International Core Journal of Engineering Volume 7 Issue 8, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

Text Generation Image Model based on GANs

Yipeng Wang[a], Ruonan Wu[b]
School of Dalian Ocean University, Dalian 116023, China.
[a]aowuuu@dingtalk.com, [b]1067234653@qq.com

Abstract

In order to solve the problems of poor quality and unstable training in the current generation of text-based images, a high-quality image generation model based on single-stage generation confrontation network (GANs) is proposed. Specifically, the attention mechanism is introduced into the GANs generator to generate fine-grained images, and the local global language representation is added to the discriminator to accurately identify the generated image and the real image; Through the game between generator and discriminator, high quality image is generated. The experimental results on benchmark data set show that compared with the latest model with multi-stage framework, the image generated by this model is more realistic and achieves the highest IS value, which can be better applied to the scene with image generated by text description.

Keywords

Text Generating Image; Generation of Countermeasure Network; Attention Mechanism.

1. Introduction

It has become a challenging task in computer vision (CV) and NLP to generate high resolution realistic images based on given text description. The subject has a variety of potential applications, such as art creation, photo editing and video games.

Recently, due to the generation of countermeasure networks (GANs) [1] in the generated image.

In 2016, reed proposed to generate a reasonable image from the text description by condition generation (cGANs) [2]; Zhang H proposed StackGAN + + [4] model by stacking multiple generators and discriminators in 2017, and 256 was generated for the first time × 256 resolution image. Currently, almost all text-generated image models are based on StackGAN. These models have multiple pairs of generators and discriminators, which generate initial images by embedding text and random noise into the first generator, In the subsequent generator, the initial image is refined and the high resolution image is finally generated. For example, AttnGAN[5] introduces a cross modal attention mechanism in each generator to help the generator synthesize images in more detail; MirrorGAN [6] regenerates the text description from the generated image to achieve the semantic consistency between text and image; DM-GAN [7] introduces dynamic memory network [8] to solve the problem of stack structure training instability.

Although stack generation counter network has achieved good results in text generation image, there are still two problems that can not be solved: firstly, training multiple networks will increase the computing time and affect the stability of the generation model; Moreover, if the generators in the previous phase do not converge to the global optimal value, the final generation network will not be improved because the final generator gradient will be difficult to return. Secondly, in the process of generating the first stage of initial image, the generator network is composed of upper sampling layer and volume layer, and lacks the process of image integration and refinement using input natural

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

language text, which makes the initial image quality poor, The final image is lack of fine-grained information.

In order to solve the above problems, this paper proposes a text generation image network based on single-stage GANs, which can fine tune the characteristic graph of each scale according to the given text description, and only a single generator and discriminator can generate high-quality images. Specifically, in the generator, a channel pixel attention module is designed, which gradually associates the channel and pixel information in the visual feature graph with the text description, and calculates the attention weight of the visual feature graph based on the global text embedding, To find the most relevant feature map of text description; In discriminator, the global text representation and local word embedding technology are used to provide fine-grained discrimination signal for discriminator, and the visual feature map of the last lower sampling block is projected to global text representation, and the visual feature map of the second bottom sampling block is projected into the local word, The discriminator can be distinguished by integrating local and global language representation as supervisory information.

2. Model method

2.1 Network structure

As shown in Figure 1, the network structure of this paper consists of text encoder, generator and discriminator. For text encoder, Bi long and short term memory network (Bi - LSTM) [9] is used to learn the semantic representation of a given text description. In Bi - LSTM, two hidden states are used to capture the semantics of words as local language representation, and the last hidden state is used to represent sentence features as global language representation. The generator is composed of seven up sampling blocks, which are responsible for different scale feature maps. Each up sampling block includes two convolution layers, two conditional batch normalization layers [10] and a channel pixel attention module. The discriminator consists of seven down sampling blocks and a local global projection block. The downsampling blocks can be regarded as image encoders, which encode the input image into a high-dimensional feature map. Each downsampling block consists of a convolution layer and an average pooling layer. The local global projection block projects the last two downsampling blocks into local and global linguistic representations respectively.

