IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2019, accepted January 3, 2020, date of publication January 9, 2020, date of current version January 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964946

Attentional Generative Adversarial Networks With Representativeness and Diversity for Generating Text to Realistic Image

ANJIE TIAN AND LU LU

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000, China

Corresponding author: Lu Lu (lul@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61370103, in part by the Guangdong Province Application Major Fund under Grant 2015B010131013, and in part by the Guangzhou Produce and Research Fund under Grant 201802020006.

ABSTRACT In recent years, with the emergence and rapid development of Generative Adversarial Networks (GANs), the generation of realistic images consistent with their semantics based on text description has become one of the most popular research directions in the field of computer vision. Although the idea of applying attention mechanism has been raised out in many implementation methods, it is required to treat the sub-regions of the generated images equally. For this reason, this paper proposes a novel generative adversarial networks, rdAttnGAN, which generate text to fine images by training multi-pair generators and discriminators. Comparing with the conventional models, it pays more attention to the representativeness and diversity of the generated images. In addition, an optimization method for calculating the similarity between the generated image and the text description is also introduced to enhance the representative judgment of the images. By paying more attention to the generation of important sub-regions of images, the model can further optimize the training of generators. In order to verify the effectiveness of our proposed framework, a comprehensive set of experiments are conducted on CUB dataset and COCO dataset. The results demonstrate viability to improve the representativeness and diversity of images with our rdAttnGAN.

INDEX TERMS GANs, representativeness and diversity, text-to-image generation.

I. INTRODUCTION

It is a very meaningful research to generate high-resolution and realistic images based on text descriptions. In industry, it not only provides assistance on a deeper visual understanding for the related research in the field of computer vision, but also has a wide range of realistic application. In photo editing and art designing, the results of this research supports the artists or designers to accelerate their work. Perhaps the algorithms in the video and image search engines that are now widely used will be replaced by this work one day. In academia, it has become one of the most popular research directions in the field of computer vision in recent years and has achieved remarkable results [1]–[8]. Recurrent Neural Networks (RNNs) and Generative Adversarial Networks (GANs) [9] are often combined to generate real images based on natural language descriptions. These methods have been able to produce satisfactory results in some areas, such as creating fine images of flowers or birds.

Although impressive results have been achieved, most of the existing image generation methodologies focus on global sentence vectors when training the conditional GANs. The helpful fine-grained image characteristics and word-level text information are ignored. At the same time, when evaluating the generated image, it is not considered that each sub-region of the image has a different influence on the entire image. Such methods can hinder the generation of high-quality images on the one hand, and reduce the diversity of generated images on the other hand. This problem becomes more serious when the scenes that need to be generated are more complex.

In order to solve these problems, inspired by AttnGAN [8] and Representativeness-Diversity Reward [10], we propose our rdAttnGAN framework, an Attentional Generative Adversarial Networks which pay more attention to the representativeness and diversity of generated images.

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

Such a network can focus on the sub-regions of the image, i.e. giving more attention to important sub-regions of the image and the sub-regions with rich content. Thus, according to the text description and the image details of the previous stages, more and more fine and diverse images are generated through multiple stages.

The main contributions of this paper are twofold. (i) An attentional GAN (rdAttnGAN) is proposed, which focuses on the representativeness and diversity of generated images. The network can recognize the significance and diversity of image sub-regions on top of word-level details in image generation. This plays an important role in synthesizing more realistic images from text descriptions. The image generated by our rdAttnGAN has more vivid details than the existing text-to-image generative models, and enhances the overall stability of the generated images. (ii) The calculation method of the similarity between the generated image and the text description is optimized to better provide additional support for training generators. The method presented in this paper has been qualitatively analyzed and quantitatively validated on two widely used public datasets. Compared with the original AttnGAN model, the inception score is increased by 3.78% on CUB [11] dataset and 4.39% on COCO [12] dataset. And the diversity of the generated images looks better, which can be verified in the experimental part.

The main structure of this paper consists of five parts. In addition to this chapter, Chapter 2 describes the work related to Generative Adversarial Networks, text-to-image generation techniques, and so on. In the third chapter, the structure of rdAttnGAN is introduced, and the realization and key algorithms of judging the representativeness and diversity of images are introduced in detail. The work related to the experiment is introduced in Chapter 4, including the evaluation of the model itself, and the qualitative analysis and quantitative comparison between the experimental results and other existing text-to-image generation models. The fifth chapter summarizes the content of the whole paper.

II. RELATED WORK

In this part, we first briefly introduce the related concepts of GANs, then show the research progress in the field of text-to-image generation, and finally share the research status of technologies related to our methods, namely diversity and representativeness.

