature information between images independently. The feature map finally developed in the initial stage integrates more space and position information in the model to generate structure. It can further improve the effect of high-resolution image synthesis and enhance the clarity and diversity of image synthesis. In addition, sentence-level and word-level text features are used simultaneously to extract more textual information and improve the semantic consistency between text and images.

As shown in Table 2, the influence of the hyperparameter λ on the two metrics with different values is also calculated. λ is the DAMSM network module on the overall network. After the values are 0.1, 1, 5, and 10, when $\lambda = 5$, the index effect is the best.

4.4. Comparison of Synthetic Effects. Figure 5 compares the experimental results between SA-AttnGAN and many usual methods. Among them, the HDGAN [10], StackGAN++ [6], and AttnGAN [7] ways use the officially implemented model to conduct experiments and test 2 933 test set texts in the same experimental environment. The HDGAN model is inspired by StackGAN [5], proposes an end-to-end model, and introduces adversarial hierarchical

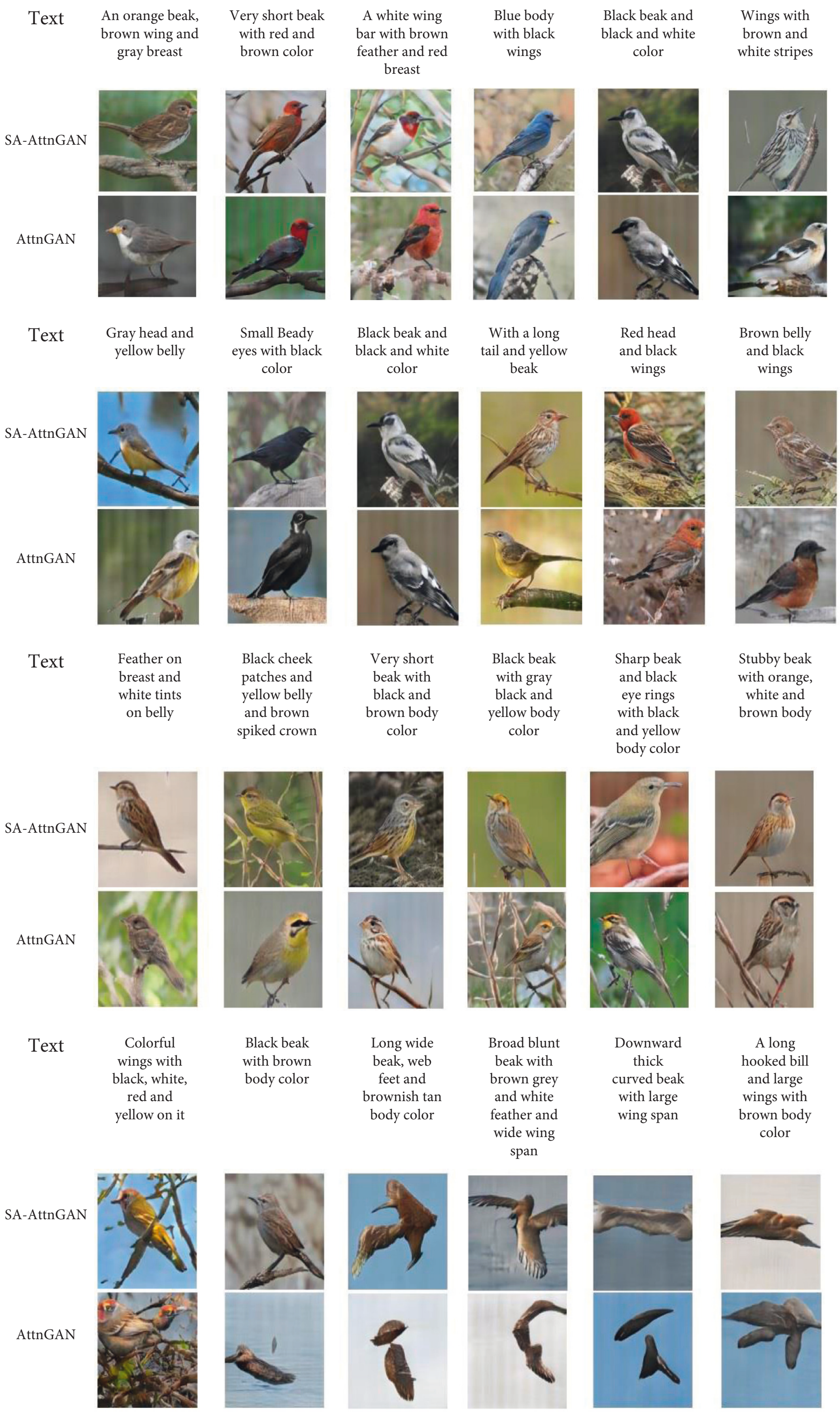
Text
An orange beak, brown wing and gray breast
Very short beak with red and brown color
A white wing bar with brown feather and red breast
Blue body with black wings
Black beak and black and white color
Wings with brown and white stripes
SA-AttnGAN
AttnGAN
Text
Gray head and yellow belly
Small Beady eyes with black color
Black beak and black and white color
With a long tail and yellow beak
Red head and black wings
Brown belly and black wings
SA-AttnGAN
AttnGAN
Text
Feather on breast and white tints on belly
Black cheek patches and yellow belly and brown spiked crown
Very short beak with black and brown body color
Black beak with gray black and yellow body color
Sharp beak and black eye rings with black and yellow body color
Stubby beak with orange, white and brown body
SA-AttnGAN
AttnGAN
Text
Colorful wings with black, white, red and yellow on it
Black beak with brown body color
Long wide beak, web feet and brownish tan body color
Broad blunt beak with brown grey and white feather and wide wing span
Downward thick curved beak with large wing span
A long hooked bill and large wings with brown body color
SA-AttnGAN
AttnGAN