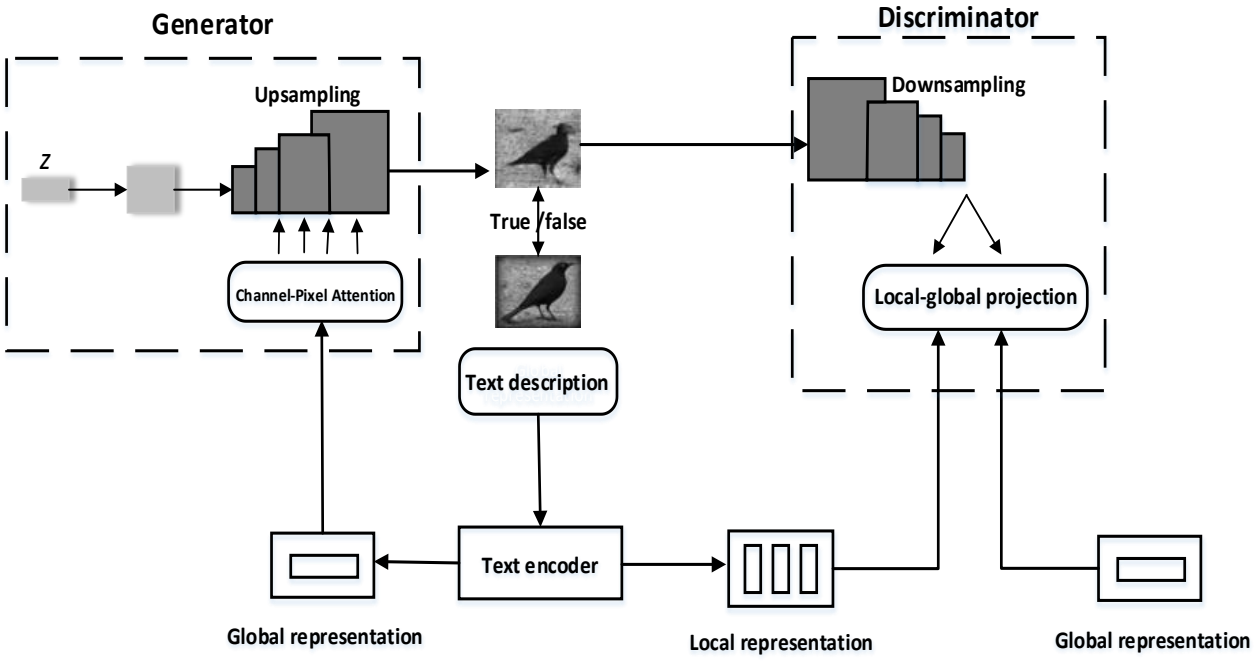

Fig. 1 Network structure of text generated image

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

2.1.1 Generator

The generator takes the global representation vector *S* and noise vector *Z* of the text as input, and consists of seven up sampling blocks, which are used to generate the visual feature map of each resolution. The whole image generation process is shown in equation (1):

$$\begin{cases} h_0 = F_0(z) \\ h_1 = F_1(h_0, s) \\ h_i = F_i(h_{i-1}, s)(i = 2,3,\cdots,7) \\ o = G_c(h_7) \end{cases} \tag{1}$$

Where *Z* is the random noise obeying normal distribution, F_0 is the fully connected layer, *Fi* is the residual layer containing channel pixel attention, G_c is the last convolution layer used to generate the final image *o*, h_0 is the hidden state of the initial fully connected layer, and h_1~h_7 is the intermediate representation of the output of the residual layer.

In order to consider the channel and spatial pixel information of the feature map of the volume layer, the channel and pixel awareness attention mechanism are introduced in the residual block. Since each feature map in convolution layer has different importance to text embedding, this paper introduces channel pixel awareness attention module to guide the generator to focus on selecting feature graph related to text and ignoring secondary feature graph. The channel aware attention module is shown in Figure 2.