A. GENERATIVE ADVERSARIAL NETWORKS

Recently, deep generative models have attracted widespread attention, including Variational Auto-encoders (VAE) [13], autoregressive approaches [14], GANs [9]. Compared with the other two deep generative models, the Generative Adversarial Networks (GANs) exhibits good performance [15], [16] in terms of generating clearer samples. It is a model put forward by Goodfellow *et al.* [9]. The original GAN model contains a generator and a discriminator. The generator is optimized to produce samples that are distributed toward real data to achieve the purpose of deceiving discriminator. The discriminator is trained in order to separate the true data distribution samples from the false samples generated by generator. There are a number of methods for studying how to better apply GANs to different fields, such as medical applications [17], [18], image synthesis [3], [16], [19], and even the art field (music and paintings generation) [20], [21].

There are many challenges in implementing GANs models. Most GANs models tend to learn only one data distribution mode, which is prone to mode collapse, that is, the generator will produce the same images every time. Although the images are clear, they have not variety. Another major challenge is that the instability of training process, and the loss obtained during training do not converge. In order to solve these problems, a wide range of schemes have been proposed, such as modifying the loss functions [22]–[24], modifying the architecture [25]–[27], modifying the optimizer [28], establishing a theoretical framework to analyze convergence and balance [23], [29], [30], etc. In addition, how to evaluate the generative models is also a very important task. In addition to the inception score proposed by Salimans *et al.* [22], Borji [31] also discussed the pros and cons of the evaluation indicators.

B. THE DEVELOPMENT OF GANS IN THE FIELD OF TEXT-TO-IMAGE GENERATION

Text-to-image generation is a very interesting research direction of GAN. Reed *et al.* [1] firstly propose a new training strategy for image-text matching, which can generate 64*64 resolution images. However, the images synthesized by this method not only lacks details and vivid parts of objects in many cases, such as the eyes and beaks of birds, but also has low resolution. In addition to the general problems mentioned in the first section of GANs training process, with the increase of the target image resolution, the difficulty of training will increase significantly.

In order to synthesize high-resolution images, many methods have been proposed. One is to add auxiliary information to the generator. Reed *et al.* [2] propose a generative adversarial what-where network (GAWWN). It shows that by adding additional conditions (such as the partial key points or bounding boxes of objects), the model can get position and content instructions to control the position of the object as it is generated, so as to improve the resolution of the image to a certain extent. Dash *et al.* [32] utilizes an auxiliary classifier (similar to [33]) to assist the training of GAN generators to achieve text-to-image synthesis. Cha *et al.* [34] discuss the usage of perceptual loss on CNN pre-trained by ImageNet.

Another way to improve image resolution is to decompose previous difficult tasks into multiple subtasks. Denton *et al.* [15] trained multiple GANs in the Laplacian Pyramid Framework (LAP-GAN). For each layer of the pyramid, the image output from the previous level is conditionally generated as a residual image and then added back to the input image to generate a new image of the next layer. However, this model can only successfully generate 96*96 low-resolution images. Inspired by the above idea, Chen and Koltun [4]

