Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 6452536, 10 pages
https://doi.org/10.1155/2020/6452536

Research Article

Text to Realistic Image Generation with Attentional Concatenation Generative Adversarial Networks

Linyan Li,[1] Yu Sun,[2] Fuyuan Hu,[2,3] Tao Zhou,[4] Xuefeng Xi,[2,5] and Jinchang Ren[6]

[1]*College of Information Technology, Suzhou Institute of Trade & Commerce, Suzhou 215009, China*
[2]*College of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*
[3]*Suzhou Key Laboratory for Big Data and Information Service, Suzhou 215009, China*
[4]*School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China*
[5]*Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, Suzhou 215009, China*
[6]*University of Strathclyde, Glasgow, UK*

Correspondence should be addressed to Fuyuan Hu; fuyuanhu@mail.usts.edu.cn

Received 2 July 2020; Revised 25 September 2020; Accepted 6 October 2020; Published 28 October 2020

Academic Editor: Longzhuang Li

In this paper, we propose an Attentional Concatenation Generative Adversarial Network (ACGAN) aiming at generating 1024×1024 high-resolution images. First, we propose a multilevel cascade structure, for text-to-image synthesis. During training progress, we gradually add new layers and, at the same time, use the results and word vectors from the previous layer as inputs to the next layer to generate high-resolution images with photo-realistic details. Second, the deep attentional multimodal similarity model is introduced into the network, and we match word vectors with images in a common semantic space to compute a fine-grained matching loss for training the generator. In this way, we can pay attention to the fine-grained information of the word level in the semantics. Finally, the measure of diversity is added to the discriminator, which enables the generator to obtain more diverse gradient directions and improve the diversity of generated samples. The experimental results show that the inception scores of the proposed model on the CUB and Oxford-102 datasets have reached 4.48 and 4.16, improved by 2.75% and 6.42% compared to Attentional Generative Adversarial Networks (AttenGAN). The ACGAN model has a better effect on text-generated images, and the resulting image is closer to the real image.

1. Introduction

In recent years, with the rise of artificial intelligence and deep learning, natural language processing and computer vision have become the hot research fields. The text to image as a basic problem in the field has also attracted the attention and research of many scholars. Text to image is the generation of a realistic image that matches a given text description, requiring processing fuzzy and incomplete information in natural language descriptions. Text to image drives the development of multimodal learning and cross-modal generation and shows great potential in applications such as cross-modal information retrieval, photo editing, and computer-aided design.

Since Goodfellow et al. [1] proposed Generative Adversarial Networks (GANs) in 2014; the network model has received extensive attention from academia and industry. With the continuous development of GAN, it has been widely used to generate realistic high-quality images based on text descriptions. The commonly used method [2–5] encodes the entire text description into a global sentence vector, which is input to the generator as a condition variable of GAN to generate an image. However, due to the large structural differences between text and images, the use of only word-level attention does not ensure the consistency of global semantics, while it is difficult to generate complex scenes; moreover, fine-grained word information is still not explicitly used for generating images.

Therefore, the generated image does not contain enough details and is still significantly different from the real image.

To address this issue, this paper proposes Attention Cascade Generative Adversarial Networks (ACGAN). The network adopts multilevel cascade structure, the generator and discriminator in each layer are composed of convolution units, and a new network layer is added layer by layer during the training process, and the generator and discriminator are added for processing the details of the higher resolution image. At the same time, the deep attentional multimodal similarity model is introduced into the network, focusing on the fine-grained information of the word level in the semantics. The word vector is used as the input of the generator, and through the constraint of the word vector, in the case of ensuring that the overall shape of the image is unchanged, the details of the generated image are emphasized, the consistency of the image and the semantic cross-modality is maintained, and the generation process is smooth. Finally, a measure of diversity is added to a layer of the discriminator to influence the discriminator's discriminant, so that the generator can obtain more diverse gradient directions, increase the diversity of generated samples, and improve the quality of generated samples.

The contribution of our method is threefold as follows:

(i) A multilevel cascade structure is proposed, which improves the resolution of the generated image, and can generate a high-resolution image of up to 1024×1024.

(ii) Introduce the attention mechanism model into the network, and make the details of the generated image richer by paying attention to the fine-grained information of the word level in the semantics.

(iii) Add the measure of diversity to the discriminator, increase the diversity of the generated samples, and improve the quality of the generated samples.

2. Related Works

Generative image modeling is a fundamental problem in computer vision. There has been remarkable progress in this direction with the emergence of deep learning techniques. Variational Auto Encoders (VAEs) [6, 7] is aimed to maximize the lower bound of the data likelihood. Autoregressive models (e.g., PixelRNN) [8] that utilized neural networks to model the conditional distribution of the pixel space have also generated appealing synthetic images. Recently, Generative Adversarial Networks (GANs) have shown promising performance for generating sharper images and video [9–11]. For example, Eghbal-zadeh et al. [12] proposed a Mixture Density Generative Adversarial Networks to improve the clarity and quality of generated images. Gecer et al. [13] combined the generated confrontation network with a deep convolutional neural network to reconstruct a 3D facial structure from a single face image. But training instability makes it hard for GAN models to generate high-resolution images. A lot of work has been proposed to stabilize the training and improve the image quality [14–19].

