ADAPTIVE FORGETTING, DRAFTING AND COMPREHENSIVE GUIDING: TEXT-TO-IMAGE SYNTHESIS WITH HIERARCHICAL GENERATIVE ADVERSARIAL NETWORKS

Yuting Xue[1], Heng Zhou[1, 2], Yuxuan Ding[1]*, Xiao Shan[1]

[1]School of Electronic Engineering, Xidian University, Xi'an, China
[2]Institute of Systems Engineering, AMS, Beijing, China

ABSTRACT

The generation task from text to image generates cross modal data with consistent content by mining the semantic consistency contained in two different modal information of text and image. Due to the differences between the two modes, the task of text to image generation faces many difficulties and challenges. In this paper, we propose to boost the text-to-image synthesis through an adaptive learning and generating generative adversarial networks (ALG-GANs). First, we propose an adaptive forgetting mechanism in the generator to reduce the error accumulation and learn knowledge flexibly in the cascade structure. Besides, to evade the mode collapse caused by a strong biased surveillance, we propose a multi-task discriminator using weak-supervision information to guide the generator more comprehensively and maintain the semantic consistency in the cascade generation process. To avoid the refine difficulty aroused by the bad initialization, we judge the quality of initialization before further processing. The generator will re-sample the noise and re-initialize the bad initializations to obtain good ones. All the above contributions have been integrated in a unified framework, which is an adaptive forgetting, drafting and comprehensive guiding based text-to-image synthesis method with hierarchical generative adversarial networks. The model is evaluated on the Caltech-UCSD Birds 200 (CUB) dataset and the Oxford 102 Category Flowers (Oxford) dataset with standard metrics. The results on Inception Score (IS) and Fréchet Inception Distance (FID) show that our model outperforms the previous methods.

KEYWORDS

Text-to-Image Synthesis, Generative Adversarial Network, Forgetting Mechanism, Semantic Consistency.

1. INTRODUCTION

In the past few years, generative adversarial networks (GANs) [1] have boomed in the deep learning tasks. Various kinds of GANs [2], [3], [4] have brought amazing results on generating natural images through random noise. In order to generate images met the desire of users, conditional generative adversarial network (CGAN) [5] sets a condition as the target for the generator. deep convolutional generative adversarial network (DCGAN) [6] combines GANs with convolutional neural network (CNN) to improve the quality of generated images. To incorporate more accurate surveillance, auxiliary classifier generative adversarial network (ACGAN) [7] requires the discriminator to output both probability and fine-grained category.

David C. Wyld et al. (Eds): EMSA, SEA, AIFU, NLCAI, BDML, BIoT, NCOM, CLOUD, CCSEA, SIPRO - 2022
pp. 273-285, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120623

Although significant progress has been made in generating visually realistic images, generating images that match the given text descriptions is still challenging. The conditional align Deep Recurrent Attention Writer (alignDRAW) [8], which is the first text-to-image generation model extending DRAW [9] to generate images from texts. However, the synthesized images are blurred with a low-resolution of 36x36. Then, Reed et al. [10] introduce GAN to text-to-image task, which follows DCGAN and CGAN to generate images from texts. As the training process is not stable, GAN-INT-CLS only generates plausible images for birds and flowers. To reduce the unstable of the training process, the popular text-to-image generation methods [11], [12], [13] mainly apply a multi-stage generator to supplement more restrictions. However, there are still three issues in the multi-stage structure. First, the cascade structure accumulates the incorrect and redundant information during the generation process. Second, the output of discriminators is the probability of reality which cannot guide the generator comprehensively. Finally, bad initialization of images has unclear parts which make the refine difficulty.

To address these issues, we propose a new text-to-image synthesis model called ALG-GAN. In the process of cognition, we ignore the redundant and incorrect information to learn and summarize knowledge more efficiently. Inspired by this, we incorporate a pair of down-sampling and up-sampling convolutional layers to the up-block to bring in the forgetting process. Thus, the model has opportunity to throw the useless and improper information away, which is called forgetting mechanism (FM). For the second problem, we design a multi-task discriminator (MTD) to fully utilize the additional weak-supervision information in the training process, which can help the discriminator to guide the generator in detail. For the last problem, multi-stage generators start from the initial images to synthesize larger images, regardless the quality of initial images. However, people do not perform in this way. Good painting usually starts from a satisfied draft. Thus, we propose to supervise the generation process of small image to guarantee its quality. It is called drafting mechanism (DM).