Figure 8: Ablation experiment with a self-attention mechanism.

adversarial objectives, focusing on improving the resolution of image generation. Still, it does not pay attention to the structural information of the generated images. As a result, some of the attributes of HDGAN are caused unnaturally. For example, in the third group of experiments, the proportion of bird eyes generated by HDGAN is inconsistent. The StackGAN++ [6] model is based on the StackGAN [5] model, changing it to an end-to-end model, adding colour regularization loss, focusing on improving the colour consistency of the multistage generated method cannot also create images. As a result, the learning of spatial and location information, such as the third group of

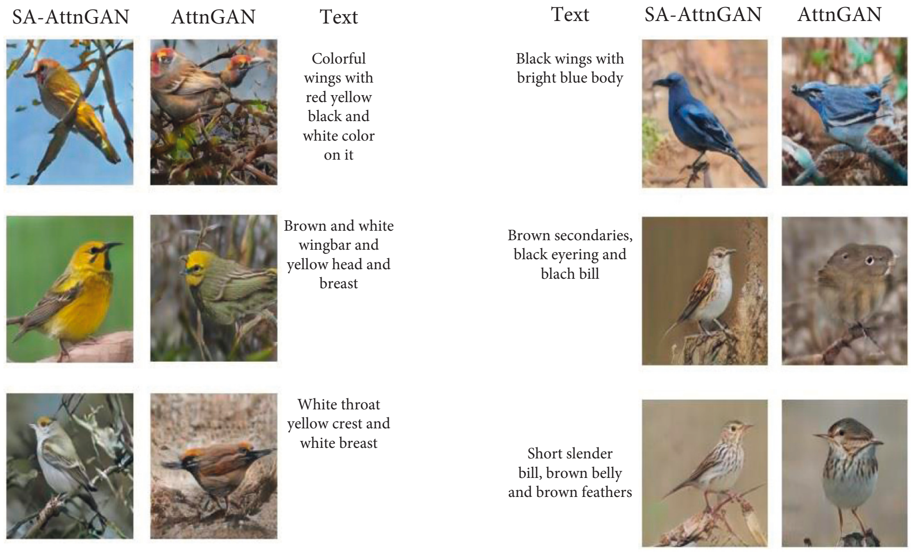
SA-AttnGAN
AttnGAN
Text
Colorful wings with red yellow black and white color on it
Brown and white wingbar and yellow head and breast
White throat yellow crest and white breast
Text
SA-AttnGAN
AttnGAN
Black wings with bright blue body
Brown secondaries, black eyering and blach bill
Short slender bill, brown belly and brown feathers