Generating high-resolution images from text descriptions, though very challenging, is important for many practical applications such as art generation and computer-aided design. Lyu et al. [9] learn joint embedding to establish the relationship between natural language and real images, and then train GAN to generate 64×64 images that are conditional on text descriptions. Cao et al. [10] proposed a Stacked Generative Adversarial Networks, which decompose the complex problem of generating high-quality images into some subproblems with better control and generate 256×256 high-resolution images.

Recently, attention models have been widely used in computer vision and natural language processing, for example, object detection [20, 21], video subtitle [22], and visual question answer [23, 24]. Xu et al. [25] introduced the attention mechanism into the GAN network and proposed Attentional Generative Adversarial Networks, which instruct the generator to focus on different word-level fine-grained information when generating different image subregions. Qiao et al. [26] proposed a global-to-local collaborative attention module that uses word attention and global sentence attention to enhance the consistency of generated images and semantics.

2.1. The Proposed Model. The Attentional Concatenation Generative Adversarial Networks model proposed in this paper consists of two parts: attentional concatenation generative adversarial networks and deep attentional multimodal similarity model. As shown in Figure 1, the Attentional Concatenation Generative Adversarial Networks model is divided into multiple levels; each layer contains a generator G and a discriminator D, using a multilevel cascade structure, increasing generators and discriminators layer by layer, and continuously adds a new residual network layer during the training process, corresponding to the generation from low-resolution to high-resolution images. The Deep Attentional Multimodal Similarity Model contains a common semantic space, mapping the subregions of the image and the word vector of the sentence into one of the semantic spaces, and measuring the word-level image and text similarity. Instead of adopting a one-step approach, the entire model's training process tries to generate low-resolution images, then continuously increase the resolution, and finally generate high-resolution and high-quality images.

2.2. Concatenation Generative Adversarial Networks. The generative network has k generators $(G_0, G_1, \ldots, G_{k-1})$, which take the hidden states $(h_0, h_1, \ldots, h_{k-1})$ as input to the generator $(G_0, G_1, \ldots, G_{k-1})$, generating images of different resolutions.

Specifically,

$$\begin{aligned} h_0 &= F_0\left(z, F^{ca}\left(c_s\right)\right), \\ h_i &= F_i\left(h_{i-1}, F_i^{\text{atten}}\left(c_w, h_{i-1}\right)\right), \quad i = 1, 2, \ldots, k-1, \\ \widehat{x}_i &= G_i\left(h_i\right). \end{aligned} \tag{1}$$