Channel awareness attention module has two inputs: feature graph *h* and global representation *s* of text. Firstly, channel features x_a and x_m are obtained by averaging pooling (GAP) and maximizing pooling (GMP) of *h*, as shown in equation (2):

$$\begin{cases} x_a = GAP(h) \\ x_m = GAP(h) \end{cases} \tag{2}$$

In the formula, GAP is used to obtain the information of the whole feature graph, while GMP is used to extract the most distinctive part of the feature graph. Then, query (*q*), key (*k*) and value (*v*) are used to capture the semantic correlation between channel and input text, where x_a and x_m are used as query, and global representation *s* is used as key and value, the process definition is shown in equation (3):

$$\begin{cases} q_a = w_a x_a \\ q_m = w_m x_m \\ k_c = w_k s \\ v_c = w_v s \end{cases} \tag{3}$$

Where w_a, w_m, w_k and w_v are the projection matrices realized by convolution, so as to realize dimension matching in the process of attention calculation. The calculation process of channel perceived attention is defined as equation (4):

$$\begin{cases} \alpha_a^c = q_a k_c^T \\ \alpha_m^c = q_m k_c^T \\ \beta_a^c = soft\,max(\alpha_a^c, v_c) \\ \beta_m^c = soft\,max(\alpha_m^c, v_c) \end{cases} \tag{4}$$

Where α_a^c, α_m^c is the semantic correlation between channel graph and global representation, β_a^c, β_m^c refers to the final channel perceived attention weight after average pooling and maximum pooling. At the same time, as shown in equation (5), an adaptive gate is designed to fuse the results of average pooling and maximum pooling.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

$$\begin{cases} g^c = \sigma(w_1 x_a + w_2 x_m) \\ o_c = g^c \beta_a^c + (1 - g^c)\beta_m^c \end{cases} \tag{5}$$

Where w_1 and w_2 are learnable matrices, σ is the sigmoid function. Finally, the final result is generated by adaptive residual connection.

The image is composed of related pixels, which are very important for the quality and semantic consistency of the composite image. Therefore, after obtaining the feature map of channel attention, the new feature map is used to calculate the pixel perceptual attention, so as to effectively model the relationship between the spatial pixels and the given natural language description, and make the important pixels receive more attention from the generator. Compared with the channel aware attention calculation, the pixel aware attention ignores the influence of the channel information of each feature map, and only focuses on the weight of the spatial information in the feature map to the visual pixels, and its calculation process is similar to the channel aware attention.