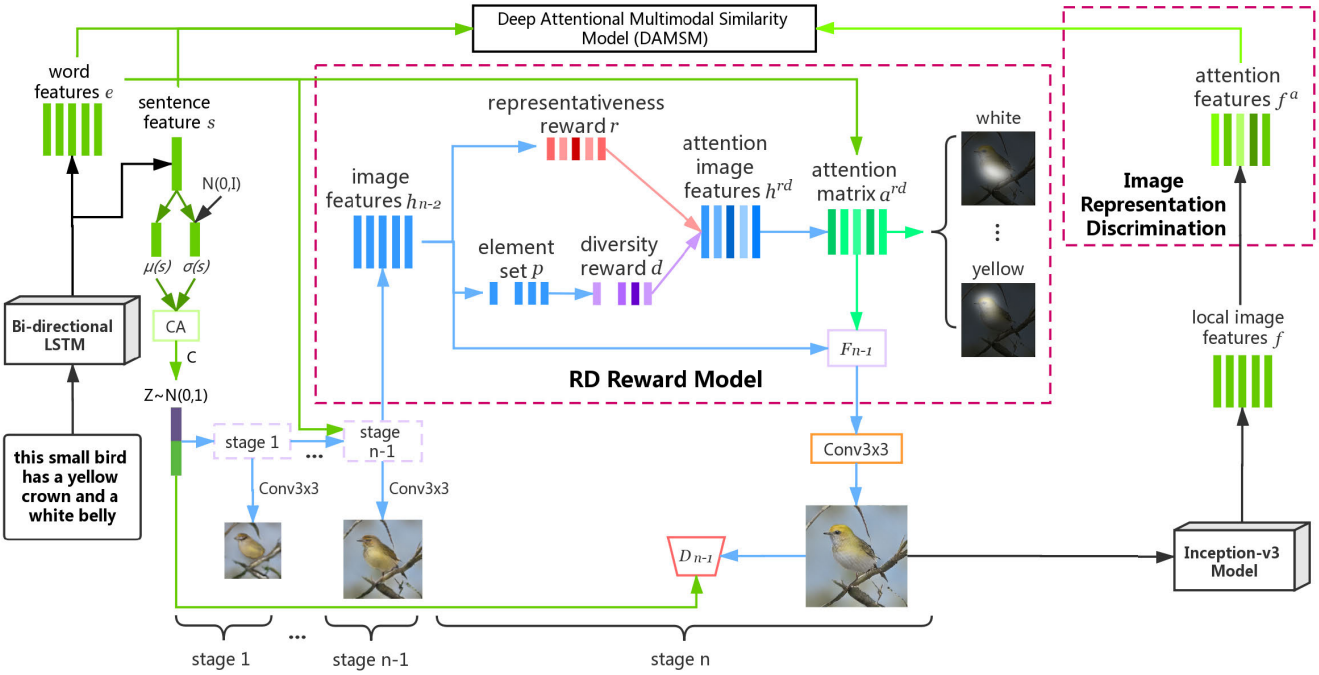

FIGURE 1. The architecture of our rdAttnGAN includes a generative network (bottom) and a DAMSM (top). The existence of the RD reward model in the generative network and the presence of image representational discrimination in the DAMSM is intended to focus more attention on the important and representative regions in which the image is generated.

present a cascaded refinement network for synthesizing high-resolution scenes from semantic maps. Subsequently, Johnson *et al.* [6] propose a method using scene graphs as an intermediate medium. First, the text description is converted into a scene graph, and then obtaining images based on the scene layout through cascaded refinement network.

Increasing the number of generator and discriminator pairs can also improve image resolution. Zhang *et al.* [3], [35] propose a multi-stage training method. It realizes text-to-image synthesis by stacking two GANs, and uses generators from different stages to generate images of different sizes. Finally, 256*256 compelling images can be generated. However, all GANs in these models are conditional on global sentence vectors, lacking word-level information for generating fine-grained images. Karras *et al.* [36] also suggest that GAN could be trained progressively to generate high-resolution images by gradually adding symmetrical generators and discriminator layers. Our approach also refers to this idea.

C. DIVERSITY AND REPRESENTATIVENESS

Attention mechanism is playing a more and more important role in both industry and academic area, especially in the area of sequence transduction models. Literature [37], [38] and [39] have successfully applied it in video classification, language understanding and representation, and financial data mining respectively. In the field of machine translation, research [40] has achieved advanced results by using only one attention model. Xu *et al.* [8] show a model consisting of multiple GANs stacked. Inspired by the idea of attention mechanism, an improved method is proposed to add attention-driven image details. This approach applied the attention mechanism for the first time to the GAN of text-to-image synthesis. However, this method only focused on multiple levels of text (such as word hierarchy and sentence hierarchy), but treated different sub-regions of the generated image equally when the generator is optimized. And we believe that the generation of certain important and representative areas in the image may need to be paid more attention to.

The research of video summarization provides new ideas for our method of optimizing text-to-image synthesis. Zhou *et al.* [10] design a reward function to consider the diversity and representativeness of generated summaries. The reward function judges the diversity and representativeness of the generated summaries, while DSN strives to obtain higher rewards by learning to generate more diverse and representative summaries. He *et al.* [41] use conditional feature selectors to guide GAN model and make it focus on the important time regions of the whole video frames. Inspired by these studies, we propose a method to evaluate the diversity and representativeness of different regions of the generated images.

III. MODEL

Fig. 1 shows the overall structure of our rdAttnGAN, which consists of an Attentional Generative Adversarial Network and a Deep Attentional Multimodal Similarity Model (DAMSM). An Representation-Diversity Reward Model (RD) is added to the generative network so that the generator can optimize key areas of the image while ensuring the diversity of the generated images. In the last stage of the

generative network, the image generated in the previous stage is processed by RD model, and the network can obtain an image representation vector and an image diversity vector. Combining these two vectors with the text context vectors generated by the attention model to form a multi-modal context vector, a new RD image sub-region feature can be generated. In addition, by identifying the representativeness of the generated image, DAMSM can evaluate the importance of the sub-regions of the generated image, thereby paying more attention to the generation effect of the important regions of the image. These two items will be described in detail in sections 3.2 and 3.3 of this chapter.

A. PRELIMINARIES

The Generative Adversarial Network (GAN) [9] contains a generator and a discriminator, which are trained alternately to compete with each other. The generator G is trained to generate an image, making it difficult for discriminator D to distinguish from the real image, and finally reproduces the true data distribution p_{data}. Meanwhile, the discriminator D is optimized to distinguish synthetic images from real images generated by G. The training process of GAN is usually similar to a two-party game which has the following objective function,

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))], \quad (1)$$

By executing Eq. (1), the best G and D can be generated eventually. Where z is a noise vector sampled from the prior distribution p_z (e.g., uniform or Gaussian distribution) and x represents the real image from the true data distribution p_{data}. In actual training, the training goal of generator $G$is not to minimize the value of $\log(1 - D(G(z)))$, but is modified to obtain the maximum value of $\log D(G(z))$. This can alleviate the problem of gradient vanishing [9]. We use this goal in all experiments.