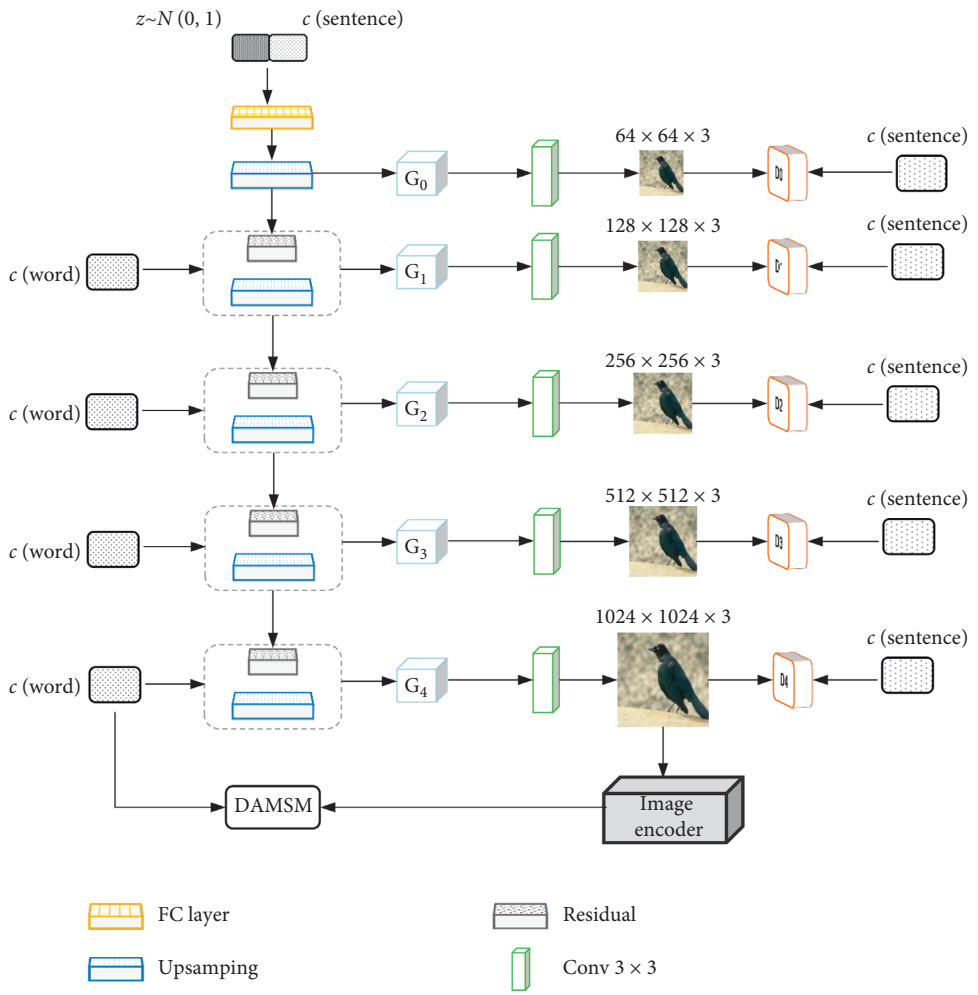

FIGURE 1: Attentional Concatenation Generative Adversarial Networks Model. Our training starts with both the generator and discriminator having a low spatial resolution of 64 × 64 pixels. As the training advances, we incrementally add layers to G and (D), thus increasing the spatial resolution of the generated images.

Here, z is a noise vector usually sampled from a standard normal distribution. c_s is a global sentence vector, and c_w is a word vector. F^{ca} represents the Conditioning Augmentation [10] that converts the sentence vector c_s to the conditioning vector. F_i^{atten} is the proposed attention model at the i^{th} stage of the attention model. The attention model $F^{\text{atten}}(c, h)$ has two inputs: the word features $c \in \mathbb{R}^{D \times T}$ and the image features $h \in \mathbb{R}^{\widehat{D} \times N}$ from the previous hidden layer.

Training starts with both the generator G and discriminator D having a low spatial resolution of 64 × 64 pixels. As the training advances, we incrementally add layers to G and D, and all existing layers remain trainable throughout the process. When doubling the resolution of the generator G and discriminator D we fade in the new layers smoothly. During the transition, we treat the layers that operate on the higher resolution like a residual block, whose weight increases linearly from 0 to 1.

Then we add a new residual layer and transform word features into semantic space of image features. Based on the hidden feature h of the image, a word context vector is calculated for each subregion of the image.

Finally, the image features and corresponding word context features are combined to generate an image in the next stage. In order to generate a real image with multiple levels (sentence level and word level) of conditions, the final objective function of the attention generation network is defined as

$$\begin{aligned} \mathscr{L} &= \mathscr{L}_G + \lambda \mathscr{L}_{\text{DAMSM}}, \\ \mathscr{L}_G &= \sum_{i=0}^{k-1} \mathscr{L}_{G_i}. \end{aligned} \tag{2}$$

Here, λ is a hyperparameter to balance the two terms of equation (2). The first term is the GAN loss that jointly approximates conditional and unconditional distributions. At the i^{th} stage of the ACGAN, the adversarial loss for is defined as

$$\mathscr{L}_{G_i} = -\frac{1}{2} E_{\widehat{x}_i \sim P_{Gi}} [\log (D_i (\widehat{x}_i))] - \frac{1}{2} E_{\widehat{x}_i \sim P_{Gi}} [\log (D_i (\widehat{x}_i, c_s))], \tag{3}$$

where the unconditional loss determines whether the image is real or fake, while the conditional loss determines whether the image and the sentence match or not.

As shown in Figure 2, for unconditional image generation, the discriminator is trained to distinguish between real images and forged images. For conditional image generation, images and variables are input to the discriminator to determine if the image matches the condition, and the bootstrap generator approximates the conditional image distribution. Discriminator D_i is trained to classify the input into the class of real or fake by minimizing the cross-entropy loss defined by

$$\begin{aligned}\mathscr{L}_{D_i} = &-\frac{1}{2}\mathbb{E}_{x_i\sim P_{\text{datai}}}[\log(D_i(x_i))]\\ &-\frac{1}{2}\mathbb{E}_{\widehat{x}_i\sim P_{Gi}}[\log(1-D_i(\widehat{x}_i))]+\\ &-\frac{1}{2}\mathbb{E}_{x_i\sim P_{\text{datai}}}[\log(D_i(x_i,c_s))]\\ &-\frac{1}{2}\mathbb{E}_{\widehat{x}_i\sim P_{Gi}}[\log(1-D_i(\widehat{x}_i,c_s))],\end{aligned} \tag{4}$$

where x_i is from the true image distribution p_{datai} at the i^{th} scale, and $\widehat{x}_i$ is from the model distribution P_{G_i} at the same scale. Discriminators D_i of the ACGAN are structurally disjoint, so they can be trained in parallel and each of them focuses on a single image scale.

2.3. Deep Attentional Multimodal Similarity Model. The Deep Attentional Multimodal Similarity Model [25] learns two neural networks that map subregions of the image and words of the sentence to a common semantic space, thus measuring the image-text similarity at the word level to compute a fine-grained loss for image generation.

This paper first uses a standard convolutional neural network to transform an image into a set of feature maps. Each feature map represents some subregions of the image. The dimension of the feature map is equal to the dimension of the word vector, and they are treated as equivalent entities. Next, based on each token in the text, attention is applied to the feature map and their weighted averages are calculated. Finally, the DAMSM is trained to minimize the difference between the attention vector and the word vector described above.

$$\mathscr{L}_1^w = -\sum_{i=1}^{k}\log P(S_i|M_i), \tag{5}$$

where "w" stands for "word".