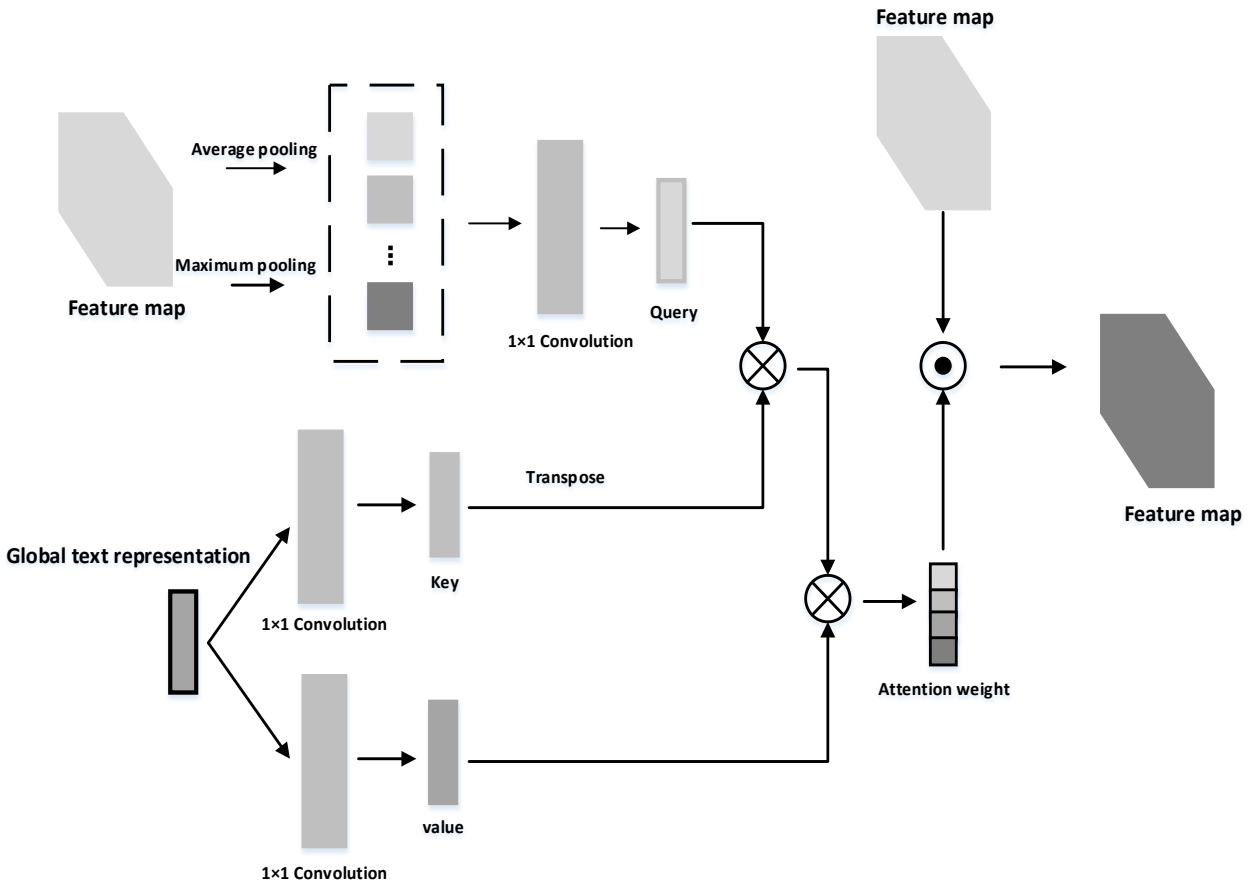

Fig. 2 Attention mechanism of channel perception

2.1.2 Discriminator

Discriminator plays two important roles. On the one hand, it is responsible for identifying whether the image is real or generated; On the other hand, it determines whether image and text descriptions are semantically related. This paper proposes a local global projection block in the discriminator to capture the correlation between vision and semantics, and its structure is shown in Figure 3. The last layer of feature graph v_D is projected to the global representation s of the text, and the penultimate layer of feature graph v_{D-1} is projected to the local representation S_l of the text. The idea behind this operation is that v_D is closer to the global semantics of the text in the high-dimensional semantics of vision; The low dimensional visual representation of v_{D-1} is more suitable for the local embedding of text. In this paper, cross modal projection is designed to correlate visual and text information.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