Figure 9: Comparison of ablation experiments in situations such as “multiple heads” and “multiple mouths.”

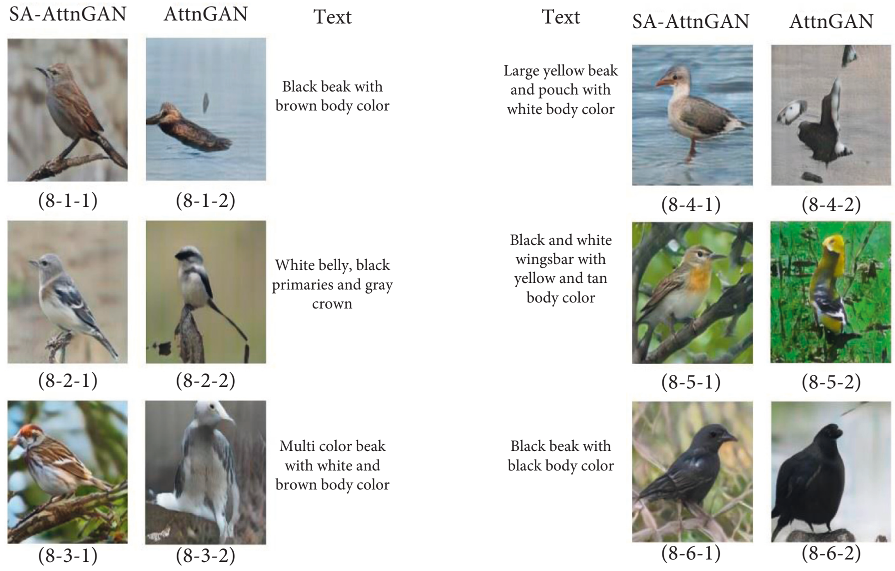
SA-AttnGAN
AttnGAN
Text
Black beak with brown body color
(8-1-1)
(8-1-2)
White belly, black primaries and gray crown
(8-2-1)
(8-2-2)
Multi color beak with white and brown body color
(8-3-1)
(8-3-2)
Text
SA-AttnGAN
AttnGAN
Large yellow beak and pouch with white body color
(8-4-1)
(8-4-2)
Black and white wingsbar with yellow and tan body color
(8-5-1)
(8-5-2)
Black beak with black body color
(8-6-1)
(8-6-2)

Figure 10: Comparison of ablation experiments for the overall generation failure of birds.

experiments, the rendered birds are integrated with the background, and the overall generation fails. AttnGAN [7] uses an attention mechanism and uses sentence-level and word-level text features to enhance the semantic alignment of text images. Still, the weights of essential attributes of birds cannot be learned well, and they pay too much attention to some details. This is because, such as the tenth group of experiments, two mouths were generated, while the third group of experiments lacked detailed information such as mouths. The SA-AttnGAN method adds a self-attention mechanism module in the initial stage so that the model can learn the correct attribute weight distribution. For example, in the third group of experiments, the generated birds are complete and natural, indicating that this method improves the text generation of single-target images—visual quality.