Symmetrically, we also minimize

$$\mathscr{L}_2^w = -\sum_{i=1}^{k}\log P(M_i|S_i), \tag{6}$$

where P is the posterior probability that sentence S_i is matched with its corresponding image M_i.

Finally, the DAMSM loss is defined as

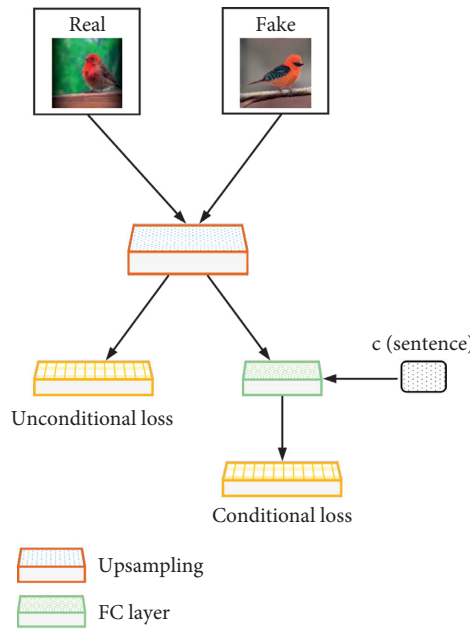

Figure 2: Discriminator model.

$$\mathscr{L}_{\text{DAMSM}} = \mathscr{L}_1^w + \mathscr{L}_2^w. \tag{7}$$

Using attention mechanism, the DAMSM is able to compute the fine-grained text-image matching loss $\mathscr{L}_{\text{DAMSM}}$. And $\mathscr{L}_{\text{DAMSM}}$ is only applied to the output of the last generator, because the ultimate goal of this paper is to generate high-resolution images through the last generator. If $\mathscr{L}_{\text{DAMSM}}$ is applied to the images generated by all generators $(G_0, G_1, ..., G_{k-1})$, the computational cost will increase greatly and the performance will not improve.

2.4. Standard Deviation of Measuring Diversity. GAN usually tends to capture only the changes found in the training data. In order to obtain more training data, this paper has greatly simplified this approach and has also improved the change based on "minibatch discrimination". Not only can feature statistics be calculated from a single image, but they can also calculate feature statistics for the entire small batch, thereby encouraging the generation of images and training images to display similar statistics. By adding a small batch layer at the end of the discriminator, the layer learns a large tensor and converts the input into a set of statistics. Finally, each instance is generated with a separate set of statistics and connected to the output of the layer so that the discriminator can use the statistics internally.

3. Experiments and Evaluation

3.1. Experimental Environment and Data. The algorithm uses the deep learning framework Tensorflow [27], and the experimental environment is Ubuntu 14.04 operating system, using four NVIDIA 1080Ti graphics processing unit (GPU) to accelerate the operation. At the same time, all models were trained on the CUB [28] and Oxford [29] datasets. As shown in Table 1, the CUB data set contains 200

Table 1: Datasets of experiment.

Dataset	CUB	CUB	Oxford	Oxford
—	Train	Test	Train	Test
Sample	8855	2933	7034	1155

Table 2: Inception scores and human rank scores for the five GAN models on the CUB and Oxford datasets.

Metric	Dataset	GAN-INT-CLS	GAWWN	StackGAN++	AttnGAN	ACGAN
Inception score	CUB	2.88 ± 0.04	3.62 ± 0.07	4.05 ± 0.05	4.36 ± 0.03	4.48 ± 0.05
	Oxford	2.66 ± 0.03	—	3.74 ± 0.03	—	3.98 ± 0.05
Human rank	CUB	2.81 ± 0.03	1.99 ± 0.04	1.37 ± 0.02	1.25 ± 0.03	1.17 ± 0.02
	Oxford	1.87 ± 0.03	—	1.13 ± 0.03	—	1.06 ± 0.02