Conditional GANs [42], [43] are an extension of GANs. Their generators and discriminators both receive additional conditioning variables c, resulting in $G(z, c)$ and $D(x, c)$. The condition GAN can be implemented to let generators to generate images conditioned by variables c.

B. GENERATIVE NETWORK WITH RD REWARD MODEL

In this section, we propose a new original representativeness-diversity reward model, which enables the generator to perform different processing on different sub-regions of the intermediate layer images, and finally generate new images according to the representativeness rewards and diversity rewards.

The proposed attentional generative network (shown in Fig. 1) has n stages, and each stage includes a generator $(G_0, G_1, \ldots, G_{n-1})$ and a discriminator $(D_0, D_1, \ldots, D_{n-1})$. These generators gradually generate images from small to large scales $(\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{n-1})$ by inputting the hidden states (or called image features) $(h_0, h_1, \ldots, h_{n-1})$.

In Fig. 1, z represents the noise vector that is typically sampled from a standard normal distribution. s represents the global sentence vector, while e represents the matrix of word vectors. CA is the Conditioning Augmentation [35] that adds conditions to s, which is a method used to enhance training data and avoid overfitting.

In the process of image generation, in order to enable the generator to focus on the generation of important image sub-regions, and take into account the diversity of the generated images, our method implements an RD reward model in the later stage of the network generation. Firstly, the output $h_{n-2} \sim N(\mu, \sigma^2)$ of the previous stage is processed to get p. Among them, p_i is the set of the remaining elements after removing part of the edge values from each image sub-region vector, $p_i \in (\mu - 3\sigma, \mu + 3\sigma)$. Then, the diversity reward matrix in the RD model can be obtained by the following formula, where $\overline{d}$ is the result of normalization.

$$d_i = \sigma_i^2, \quad where\ p_i \sim N(\mu_i, \sigma_i^2);$$
$$\overline{d}_i = \frac{\exp(d_i)}{\sum_{j=0}^{N-1} \exp(d_j)}. \quad (2)$$

Secondly, before calculating the attention image features h^{rd}, it is judged whether or not each of the sub-regions of the image feature h_{n-2} generated in the previous stage is important. This idea is implemented by the following formula.

$$r_i = \sum_{j=0 \cap j \neq i}^{N-1} \frac{h_i^T h_j}{\|h_i\| \|h_j\|},$$
$$\overline{r}_i = \frac{\exp(r_i)}{\sum_{j=0}^{N-1} \exp(r_j)}. \quad (3)$$

where $r \in \mathbb{R}^{N\times N}$ is the representativeness reward matrix with diagonal elements all zero. r_i is the degree of attention of the ith sub-region in the image features. In order to prevent over-fitting caused by excessive tensor, $\overline{r}_i$ is obtained by normalizing the representation matrix. After that, h_{n-2} is updated by the image attention weight w.

$$w_{ij} = \begin{cases} \overline{d}_i + \overline{r}_i, & i = j, \\ 0, & i \neq j. \end{cases} \quad (4)$$

According to the updated h_{n-2} and word features e, the context matrix a^{rd} of image features can be calculated. Then enter it with the image features into the final layer to obtain the final image. It is worth mentioning that in the generative network, on the one hand, the generation of high-resolution images requires more fine-grained details to be concerned, but the image output by the generators in the intermediate stages cannot provide sufficiently fine-grained details. On the other hand, if one step is added to calculate the image attention features at each stage, it will greatly increase the time it takes to train the model, especially when the model is trained on a large dataset, e.g. COCO dataset. Therefore, our method only implements the RD reward model in the final stage to optimize the rendering process of details.

Finally, in order to make the generated realistic image have not only word-level conditions but also sentence-level conditions, the objective function of rdAttnGAN is defined as the sum of generative network loss, CA loss and DAMSM loss.

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_{CA} + \lambda \mathcal{L}_{DAMSM},$$
$$where\ \mathcal{L}_G = \sum_{i=0}^{n-1} \mathcal{L}_{G_i}. \tag{5}$$

Here, λ is a weight that determines the ratio of the first two terms in the above formula to L_{DAMSM}. With the help of trained discriminators, the generators are optimized to minimize the loss of the attentional generative network by minimizing the $\mathcal{L}_G$ to jointly approximate the conditional distributions and unconditional distributions [3]. The adversarial loss of generator $\mathcal{L}_{G_i}$ is designed as

$$\mathcal{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log D_i(\hat{x}_i, s)] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log D_i(\hat{x}_i)] \tag{6}$$

The first term in Eq. (6) is conditional loss and the second term is unconditional loss. Whether the image and text description match is determined by conditional loss. And whether the image is real or fake is determined by unconditional loss. At the same time, the cross-entropy loss of the discriminator $\mathcal{L}_{D_i}$ is defined as

$$\begin{aligned}\mathcal{L}_{D_i} = &-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i, s)] \\ &-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i, s))] \\ &-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i)] \\ &-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - (D_i(\hat{x}_i)))],\end{aligned} \tag{7}$$

Similar to Eq. (6), the first two terms in Eq. (7) are conditional loss and the latter two are unconditional loss.