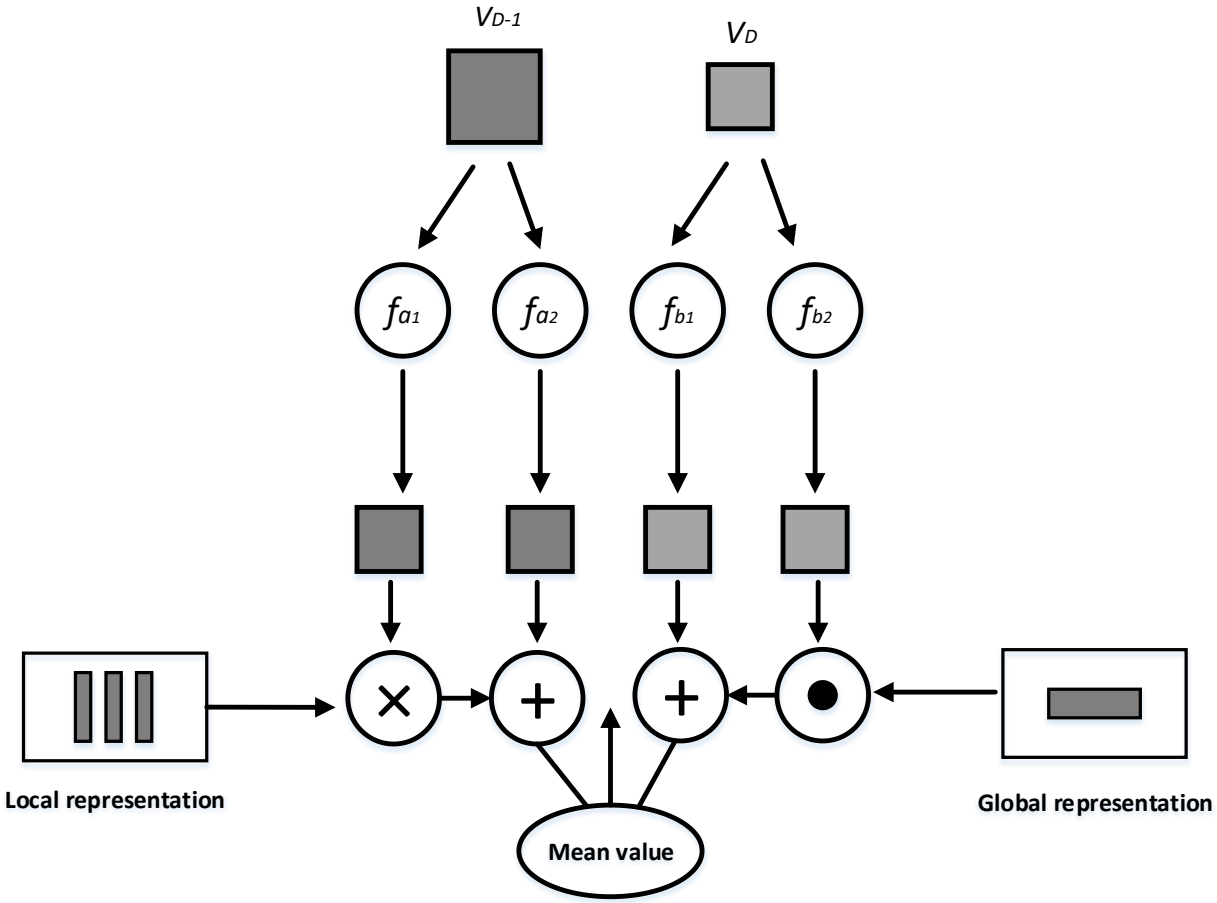

Fig. 3 Local global projection structure

Specifically, the projection operation first copies the feature graph, and then feeds the original feature graph and the copied feature graph into two fully connected layer networks. One of the output results is multiplied by the language representation, and finally outputs the mean value of the two layer feature graphs after projection operation. Since the global and local language representations have two different formats, one is a vector and the other is a matrix, the projection methods of matrix multiplication and element by element multiplication are adopted for each format, as shown in equations (6):

$$\begin{cases} P(v_{D-1}, s_l) = f_{a1}(v_{D-1}) \times s_l + f_{a2}(v_{D-1}) \\ Q(v_D, s) = f_{b1}(v_D) \times s + f_{b2}(v_D) \end{cases} \tag{6}$$

Where $f_{a1}()$ and $f_{a2}()$ are for the two fully connected layers of v_{D-1}, $f_{b1}()$ and $f_{b2}()$ are for the two fully connected layers of v_D. The total discriminator output is shown in (7):

$$D(I, s_l, s) = \frac{1}{N_p}\sum_{i=1}^{N_p} P_i + \frac{1}{N_Q}\sum_{j=1}^{N_Q} Q_j \tag{7}$$

The dimensions of the two projection vectors P and Q are N_P and N_Q respectively, and the subscripts i and j represent the indexes of the dimensions. I contains real images and generated images.