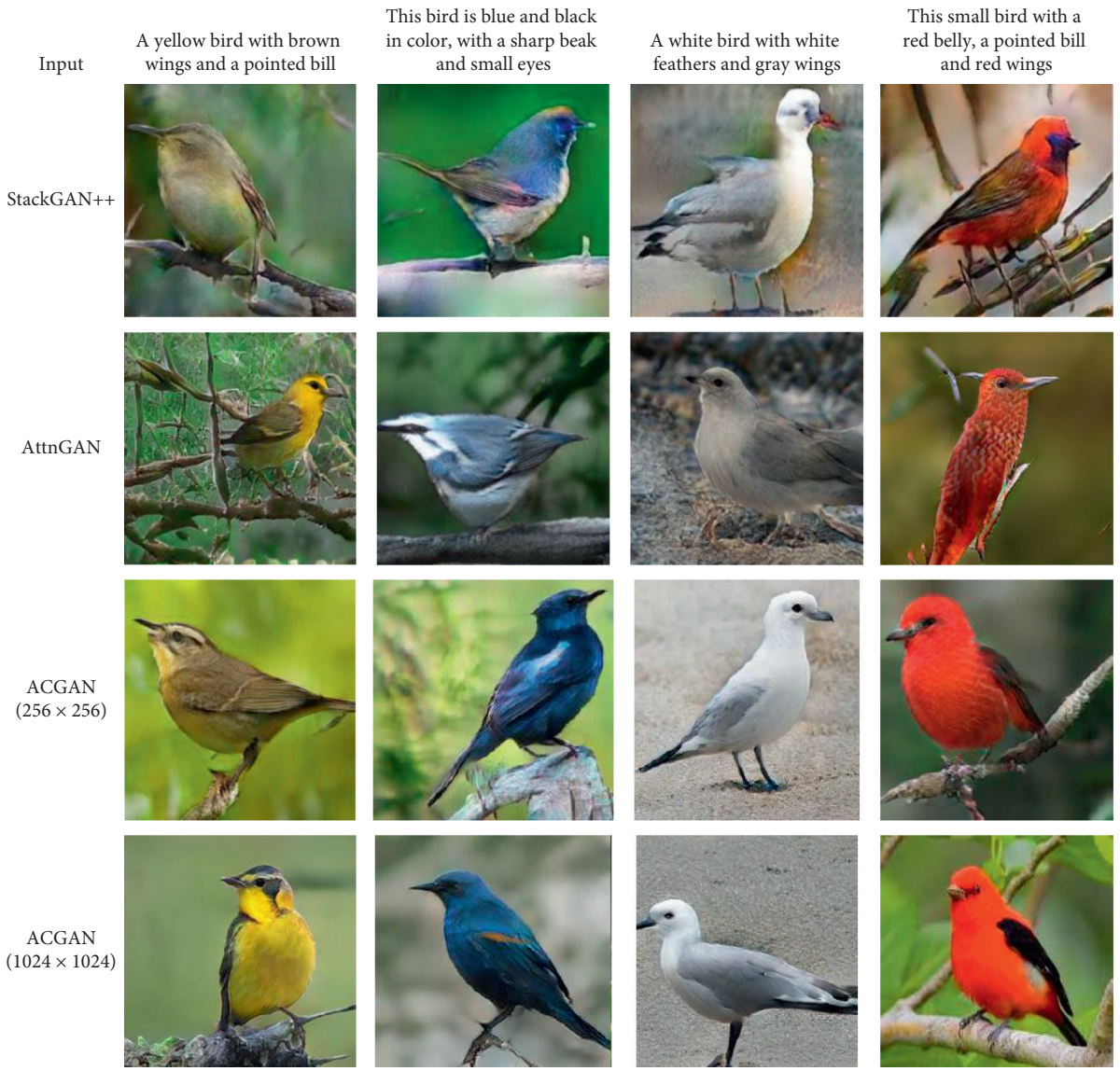

Figure 3: Images generated from descriptions using three GAN models trained on CUB test set.

species of birds with a total of 11,788 images. In this paper, 8855 images are used as training datasets and 2933 images as test datasets. Since the target area of 80% of the bird images in the dataset is less than 0.5 [28], we preprocess all images before training to ensure that the proportion of the bird target area is greater than 0.75 of the image size. The Oxford