The second term of Eq. (5), $\mathcal{L}_{CA}$, is defined as the KL divergence between the standard Gaussian distribution and the Gaussian distribution of the training data. At the same time, in order to prevent this item from having too big or too small influence on the loss function, we also normalize the item to get NI.

$$\mathcal{L}_{CA} = NI, \quad where$$
$$NI_i = \frac{KL(\mathcal{N}(\mu(s), \sigma^2(s))||\mathcal{N}(0, I))_i}{\frac{1}{D}\sum_{j=1}^{D} KL(\mathcal{N}(\mu(s), \sigma^2(s))||\mathcal{N}(0, I))_j}, \tag{8}$$

The third term of Eq. (5), $\mathcal{L}_{DAMSM}$, represents the matching loss of image-text at the word level with the addition of image representative discrimination. It will be described in detail in Section 3.3.

C. RDATTNGAN WITH IMAGE REPRESENTATIVENESS DISCRIMINATION

After learning a text coder (bi-directional Long Short-Term Memory [44]) and an image coder (pre-trained Inception-v3 model [45], it is essentially a convolutional neural network), DAMSM can map words in a sentence and sub-regions of an image to a common semantic space. The fine-grained loss between the generated image and the textual description is then calculated by measuring the word-level image-text similarity. This loss can help optimize the training of generators.

$f \in \mathbb{R}^{768\times289}$ is defined as the local feature matrix generated by Inception-v3 model. Each column in the matrix represents a feature vector of a sub-region in the image. 768 and 289 represent the dimension of local feature vector and the number of sub-regions of the image, respectively. We judge the importance of the sub-region vectors in the extracted local feature matrix f by the following formula, so as to more significantly represent the generation effect of the important image sub-regions when calculating the loss of the DAMSM.

$$a_i = \sum_{j=0 \cap j \neq i}^{288} \frac{f_i^T f_j}{\|f_i\| \|f_j\|};$$
$$\bar{a}_i = \frac{\exp(a_i)}{\sum_{j=0}^{288} \exp(a_j)}. \tag{9}$$

Here, $a \in \mathbb{R}^{289\times289}$ is a matrix in which all diagonal elements are zero. We represent a_{ij} by calculating the cosine similarity between the *ith* sub-region and the *jth* sub-region. Then we similarly normalize the attention matrix. Thereafter, by using $f^a = f\bar{a}$, the attention matrix $\bar{a}$ is added to the local feature matrix f to measure the representation of the generated images.

Ultimately, the loss of DAMSM can be expressed as

$$\mathcal{L}_{DAMSM} = \alpha_1 \mathcal{L}_1^w + \alpha_1 \mathcal{L}_2^w + \alpha_2 \mathcal{L}_1^s + \alpha_2 \mathcal{L}_2^s, \tag{10}$$

where $\mathcal{L}_1^w$ and $\mathcal{L}_2^w$ are the loss functions calculated by the sub-regions of the image and the words of the sentence, while $\mathcal{L}_1^s$ and $\mathcal{L}_2^s$ are calculated by the global image vector $\bar{v}$ and the global sentence vector s. α_1 and α_2 correspond to the weights of the two loss functions, respectively. Because we believe that when calculating the matching loss of fine-grained image-text, the sentence-related and word-related loss functions should have different importance.

D. IMPLEMENTATION DETAILS

After repeated experiments, the hyperparameters in this section are set to $\alpha_1 = 1.1$, $\alpha_2 = 0.9$. The learning rate is set to 0.0002. At the same time, set $\lambda = 5$ on the CUB dataset to train 250 epochs, while set $\lambda = 50$ on the COCO dataset to train 120 epochs.

IV. EXPERIMENTS

Several experiments have been carried out to evaluate the image generation effect of rdAttnGAN, which emphasizes image representativeness and diversity. Firstly, we separately study the two components of rdAttnGAN: the generative network and the DAMSM. After that, we compare our model

FIGURE 2. Sample images of different scales by our proposed rdAttnGAN conditioned on text descriptions from CUB and COCO test sets.