The projection module provides local and global linguistic representations as conditional information embedding discriminators. This method provides fine-grained gradients for training the whole text to image generation model, so as to obtain the correlation between visual and linguistic representations.

2.2 Loss function.

Anti loss is used to match the generated sample with the given text description. In this paper, hinge loss [11] is used to stably generate the training of the confrontation network. The basic idea is to keep the generated negative samples and the real samples in a decision interval, so as to avoid the gradient

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

oscillation when the two samples deviate too much. The counter loss function of the discriminator is shown in (8):

$$L_{adv}^{D} = E_{x\to p}\left[max\left(0,1-D\left(x,s_l,s\right)\right)\right]+\frac{1}{2}E_{x\to p_c}\left[max\left(0,1+D\left(\hat{x}\left(z,s\right),s_l,s\right)\right)\right]+\frac{1}{2}E_{x\to p}\left[max\left(0,1+D\left(x,\hat{s}_l,\hat{s}\right)\right)\right] \quad (8)$$

Where x is the sample from the real image data distribution p, $\hat{x}$ (z, s) is the sample generated from the generated data distribution p_G, and $\hat{s}_l$ and $\hat{s}$ are text representations that do not match the image. The corresponding generator loss function is shown in equation (9):

$$L_G = E_{x\to p_G}[D(x, s_l, s)] \quad (9)$$

At the same time, in order to improve the semantic consistency of the generated image, MA-GP loss [12] is added to the discriminator to optimize the gradient of the real image and the given text description. MA-GP loss [12] is a zero centered gradient penalty, which makes the generated data distribution more likely to converge to the real distribution. The expression is shown in (10):

$$L_{MA-GP} = E_{x\to p}[(\|\nabla_x D(x,s)\|_2 + \|\nabla_s D(x,s)\|_2)^T] \quad (10)$$

Therefore, the total loss function of the discriminator is shown in equation (11):

$$L_D = L_{adv}^{D} + \lambda_1 L_{MA-GP} \quad (11)$$

Among them γ and λ_1 is the super parameter, which is set to 6 and 0.1 respectively.

3. Experiment and analysis

This section mainly introduces the data sets used in the experiment, model training details and evaluation indicators, and then evaluates the proposed model quantitatively and qualitatively.

3.1 Data sets and training details

In this paper, a model evaluation experiment was carried out on cub bird data set [13]. The dataset contains 11 788 images of 200 bird species, each with 10 descriptions in English. The data set was preprocessed according to DM-GAN [7] method. Among them, 8855 images of 150 bird species were used as training set, and 2933 images of the remaining 50 bird species were used as test set. In this paper, Adam [14] optimizer is used to optimize the model network. At the same time, according to TTUR [15], the learning rate of generator is set to 0.0001, and the learning rate of discriminator is set to 0.0004.

3.2 Evaluation index

According to the previous work [5, 7], this paper selects the perception score (IS) [16] to evaluate the network performance proposed in this paper. The definition of IS is shown in equation (12):

$$IS = exp\left(E_x D_{KL}\left(p(y|x)\|p(y)\right)\right) \quad (12)$$

Where x is the generated image and y is the label generated through the pre training of the perception v3 network [17], Is calculates the KL divergence between the conditional distribution $p(y|x)$ and the marginal distribution $P(y)$. If the model can generate a variety of images that match the text, the larger the KL divergence is. The higher the is value, the higher the image quality and the more diverse the images belong to the same category. Since the CUB bird data set used in this paper is disjoint in the training set and the test set, but the perception v3 network has been pre trained in the test set, the is value on the CUB bird test data set can be used to evaluate the semantic consistency of text images.