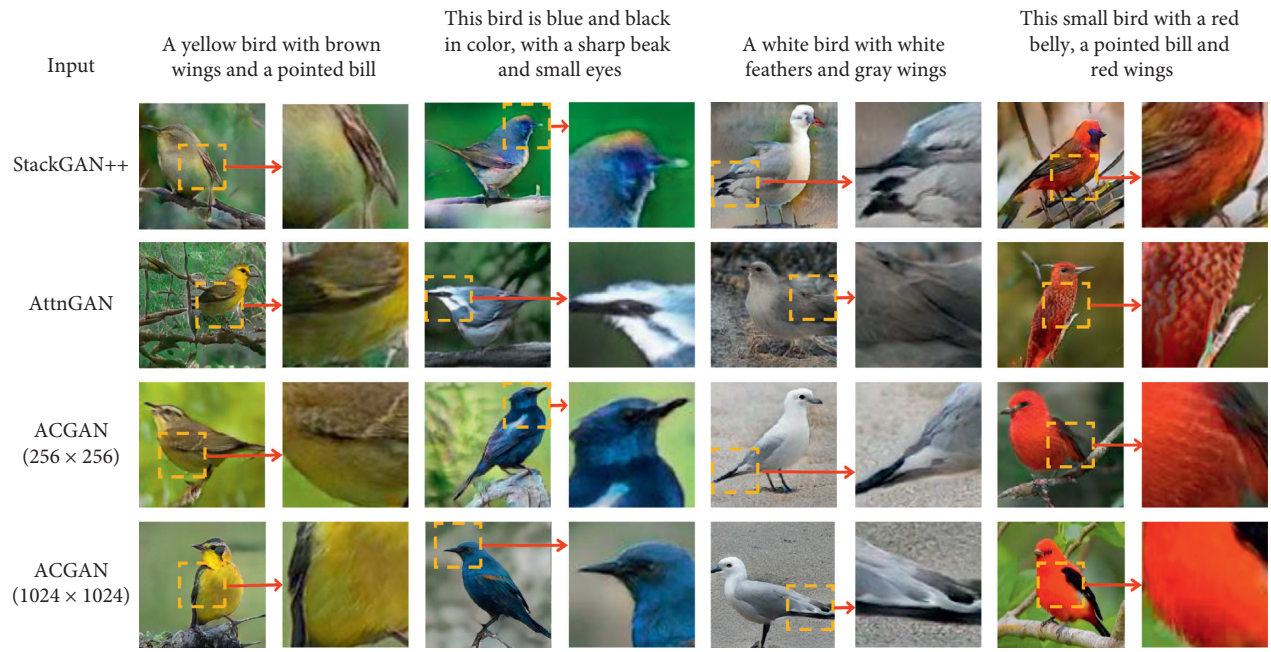

Figure 4: Details (beak, wings) comparison of the images generated from descriptions using three GAN models trained on CUB test set.

dataset contains 102 flower categories with a total of 8189 images. This article uses 7034 pictures as the training data set and 1155 pictures as the test data set.

3.2. Evaluation Metrics. For the evaluation of the GAN model, qualitative evaluation is usually used; that is, the visual fidelity of the image generated by manual inspection is required. This method is time-consuming and subjective and is somewhat misleading. Therefore, this paper mainly uses two evaluation criteria to evaluate the quality and diversity of generated images.

3.2.1. Inception Score. We choose numerical assessment approach "inception score" [16] for quantitative evaluation,

$$I = \exp\left(E_x D_{KL}\left(p\left(y \mid x\right) \| p\left(y\right)\right)\right), \tag{8}$$

where x denotes one generated sample, and y is the text label corresponding to the sample, $p(y)$ is the marginal distribution, and $p(y|x)$ is the conditional distribution. The KL divergence between the marginal distribution $p(y)$ and the conditional distribution $p(y|x)$ should be large, so that a variety of high-quality images can be generated. In the experiments, an inception model was given to the CUB data sets, and samples of each model were evaluated.

3.2.2. Human Rank. Human rank for qualitative assessment 50 text descriptions was randomly selected in the CUB and Oxford test sets, and for each sentence, the generated model generated 5 images. The five images and corresponding texts are described to different people to rank the image quality in different ways, and finally the average ranking is calculated to evaluate the quality and diversity of the generated images.

4. Experimental Result

The comparisons between the inception score and human rank results of various models on the CUB and Oxford datasets are presented in Table 2. As can be seen from the table, compared to the inception score of the AttnGAN model, the inception score of the ACGAN model on the CUB dataset has increased by 2.75% (Inception score increased from 4.36 to 4.48). Through the analysis of experimental results, ACGAN scores higher in Inception score than other GAN models; from an intuitive visual point, Human rank score is lower than other GAN models. It shows that the quality and diversity of the sample images generated by the model in this paper have been enhanced, and it is closer to the real image.

Subjective visual comparisons between the three models of StackGAN++, AttnGAN, and ACGAN on the CUB dataset are presented in Figure 3. It can be seen that the image details generated by StackGAN++ and AttnGAN are lost, colors are inconsistent with the text descriptions (1st and 2nd row), and the shape looks strange (2nd and 3rd column) for some examples. ACGAN achieved better results with more details and consistent colors and shapes compared to AttnGAN. For example, the wings are vivid in the 3rd and 4th row. By comparing ACGAN with AttnGAN, we can see that ACGAN contributes to producing fine-grained images with more details and better semantic consistency. For example, the color of the bird in the 2nd column was corrected to black. By comparing ACGAN (256 × 256) with ACGAN (1024 × 1024), we can see that the images generated by ACGAN (1024 × 1024) have higher definition, more vivid colors, and more lifelike details. Generally, content in the CUB dataset is less; therefore, it is easier to generate visually realistic and semantically consistent results on CUB. These results confirm that ACGAN