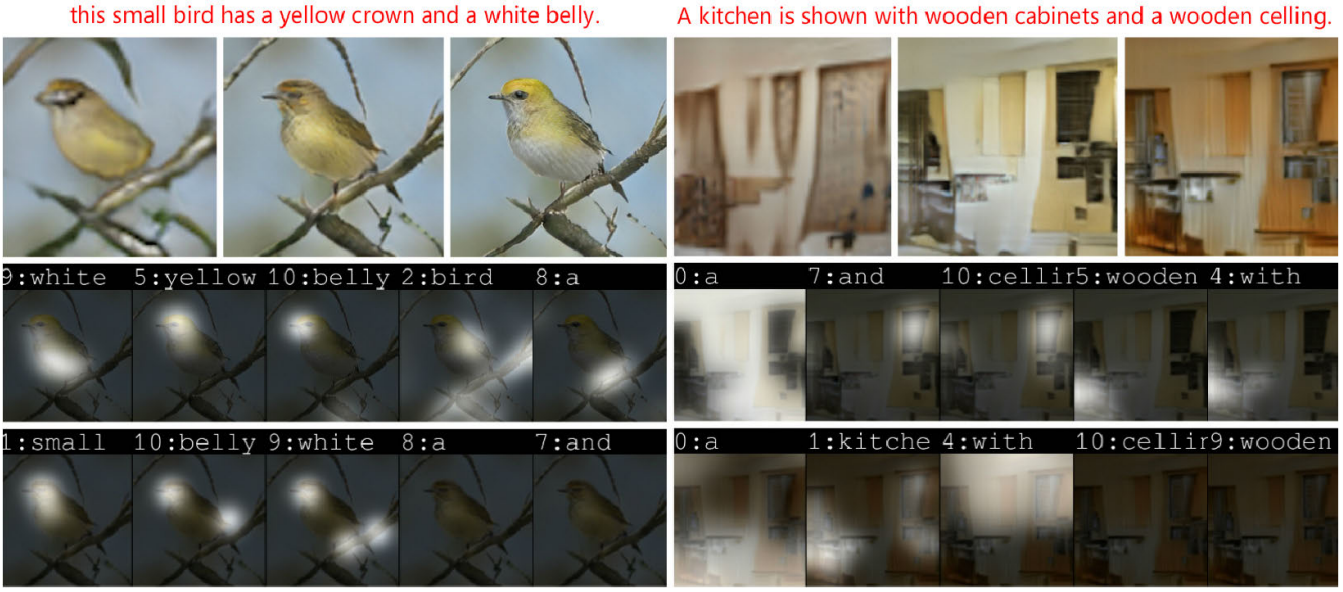

FIGURE 3. The intermediate results of our method on CUB and COCO test sets. In each section, the 64*64 image, 128*128 image, and 256*256 image output by rdAttnGAN are given in the first row by G_0, G_1, and G_2, respectively. The top 5 most-focused words in the last two stages of the rdAttnGAN are displayed in the second and third row, respectively.

with other models [1]–[3], [8], [35], [46], which are advanced GANs previously used for text-to-image synthesis, for qualitative analysis and quantitative results. In the experiments, our goal is to prove that our method can generate real images more efficiently and better.

Dataset: Like the dataset used in the previous text-to-image synthesis methods, our method is also validated on CUB [11] dataset and COCO dataset [12]. Table 1 shows the statistics for the two datasets.

Evaluation: Following the method of Zhang et al. [35], the inception score [22] is used by us as the quantitative evaluation measure for image generation. In order to compute the inception score, each model randomly selects invisible text descriptions and generates 30,000 images. Each candidate text description contains a ground truth and 99 mismatching descriptions of random selection.

TABLE 1. Statistics of CUB and COCO dataset.

dataset		train	test
CUB	#samples	8,855	2,933
	captions/image	10	10
COCO	#samples	82,783	40,504
	captions/image	5	5

In addition, we also perform qualitative examinations and analysis on samples generated by our method. Specifically, the intermediate results of the generative network are checked by the attention visualization (Fig. 3). Due to space limitations, only the top 5 most-focused words are displayed in each attention model.

FIGURE 4. The inception scores of our method (i.e., rdAttnGAN2) and the original AttnGAN at different epochs on the CUB (left) and COCO (right) test sets.

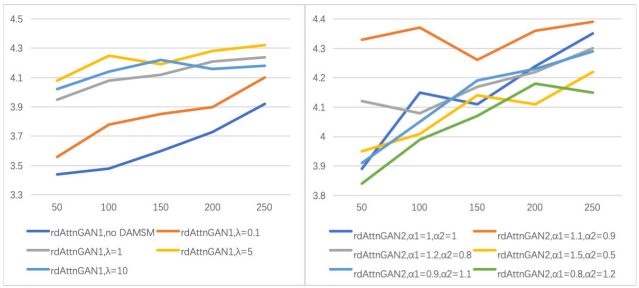

FIGURE 5. The inception scores of our method on the CUB dataset with different hyperparameters set. On the left is the result of the rdAttnGAN1 model, we set $\alpha_1 = 1$ and $\alpha_2 = 2$. On the right is the result of the rdAttnGAN2 model, we set $\lambda = 5$.