3.3 Quantitative analysis

In this paper, StackGAN++ [4], AttnGAN[5] and DM-GAN [7] are selected as the best models for multi-stage stack structure text generation in recent three years. As shown in Table 1, the proposed model based on single-stage GANs has the highest is value on the CUB data set. Higher IS value on CUB test set means higher image quality and better matching of image text semantics. Compared with AttnGAN [5] only uses pixel attention in the front layer of the full connection layer of each generator, the single stage GANs in this paper uses channel and pixel attention to each residual block

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

at the same time, and is value is is value is increased from 4.36 to 4.88; Compared with DM-GAN [7], the fuzzy images generated in each stage are refined by introducing additional dynamic memory network. In this paper, the IS value is increased from 4.75 to 4.88 by local global representation of discriminator. The quantitative comparison of initial score (IS) shows that the single-stage GANs model proposed in this paper can synthesize more realistic images and have better semantic consistency of text images.

Table 1. Comparison of IS scores of different models on CUB dataset

Model method	IS
StackGAN++	4.04±0.06
AttnGAN	4.36±0.03
MirrorGAN	4.56±0.05
DM-GAN	4.75±0.07
this paper	4.88±0.03

Fig. 4 Generate image contrast

3.4 Qualitative analysis

As shown in Figure 4, from top to bottom are StackGAN++ [4], AttnGAN [5], DM-GAN [7] and the visual effect of the image generated by the model in this paper according to the text. It can be found in the figure that the images generated by StackGAN++ and AttnGAN lack visual authenticity and are more like the stacking of some simple text attributes. The reason is that the stacking of multiple generators and discriminators causes the gradient to disappear, And both models only use the spatial attention mechanism of visual features and ignore the channel attention between each feature.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

Although the introduction of dynamic memory network into DM-GAN further alleviates the problem that the generated images seem to be simple combination and lack of visual authenticity, there is still a lack of coherence between visual pixels (for example, the surface skin of birds generated by the first column of DM-GAN is rough).

The proposed model only uses a pair of generated countermeasure networks with residual structure by removing the stacked structure, and introduces channel attention in the generator and local global projection in the discriminator, It makes the generated image more realistic and diverse rather than the stacking of various attributes.

4. Conclusion

This paper presents a new method to generate the confrontation network based on the deep fusion of single stage, which is used for text to image generation. Compared with the previous multi-stage model, the model can directly synthesize more realistic and text semantic consistent images, and does not need to stack multiple generation countermeasures network. In addition, a attention mechanism combining channel and pixel is proposed to guide the generator to synthesize realistic images, and the local and global language representation is embedded in discriminator to cooperate with the generator to generate images. The experiment shows that the model proposed in this paper has achieved remarkable results in the CUB data set, and is superior to the latest model in quantitative and qualitative results.

References

[1] Li Chong-yang, Wang Jian. The effect of atomistic substitution on thermal transport in large phonon bandgap GaN[J]. Japanese Journal of Applied Physics, 2021, 60(7). China National Standardization Management Committee. Specifications of Crane Design (China Standardization Press, China 2008), p. 16-19.

[2] Matys Maciej, Ishida Takashi, Nam Kyung Pil, Sakurai Hideki, Kataoka Keita, Narita Tetsuo, Uesugi Tsutomu, Bockowski Michal, Nishimura Tomoaki, Suda Jun, Kachi Tetsu. Design and demonstration of nearly-ideal edge termination for GaN p–n junction using Mg-implanted field limiting rings[J]. Applied Physics Express, 2021,14(7).

[3] Nakamura Toshihiro, Nishimura Tomoaki, Kuriyama Kazuo, Nakamura Tohru, Kinomura Atsushi. Gamma-ray induced photo emission from ZnO single crystal wafer: Comparison with GaN[J]. Solid State Communications, 2021,336.

[4] Uedono Akira, Takino Junichi, Sumi Tomoaki, Okayama Yoshio, Imanishi Masayuki, Ishibashi Shoji, Mori Yusuke. Vacancy-type defects in bulk GaN grown by oxide vapor phase epitaxy probed using positron annihilation[J]. Journal of Crystal Growth, 2021,570.

[5] Fukuhara Noboru, Horikiri Fumimasa, Narita Yoshinobu, Isono Ryota, Tanaka Takeshi. Substrate off-angle dependency of Al content in AlGaN/GaN high-electron-mobility transistor structures on free-standing GaN substrates[J]. Japanese Journal of Applied Physics, 2021,60(7).