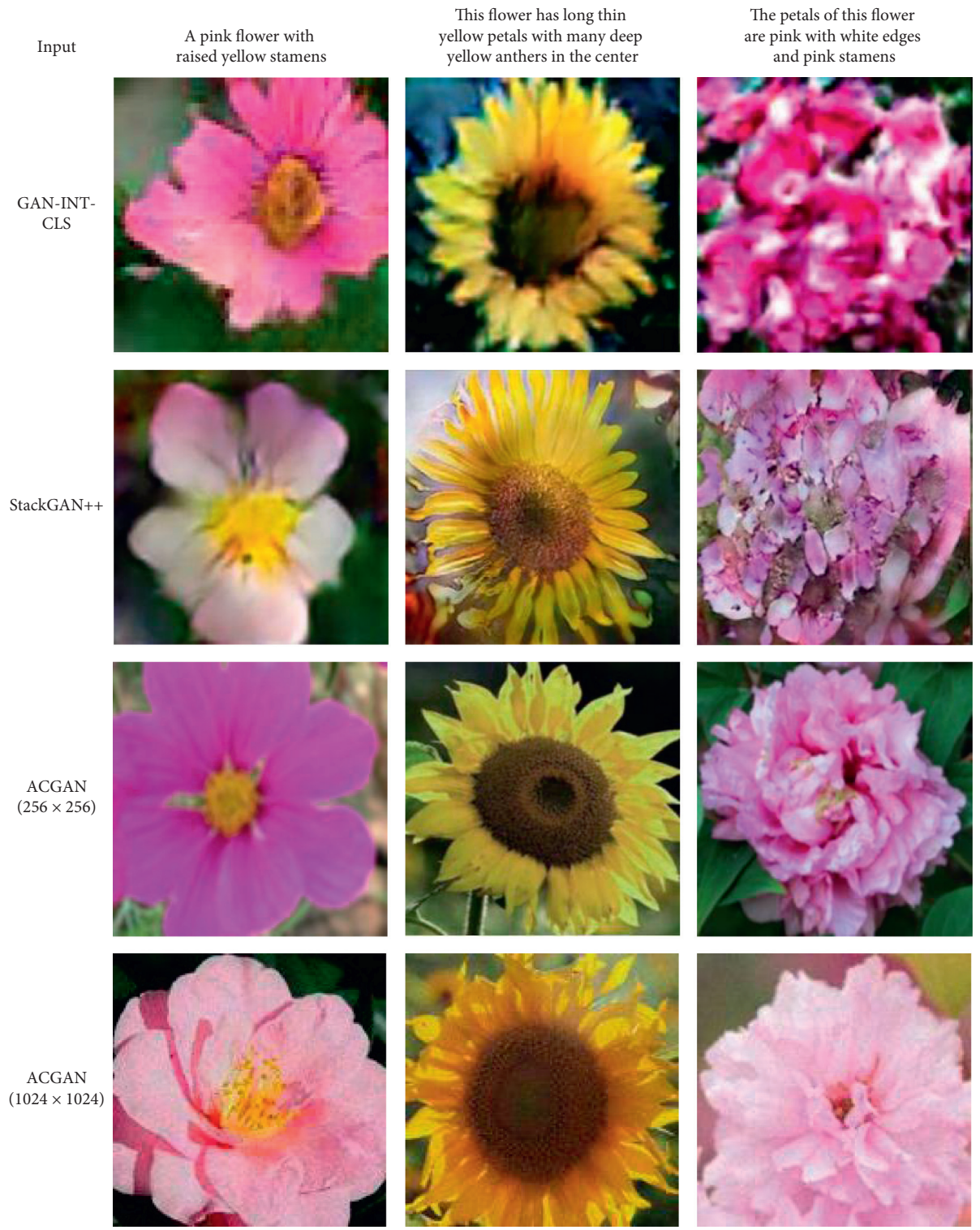

Figure 5: Images generated from descriptions using three GAN models trained on Oxford test set.

is better than AttnGAN, and the generated image is closer to the real image.

Detailed (beak, wings) comparisons of the results between the three models of StackGAN++, AttnGAN, and ACGAN on the CUB dataset are presented in Figure 4. It can be seen that the beak, wings, and feet of the bird are clearer, and the edges and details are more realistic in the images generated by the ACGAN in this paper. For example, the beak of a bird is more vivid and conforms to the text description in the 4th column. Compared with StackGAN++ and AttnGAN, it has achieved better results.

Figure 6: Details (petals) comparison of the images generated from descriptions using three GAN models trained on Oxford test set.

Subjective visual comparisons between the three models of GAN-INT-CLS, StackGAN++ and ACGAN on the Oxford dataset are presented in Figure 5. Details (petals) comparison of the results are presented in Figure 6. It can be seen that the image details generated by GAN-INT-CLS and StackGAN++ are lost, and the shape looks strange (1st and 2nd row) for some examples. ACGAN achieved better results with more details and consistent colors and shapes compared to StackGAN++. For example, the overall shape of the flowers is clearer, and the details of the petals are more obvious in the 4th row. These results confirm that ACGAN is better than StackGAN++, and the generated image is closer to the real image.

5. Conclusions

This paper adds attention mechanism and multilevel cascade structure to generate adversarial network, uses attention mechanism to pay attention to the fine-grained information of word level in semantics, enriches the details of generated images, and generates through cascade structure Higher resolution images. Experiments have shown that, on the same data set, the Attentional Concatenation Generative Adversarial Networks have clearer edge details and local textures against the image generated by the network, making the generated image closer to the real image. Although this method has achieved good results in generating images, it is still difficult to model complex scenes in life. How to deal with this problem needs further study. At the same time, the generated image is similar to the training data, lacking diversity. Therefore, it is intended to combine the zero shot learning and the generative adversarial networks to synthesize the new category image, which will be the focus of the next step.

Data Availability

The basic data used in this article was downloaded from the Internet. There are two-part datasets: (1) the CUB is a public dataset that can be downloaded from http://www.vision.caltech.edu/visipedia/CUB-200-2011.html. (2) The Oxford is a public dataset that can be downloaded from http://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Linyan Li and Yu Sun contributed equally to this work.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (61876121, 62002254, 61801323, and 62062003), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (19KJB520054), Research Fund

of Suzhou Institute of Trade and Commerce (KY-ZRA1805) , Primary Research and Development Plan of Jiangsu Province (BE2017663), Foundation of Key Laboratory in Science and Technology Development Project of Suzhou (SZS201609), and Graduate Research and Innovation Plan of Jiangsu Province (KYCX18_2549).