Figure 6 shows the ablation experiment adding the self-attention mechanism module. SA-AttnGAN indicates that the self-attention mechanism is used, and AttnGAN means that the self-attention tool is not used. Figure 6 is divided into four groups, each with six groups of comparative experiments. The first three groups show that both SA-AttnGAN and AttnGAN methods synthesize realistic and natural bird pictures. For example, the renderings of the first sentence of the third group of text synthesis conform to text semantics, including details such as “brown bird” and “white belly.” *Information.* The fourth set of experiments shows partially generated images. For example, the second sentence of the fourth group of text AttnGAN synthesized a “multiheaded” bird, the second sentence AttnGAN failed to create birds, and SA-AttnGAN synthesized the correct bird photos. The experimental analysis will focus on these two parts later. *Illustrate*. In addition, the fourth group of experiments also showed that the SA-AttnGAN and AttnGAN methods failed to generate some images, such as the third, fourth, fifth, and sixth sentences of the fourth group of experiments. The analysis is that "Large bird" text descriptions such as "Large wings" will synthesize photos of flying birds. Still, since there are few photos of soaring postures in the dataset, the model does not thoroughly learn the distribution of such images, ultimately affecting image generation results.

4.4.1. Analysis of Inappropriate Images such as "Long Heads" and "Multiple Mouths". Figure 7 shows some well-generated synthetic photos of birds, but the AttnGAN method will also synthesize inappropriate images during the test. Figure 8 shows six sets of AttnGAN models and high-resolution images generated by the model in this article. It can be found that AttnGAN often causes bird pictures that are not normal, such as "multiple heads," "multiple mouths," and "multiple eyes." For example, a bird head, 7-2-2, 7-6-2 generate two beaks, and 7-5-2 causes multiple eyes. The method in this article uses a self-attention mechanism in the initial stage so that the model can not only By learning pixel information such as background and other colours; the model can also capture the structural information of the target, correctly generate the position and number of bird heads, beaks, and bird eyes, and improve the synthesis of bird images that AttnGAN does not match with text features.

4.4.2. Image Analysis of Bird's Overall Generation Failure. The self-attention mechanism can learn important spatial and positional information, improve errors such as "multiple heads" and "multiple mouths," improve the stability of the t2i model, and generate more realistic bird pictures. As shown in Figures 9 and 10, the pictures caused by AttnGAN cannot be seen as birds, and the generated birds do not match the real birds. Compared with the shape of the class, the model proposed in this article can synthesize images with solid correlation with the textual feature information. Taking the third set of comparative experiments in Figure 10 as an example, the picture synthesized by this model can correctly reflect the text attributes such as "white and brown" and "multicoloured beak" and the composition of the composition ensures the content of the synthesized image. Consistency with text descriptions and high discrimination with background image features, while the images synthesized by the AttnGAN method are distorted and do not correctly generate textual semantic information.

5. Conclusion

A multistage generative adversarial network is generating a lot of buzz these days because it is being used in AI techniques and computing models to explore anatomical visuals, images of somatic cells, and diverse human organ prosecutions, together with fingerprint scanning. As a result, this strategy is playing a key role in assessing images for diverse medical specialties and criminology for detecting fatal ailments and verifying new sequences of image clues for criminal justice. This article proposes a GAN-based t2i network model. By introducing the self-attention mechanism, the stability of the model is improved. Furthermore, the IS index and the FID index are optimized on the CUB dataset. Image synthesis by integrating AI techniques is employed. The experimental results show that the network proposed in this article can generate clear, natural, realistic, and diverse single-target images and has a specific generalization. In addition, the Chinese t2i dataset is further enriched. Future research will focus on the controllability of text-generated clothing images and apply them in clothing generation and design.

Data Availability

The data shall be made available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Jiang, M. Zhong, and F. Qiu, "Single-image super-resolution based on a self-attention deep neural network," in *Proceedings of the 2020 13th International Congress on Image and Signal Processing BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 387–391, Chengdu, China, October 2020.

[2] L. Wu, T. Tong, M. Du, and Q. Gao, "Image colorization algorithm based on self-attention network," in *Proceedings of the 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*, pp. 1–3, Fuzhou, China, December 2020.