A. COMPONENT ANALYSIS

In this section, we perform quantitative and qualitative analysis of our methods separately. Firstly, rdAttnGAN and its variants are quantitatively evaluated (see Fig. 4,5 and Table 2 for the results). "rdAttnGAN1" in Table 2 represents an attention model in which two generators are stacked, while "rdAttnGAN2" represents a structure that contains two attention models and three generators stacked (as shown in Fig. 1). In addition, as illustrated in Fig. 2, Fig. 3 and Fig. 6, the images generated by our method are qualitatively checked, and the effects of image generation are also evaluated.

In order to test the proposed $\mathcal{L}_{DAMSM}$ emphasis on image representativeness, we adjust the value of λ in Eq. (5) and the values of α_1, α_2 in Eq. (10). "rdAttnGAN1" is used to search for the best value of λ while "rdAttnGAN2" model is used to determine the proportion of word loss and sentence loss in DAMSM loss, and ultimately generate more detailed images. To ensure the quality of image generation as much as possible, we find its optimal value by increasing the value of λ until the inception score of the model begins to decline on the validation set. Through experiments we can find that a larger λ can significantly improve inception score (see Table. 2 and Fig. 5). When increasing the value of λ from 0.1 to 5, the inception score of the rdAttnGAN1 can be found to be increased from 4.10 to 4.32 on CUB dataset. At the same time, as the α_1 value increases and the α_2 value decreases in rdAttnGAN2, the inception score on CUB dataset has a certain increase from 4.35 to 4.39.

TABLE 2. The inception score of each rdAttnGAN model on the CUB test set (first 7 lines) and coco test set (last line).

Method	Inception Score
rdAttnGAN1, no DAMSM	3.92 ±.04
rdAttnGAN1, $\lambda = 0.1$, $\alpha_1 = 1$, $\alpha_2 = 1$	4.10 ±.04
rdAttnGAN1, $\lambda = 1$, $\alpha_1 = 1$, $\alpha_2 = 1$	4.24 ±.04
rdAttnGAN1, $\lambda = 5$, $\alpha_1 = 1$, $\alpha_2 = 1$	4.32 ±.04
rdAttnGAN1, $\lambda = 10$, $\alpha_1 = 1$, $\alpha_2 = 1$	4.18 ±.05
rdAttnGAN2, $\lambda = 5$, $\alpha_1 = 1.1$, $\alpha_2 = 0.9$	**4.39 ± .03**
rdAttnGAN2, $\lambda = 5$, $\alpha_1 = 1.5$, $\alpha_2 = 0.5$	4.22 ±.03
rdAttnGAN2, $\lambda = 50$, $\alpha_1 = 1.1$, $\alpha_2 = 0.9$	**26.39 ± .41**

These comparisons on the one hand indicate that an appropriate increase in the weight of the $\mathcal{L}_{DAMSM}$ helps to generate higher quality images using given textual descriptions. On the other hand, it also shows that under the condition of strengthening the important regions of the image, it is more helpful to optimize the generation of image details by appropriately increasing the proportion of the fine-grained detail loss of the word level. Besides, owing to the limitation of GPU memory, we did not build another model which contains more attention models.

Next, the qualitative analysis of our method is carried out. The intermediate results of the method are shown in Fig. 3. Because only global sentence vectors are used as input and the details described by words are lacking in the first stage, the output image resolution is low and can only show the outline and colors of objects roughly. The G_1 and G_2 stages add more details by adding word vectors to generate higher resolution images. In the latter two stages, some sub-regions of images can be inferred from the output image of the previous stage. For these sub-areas, reduce their importance and assign the same attention to all words. Such sub-regions are shown in black in the intermediate result graph. For other sub-areas that have corresponding semantic meanings in the text description, attention is assigned to the words most relevant to them. These sub-regions are shown as bright areas in Fig. 3. The remaining unrecognized sub-regions in the image are not representative regions (such as background regions), and their representation is also reduced.

According to the intermediate results produced by the model, the words "bird" and "this" are commonly used to locate objects on CUB dataset. While some words that describe the properties of objects, such as the color of a bird and the size of the wings, are used to improve the detail of image drawing. These observations on the one hand prove that rdAttnGAN can truly understand the detailed semantics of the text description, while reducing the attention to some less important sub-regions in the image. Another conclusion of the observations is that the latter attention model can deal with some new words missing from the previous attention model. This shows that rdAttnGAN can provide richer information in the later stage and generate higher resolution images.

B. COMPARISON WITH PREVIOUS METHODS

We use a variety of methods to generate images on the CUB and COCO test sets based on textual descriptions to

FIGURE 6. A comparison of the sample images generated by the original AttnGAN and our rdAttnGAN on the CUB test set. Among them, the upper part of each module is the result of AttnGAN's, and the lower part is the image generated by our method. It can be seen that the results of our method are richer in detail and more diverse.