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the Advanced Neural Information Processing Systems (NIPS)*, pp. 2672–2680, Montreal, Canada, December 2014.

[2] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proceedings of the Advanced Neural Information Processing Systems (NIPS)*, pp. 217–225, Barcelona, Spain, October 2016.

[3] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *Proceedings of the Advanced International Conference on Machine Learning (ICML)*, pp. 1060–1069, New York, NY, USA, June 2016.

[4] H. Zhang, T. Xu, H. Li et al., "StackGAN++: realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[5] H. Zhang, T. Xu, H. Li et al., "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the Advanced International Conference on Computer Vision (ICCV)*, pp. 5907–5915, Venice, Italy, August 2017.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the Advanced International Conference on Learning Representations (ICLR)*, Banff, Canada, May 2014.

[7] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the Advanced International Conference on Machine Learning (ICML)*, pp. 1278–1286, Beijing, China, May 2014.

[8] A. Oord, N. Kalchbrenner, and N. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the Advanced International Conference on Machine Learning (ICML)*, pp. 1747–1756, New York, NY, USA, August 2016.

[9] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: multi-label image classification with visual attention and recurrent neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1971–1981, 2019.

[10] Y.-J. Cao, L.-L. Jia, Y.-X. Chen et al., "Recent advances of generative adversarial networks in computer vision," *IEEE Access*, vol. 7, pp. 14985–15006, 2018.

[11] S. Y. Zhao and J. W. Li, "Generative adversarial network for generating low-rank images," *Journal of Acta Automatica Sinica*, vol. 44, no. 5, pp. 829–839, 2019.

[12] H. Eghbal-zadeh, W. Zellinger, and G. Widmer, "Mixture density generative adversarial networks," in *Proceedings of the Advanced Computer Vision and Pattern Recognition (CVPR)*, pp. 5820–5829, Long Beach, CA, USA, November 2019.

[13] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proceedings of the Advanced Computer Vision and Pattern Recognition (CVPR)*, pp. 1155–1164, Long Beach, California, USA, April 2019.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proceedings of the Advanced International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, January 2016.

[15] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proceedings of the Advanced Neural Information Processing Systems (NIPS)*, pp. 2234–2242, Barcelona, Spain, June 2016.

[16] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *Proceedings of the Advanced International Conference on Learning Representations (ICLR)*, Toulon, France, May 2017.

[17] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," in *Proceedings of the Advanced International Conference on Learning Representations (ICLR)*, pp. 1–13, Toulon, France, March 2017.

[18] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug& play generative networks: conditional iterative generation of images in latent space," in *Proceedings of the Advanced Computer Vision and Pattern Recognition (CVPR)*, pp. 4467–4477, Honolulu, Hawaii, April 2017.

[19] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proceedings of the Advanced International Conference on Learning Representations (ICLR)*, Toulon, France, May 2017.

[20] J. Liu, C. Q. Gao, D. Y. Meng, and A. G. Hauptmann, "Decidenet: counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the Advanced International Conference on Computer Vision (ICCV)*, pp. 5197–5206, Santa Rosa, CA, USA, March 2018.

[21] X. N. Zhang, T. T. Wang, J. Q. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the Advanced International Conference on Computer Vision (ICCV)*, pp. 714–722, Salt Lake City, Utah, June 2018.

[22] J. W. Wang, W. H. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the Advanced International Conference on Computer Vision (ICCV)*, pp. 7190–7198, Salt Lake City, Utah, April 2018.

[23] P. Anderson, X. D. He, C. Buehler et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the Advanced International Conference on Computer Vision (ICCV)*, pp. 6077–6086, Salt Lake City, Utah, July 2018.

[24] H. J. Xu and K. Saenko, "Ask, attend and answer: exploring question-guided spatial attention for visual question answering," in *Proceedings of the Advanced European Conference on Computer Vision (ECCV)*, pp. 451–466, Amsterdam, Netherlands, March 2016.

[25] T. Xu, P. C. Zhang, Q. Y. Huang et al., "AttnGAN: fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the Advanced Computer Vision and Pattern Recognition (CVPR)*, pp. 1316–1324, Honolulu, Hawaii, November 2017.

[26] T. T. Qiao, J. Zhang, D. Q. Xu, and D. Tao, "MirrorGAN: learning text-to-image generation by redescription," in *Proceedings of the Advanced Computer Vision and Pattern Recognition (CVPR)*, pp. 1505–1514, Los Angeles, California, USA, March 2019.

[27] M. Abadi, P. Barham, J. Chen et al., "TensorFlow: a system for large-scale machine learning," in *Proceedings of the Advanced Operating Systems Design and Implementation (OSDI)*, pp. 265–283, Savannah, GA, USA, November 2016.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset: Computation & Neural Systems Technical Report*, California Institute of Technology, Pasadena, CA, USA, 2011.

[29] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the Advanced Computer Vision, Graphics & Image Processing (CVGIP)*, pp. 722–729, Marseille, France, December 2008.