[6] Khan Muhammad Saddique Akbar, Liao Hui, Yu Guo, Iqbal Imran, Lei Menglai, Lang Rui, Mi Zehan, Chen Huanqing, Zong Hua, Hu Xiaodong. Reduction of threading dislocations in GaN grown on patterned sapphire substrate masked with serpentine channel[J]. Materials Science in Semiconductor Processing, 2021,134.

[7] Monish Mohammad, Nayak C, Sutar D S, Jha S N, Bhattacharyya D, Major S S. X-ray absorption study of defects in reactively sputtered GaN films displaying large variation of conductivity[J]. Semiconductor Science and Technology, 2021,36(7).

[8] Yifan Lu, Siyuan Fu, Xiao Hua Zhang, Ning Xie. Denoising Monte Carlo renderings via a multi-scale featured dual-residual GAN[J]. The Visual Computer, 2021(prepublish).

[9] Dub Maksym, Sai Pavlo, Sakowicz Maciej, Janicki Lukasz, But Dmytro B., Prystawko Paweł, Cywiński Grzegorz, Knap Wojciech, Rumyantsev Sergey. Double-Quantum-Well AlGaN/GaN Field Effect Transistors with Top and Back Gates: Electrical and Noise Characteristics[J]. Micromachines,2021,12(6).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202108_7(8).0068

[10] Gokhale Vikrant J, Downey Brian P, Roussos Jason A, Scott Katzer D, Meyer David J. Passive high power RF comb filters using epitaxial GaN/NbN/SiC HBARs. [J]. IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 2021, PP.

[11] Reddeppa Maddaka, Nam DongJin, Bak NaHyun, Pasupuleti Kedhareswara Sairam, Woo Hyeonseok, Kim SongGang, Oh JaeEung, Kim MoonDeock. Proliferation of the Light and Gas Interaction with GaN Nanorods Grown on a V-Grooved Si (111) Substrate for UV Photodetector and NO2 Gas Sensor Applications. [J]. ACS applied materials & interfaces, 2021.

[12] Li Ban, Luo Senlin, Qin Xiaonan, Pan Limin. Improving GAN with inverse cumulative distribution function for tabular data synthesis[J]. Neurocomputing, 2021,456.

[13] Zhou Dejin,Xu Hong,Chen Leilei,Lu Hong Liang,Huang Wei,Zhang David Wei,Yan Dawei. Temperature dependent characteristics of Ti/Al/Ni/Au Ohmic contact on lattice-matched In0.17Al0.83N/GaN heterostructures[J]. Solid State Electronics,2021,183.

[14] Mishra Puneet, Herrmann Ittai. GAN meets chemometrics: Segmenting spectral images with pixel2pixel image translation with conditional generative adversarial networks [J]. Chemometrics and Intelligent Laboratory Systems, 2021, 215(prepublish).

[15] Ji Keyu, Cui Xiao, Chen Jiwei, Guo Qi, Jiang Bing, Wang Bingjun, Sun Wenhong, Hu Weiguo, Hua Qilin. Effect of backside dry etching on the device performance of AlGaN/GaN HEMTs[J]. Nanotechnology, 2021, 32(35).

[16] Imai Daichi, Murakami Yuto, Miyata Rino, Toyoda Hayata, Yamaji Tomoaki, Miyoshi Makoto, Takeuchi Tetsuya, Miyajima Takao. Corrigendum: “Analysis of the optical constants and bandgap energy in AlInN alloys grown on a -plane freestanding GaN substrate by using spectroscopic ellipsometry” [J]. Japanese Journal of Applied Physics, 2021,60(7).

[17] Yuan Tao, Tan Jin, Han Cong, Xiong Ao. Highly sensitive and self-powered ultraviolet photodetector based on GaN/poly (styrenesulfonate)/ polyaniline hybrid heterojunction[J]. Materials Letters,2021,299.