TABLE 3. Inception scores by previous advanced GANs models [1]–[3], [8], [35] and our method on CUB and COCO test sets.

GANs Models	Inception Score	
	CUB	COCO
GAN-INT-CLS [1]	2.88 ±.04	7.88 ± .07
GAWWN [2]	3.62 ±.07	/
StackGAN [35]	3.70 ±.04	8.45 ± .03
StackGAN++ [3]	3.82 ±.06	/
AttnGAN [8]	4.23 ±.03	25.28 ± .45
Our rdAttnGAN	**4.39 ±.03**	**26.39 ±.41**

compare our rdAttnGAN with the previous GANs models. As shown in Table 3, compared with the previous optimal inception score of 4.23 on CUB test set, our method obtains an inception score of 4.39 under the same experimental environment, while the inception score on COCO test set also increases from 25.28 to 26.39. The experimental results show that rdAttnGAN, which pays more attention to the representativeness and diversity of the generated images, can generate high-resolution images with fine details more efficiently. This is because it applies a new RD reward model, which enables the generative network to better capture more representative sub-region levels and fine word-level information from text to image in the process of image generation. Moreover, by comparing the calculated standard deviations of the output results, we can find that the image generated by our method has smaller standard deviation, which indicates that the image generated by our method is more stable.

Not only that, by comparing the image generated by our method with those generated by the original AttnGAN (as shown in Fig. 6), it is easy to see that the content of the image generated by our method is more abundant. This proves that the images generated by our method have better diversity. And because the RD reward model is only added in the last stage of the generative network, it does not have a great impact on the training efficiency of the entire method. Under the same hardware environment, the training time of our method is only less than 2% longer than the original AttnGAN. This shows that our method can generate higher quality images without affecting the efficiency.

V. CONCLUSION

In this paper, an attentional generative adversarial network, which pays more attention to the representativeness and diversity of generated images, is proposed to realize fine image synthesis based on textual descriptions. Firstly, the RD reward model is implemented in the attentional generative network for generating high-precision images by optimizing more important and representative regions in the image. Secondly, the characterization evaluation of each sub-region of the image is introduced when calculating the word-level image-text matching loss to optimize the training of rdAttnGAN generators. Our method has been proved to be superior to previous models. The inception score has been increased by 3.78% on CUB test set and 4.39% on COCO test set. In addition, compared to the original AttnGAN, the images generated by our method are also more diverse. A large number of experimental results demonstrate the validity of the concept of image representativeness and diversity proposed in this paper in rdAttnGAN.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: https://arxiv.org/abs/1605.05396

[2] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[4] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1511–1520.

[5] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5630–5639.

[6] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[7] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[10] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep., 2011.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114

[14] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1747–1756.

[15] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[16] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[17] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.

[18] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[19] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[20] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[21] Y. Liu, Z. Qin, T. Wan, and Z. Luo, "Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks," *Neurocomputing*, vol. 311, pp. 78–87, Oct. 2018.

[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[23] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3478–3487.

[24] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving gans using optimal transport," 2018, *arXiv:1803.05573*. [Online]. Available: https://arxiv.org/abs/1803.05573

[25] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. Torr, and P. K. Dokania, "Multi-agent diverse generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8513–8521.

[26] C. Wang, C. Xu, X. Yao, and D. Tao, "Evolutionary generative adversarial networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 6, pp. 921–934, Dec. 2019.

[27] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Bidirectional conditional generative adversarial networks," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 216–232.

[28] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training GANs with optimism," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[29] S. Arora, A. Risteski, and Y. Zhang, "Do GANs learn the distribution? Some theory and empirics," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[30] F. Farnia and D. Tse, "A convex duality framework for GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5248–5258.

[31] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understand.*, vol. 179, pp. 41–65, Feb. 2019.

[32] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN-text conditioned auxiliary classifier generative adversarial network," 2017, *arXiv:1703.06412*. [Online]. Available: https://arxiv.org/abs/1703.06412

[33] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 2642–2651.

[34] M. Cha, Y. Gwon, and H. T. Kung, "Adversarial nets with perceptual losses for text-to-image synthesis," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[35] H. Zhang, T. Xu, and H. Li, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[36] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: https://arxiv.org/abs/1710.10196

[37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[39] W. Jiang, L. Xu, J. Yu, and G. Zhang, "Research and application of mapping relationship based on learning attention mechanism," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2018, pp. 310–321.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[41] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, 2019, pp. 2296–2304.

[42] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Convolutional Neural Netw. Visual Recognit.*, vol. 2014, no. 5, p. 2, 2014.

[43] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784

[44] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[46] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4467–4477.

ANJIE TIAN is currently pursuing the master's degree with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include text-to-image generation and computer vision.

LU LU received the Ph.D. degree from Xi'an Jiaotong University, in 1999. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, China. His main research interests are software engineering, software testing, and software architecture design.