[3] J. Feng, Z. Ye, D. Li, Y. Liang, X. Tang, and X. Zhang, "Hyperspectral image classification based on semi-supervised dual-branch convolutional autoencoder with self-attention," in *Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1267–1270, Waikoloa, HI, USA, September 2020.

[4] Y. Yang, H. Liang, Y. Yang, and W. Xu, "Image arbitrary style transfer via self-attention mechanism based on feature fusion," in *Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, pp. 58–63, Dali, China, June 2021.

[5] S. Xue, H. Hou, and K. Hui, "SARANIQA: self-attention restorative adversarial network for No-reference image quality assessment," in *Proceedings of the 2020 5th International Conference on Communication, Image and Signal Processing (CCISP)*, pp. 270–274, Chengdu, China, November 2020.

[6] T. Kang and K. H. Lee, "Unsupervised image-to-image translation with self-attention networks," in *Proceedings of the2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 102–108, Busan, Korea (South), February 2020.

[7] M. Xu, K. Huang, Q. Chen, and X. Qi, "Mssa-net: multi-scale self-attention network for breast ultrasound image segmentation," in *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 827–831, Nice, France, April 2021.

[8] Y. Li, J. Ni, A. Elazab, and J. Wu, "Multiple self-attention network for intracranial vessel segmentation," in *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Shenzhen, China, July 2021.

[9] Y. Gao, H. Luo, W. Zhu, F. Ma, J. Zhao, and K. Qin, "Self-attention underwater image enhancement by data augmentation," in *Proceedings of the 2020 3rd International Conference on Unmanned Systems (ICUS)*, pp. 991–995, Harbin, China, November 2020.

[10] M. Kim, T. Kim, and D. Kim, "Spatio-temporal slowfast self-attention network for action recognition," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2206–2210, Abu Dhabi, United Arab Emirates, October 2020.

[11] J. Zhang, P. Yang, W. Wang, Y. Hong, and L. Zhang, "Image editing via segmentation guided self-attention network," *IEEE Signal Processing Letters*, vol. 27, pp. 1605–1609, 2020.

[12] Q. Zhao, W. Yang, and Q. Liao, "Adasan: adaptive cosine similarity self-attention network for gastrointestinal endoscopy image classification," in *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1855–1859, Nice, France, April 2021.

[13] H. Tang, X. Qi, D. Xu, P. H. S. Torr, and N. Sebe, *Edge Guided GANs with Semantic Preserving for Semantic Image Synthesis*, https://doi.org/10.48550/ARXIV.2003.13898 (Version 1) arXiv, 2020.

[14] Y. Luo, D. Nie, B. Zhan et al., "Edge-preserving MRI image synthesis via adversarial network with iterative multi-scale fusion," in *Neurocomputing*vol. 452, pp. 63–77, Elsevier BV, 2021, https://doi.org/10.1016/j.neucom.2021.04.060.

[15] K. Huang, X. Deng, J. Geng, and W. Jiang, "Self-attention and mutual-attention for few-shot hyperspectral image classification," in *Proceedinds of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 2230–2233, Brussels, Belgium, July 2021.

[16] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2289–2301, 2020.

[17] H. Li, J.-Z. Cheng, Y.-H. Chou, J. Qin, S. Huang, and B. Lei, "AttentionNet: learning where to focus via attention mechanism for anatomical segmentation of whole breast ultrasound images," in *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1078–1081, Venice, Italy, April 2019.

[18] Z. Huang, Z. Chen, Q. Zhang et al., "CaGAN: a cycle-consistent generative adversarial network with attention for low-dose ct imaging," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1203–1218, 2020.

[19] H. Li and J. Tang, "Dairy goat image generation based on improved-self-attention generative adversarial networks," *IEEE Access*, vol. 8, pp. 62448–62457, 2020.

[20] A. Mathew, A. P. Patra, and J. Mathew, "Self-attention dense depth estimation network for unrectified video sequences," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2810–2814, Abu Dhabi, United Arab Emirates, October 2020.

[21] H. Mei, H. Zhang, and Z. Jiang, "Self-attention fusion module for single remote sensing image super-resolution," in *Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 2883–2886, Brussels, Belgium, July 2021.

[22] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Bi-SAN-CAP: Bi-directional self-attention for image captioning," in *Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7, Perth, WA, Australia, December 2019.

[23] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3192–3208, April 2021.

[24] C. Su, R. Huang, C. Liu, T. Yin, and B. Du, "Prostate MR image segmentation with self-attention adversarial training based on wasserstein distance," *IEEE Access*, vol. 7, pp. 184276–184284, 2019.

[25] S. Cui, Z. Zhou, L. Li, and E. Fei, "Unsupervised infrared and visible image fusion with pixel self-attention," in *Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC)*, pp. 437–441, Kunming, China, May 2021.

[26] Z. Li, C. Yuan, Y. Sun, and J. Wang, "Monocular depth recovery based on self-attention mechanism and transfer learning," in *Proceedings of the 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 75–80, Yibin, China, August 2021.

[27] A. Gupta and P. Prabhat, "Novel approaches in network fault management," *INTERNATIONAL JOURNAL OF NEXT-GENERATION COMPUTING*, vol. 8, no. 2, 2017.

[28] L. Zong and L. Chen, "Single image super-resolution based on self-attention," in *Proceedings of the 2019 IEEE International Conference on Unmanned Systems and Artificial Intelligence (ICUSAI)*, pp. 56–60, Xi'an, China, November 2019.

[29] J. Chen, L. Chen, and M. Shabaz, "Image fusion algorithm at pixel level based on edge detection," *Journal of Healthcare Engineering*, vol. 2021, Article ID 5760660, 10 pages, 2021.

[30] Y. Li, H. Huang, L. Zhang, G. Wang, H. Zhang, and W. Zhou, "Super-resolution and self-attention with generative adversarial network for improving malignancy characterization of hepatocellular carcinoma," in *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1556–1560, Iowa City, IA, USA, April 2020.

[31] R. Zhang, M. Xu, Y. Shi, J. Fan, C. Mu, and L. Xu, "Infrared target detection using intensity saliency and self-attention," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1991–1995, Abu Dhabi, United Arab Emirates, October 2020.

[32] C. Sharma, A. Bagga, R. Sobti, M. Shabaz, and R. Amin, "A robust image encrypted watermarking technique for neurodegenerative disorder diagnosis and its applications," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 8081276, 14 pages, 2021.

[33] Z. Li, L. Yuan, H. Xu, R. Cheng, and X. Wen, "Deep multi-instance learning with induced self-attention for medical image classification," in *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 446–450, Seoul, Korea (South), December 2020.

[34] H. Wu, S. Zhao, L. Li, C. Lu, and W. Chen, "Self-attention network with joint loss for remote sensing image scene

classification," *IEEE Access*, vol. 8, pp. 210347–210359, 2020.

[35] L. Kapoor, S. Bawa, and A. Gupta, "Peer clouds: a P2P-based resource discovery mechanism for the Intercloud," *International Journal of Next-Generation Computing*, , pp. 153–164, Perpetual Innovation Media Pvt. Ltd, 2015.

[36] F. Chaabane, S. Rejichi, and F. Tupin, "Self-attention generative adversarial networks for times series VHR multispectral image generation," in *Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 4644–4647, Brussels, Belgium, July 2021.

[37] A. Kapoor, A. Gupta, R. Gupta, S. Tanwar, G. Sharma, and I. E. Davidson, "Ransomware detection, avoidance, and mitigation scheme: a review and future directions," *Sustainability*, vol. 14, no. 1, p. 8, 2021.

[38] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8671–8685, 2021.