The main contributions of this paper are summarized as follows:

- Forgetting mechanism: we propose a down-up sampling dual structure, which allows the network to forget information during the generation process.
- Comprehensive guiding discriminator: it guarantees the comprehensive guidance by additional weak-supervision.
- Drafting mechanism: we supervise the generation process through discriminator to guarantee the quality of initialization.

We conduct experiments to evaluate the proposed ALG-GAN model on the Caltech-UCSD Birds 200 (CUB) dataset [14] and the Oxford 102 Category Flowers (Oxford) dataset [15]. The quality of generated images is measured using the inception score (IS) [16] and the Fréchet inception distance (FID) [17]. The experimental results indicate that our ALG-GAN model performs better than the state-of-the-art text-to-image synthesis methods. On CUB, we improve IS from 4.36 to 4.62. FID decreases from 16.899 to 16.500. On Oxford, the result of IS is 4.10. FID achieves 44.307. It proves that our model generates more realistic images.

The remainder of this article is organized as follows. In Section II, the related works of image generation are introduced from two aspects: multi-stage generator and multi-task discriminator. In Section III, we introduce AttnGAN as the baseline of our model. Section IV, the text-to image generation model ALG-GAN we proposed is introduced in detail. ALG-GAN mainly includes three parts: an adaptive forgetting and drafting generator and a comprehensive guiding discriminator. In Section V, compared with the state-of-the-art methods on the public dataset, it shows that this method has superior performance. The effectiveness of ALG-GAN is proved by a large number of ablation experiments. The conclusions and future work are shown in Section VI.

2. RELATED WORKS

Our work mainly refers to the hierarchical structure of generator and the auxiliary task of discriminator. I will introduce the related work from the following two aspects.

2.1. Multi-Stage Generator

In the text-to-image task, Reed et al. [10] propose a structure called GAN-INT-CLS based on CGAN and DCGAN, which uses sentence embedding as the condition to generate images. Since it is difficult to control the generation process, generated images are small with 64×64 resolution which lack details and easily suffer from mode collapse. To address the size limitation, Zhang et al. [11], [12] extend single-stage to multi-stage in their StackGAN and StackGAN++. The stacking structure makes the generation process more controllable to synthesize large images. Considering that the word level information can guide the local information generation, Xu et al. [13] introduce AttnGAN with deep attentional multimodal similarity model (DAMSM) to refine images which takes into account of both local and global information. However, the multi-stage structure has pros and cons. On the one hand, stacking structure addresses the size limitation by means of bit by bit surveillance. We also incorporate drafting mechanism in the stack structure to guarantee a better initialization. On the other hand, multi-stage incorporates restrictions in each stage. However, the biased surveillance from the discriminator at each stage will be accumulated. We propose a forgetting mechanism to promote the model learn adaptively. Meanwhile, the model is prone to collapse when it is supervised by a biased strong surveillance. Thus, building a comprehensive surveillance is also crucial for the model to be success.

2.2. Multi-Task Discriminator

The original discriminator only judges the reality of input images. To guide the generator better, August us et al. [7] propose ACGAN by adding an auxiliary classifier to the discriminator of GAN, which achieves the start-of-the-art results. It confirms that additional classification task in the discriminators can guide a better generator. Following ACGAN, Ayushman et al. [18] propose TAC-GAN where the auxiliary classifier classifies the category of birds. It obtains good results. Following TAC-GAN, Cha et al. [19] propose Text-SeGAN adding a semantic classifier to guarantee the semantic consistency. Following TAC-GAN, we assume that not only category label, but also some other fine-grained weak-supervision information like attributes also meet this end.

3. THE PRELIMINARY: ATTENTION BASED HIERARCHICAL GENERATIVE ADVERSARIAL NETWORK

The text-to-image model AttnGAN [13] consists of an attention based hierarchical generator G and a discriminator D. G has two main components: The initialization and DAMSM based up-block. In the initialization, firstly, the input text description is transformed by a text encoder into the word-level representations and a global feature, which is used as the sentence condition. Then, G predicts the rough sketch of image I0according to a random noise vector with the sentence condition after conditioning augmentation (CA). The noise vector is normally distributed. After initialization, more fine-grained visual contents are supplemented to the initial image by up-block, which makes it more photo-realistic. D distinguishes not only the real data from synthesized images, but also the matched sentence conditions from mismatched conditions. During training, G and D are following the two-player min-max game with value function : $V(G,D)$:

$$\min_G \max_D V(D,G) = \mathrm{E}_{I\sim p_{\text{data}}}[\log D(I)] + \mathrm{E}_{I\sim p_{\text{data}}}[\log D(I,\bar{e})] + \mathrm{E}_{\hat{I}\sim G}[\log(1-D(\hat{I}))] + \mathrm{E}_{\hat{I}\sim G}[\log(1-D(\hat{I},\bar{e}))] + L_{CA} + L_{DAMSM}$$

where E means the expectation. $D(I)$ and $D(\hat{I})$ compute the probability of reality, while $D(\hat{I},\bar{e})$ and $D(I,\bar{e})$ compute the probability of matching between text and image. L_{CA} represents the K-L divergence between the standard Gaussian distribution. L_{DAMSM} measures the correlation between images and corresponding text descriptions. Both L_{CA} and L_{DAMSM} are only related to *G*.

4. THE PROPOSED ALG-GAN

Our text-to image generation model ALG-GAN is demonstrated in Figure 1, which consists of an adaptive forgetting and drafting generator and a comprehensive guiding discriminator.

4.1. Adaptive Forgetting and Drafting Generator

To address the cons of hierarchical structure, we incorporate two new mechanism in our generator.

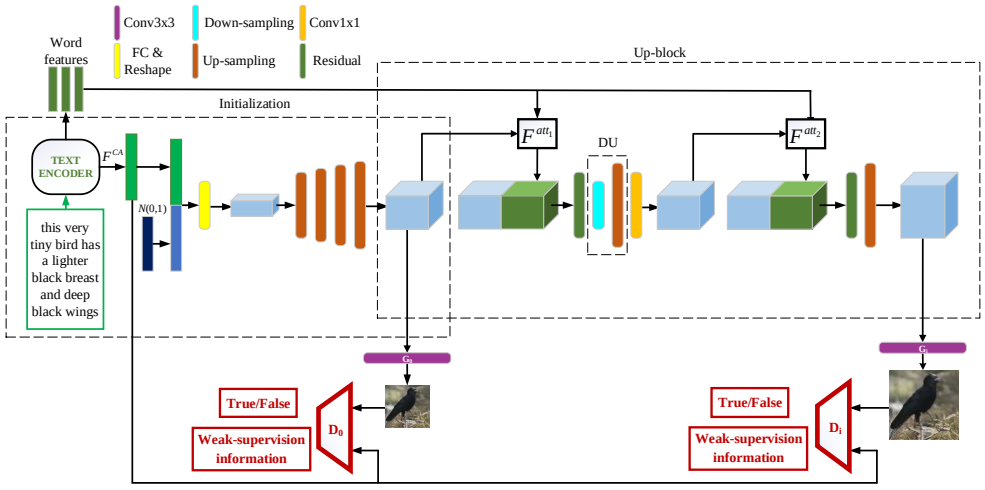

Figure 1: The architecture of the proposed ALG-GAN

Forgetting Mechanism To handle the redundant and incorrect information during the generation process, we propose a novel up-block module, which utilizes a down-up sampling dual structure (DU) to learn information adaptively.

Redundant information arises in the feature map after we first refining images with DAMSM $F_i^{att_1}$ in the up-block. So, we utilize the down-sampling operation to forget some information. To keep the size, we combine the down-sampling operation with an up-sampling operation. Then, we utilize the second DAMSM $F_i^{att_2}$ to enhance the learnt knowledge. The new image features are obtained by:

$$
\begin{aligned}
h_i^{temp} &= F_{DU}\left(h_i, F_i^{att_1}\left(e, h_i\right)\right) \\
h_i^{new} &= \phi\left(h_i^{temp}, F_i^{att_2}\left(h_i^{temp}, e\right)\right)
\end{aligned}
\tag{2}
$$

where h_i is image feature, $F_{DU}(\cdot)$ represents down-up sampling operation, $\phi(\cdot)$ is implemented as a residual module with 1×1 convolution to adjust the number of channels.

Drafting Mechanism In view of the painting process of human being, generator just likes a painter, guided by the discriminator. Actually, even professional painters still make drafts. They will judge their draft before they draw further. Such a judgment is also necessary in our generation process. However, it is difficult for the generator to judge the quality of the initialized image. We use the discriminator to conduct this task. In our DM, when the initialization of the image does not reach the judgment, the model will re-sample the noise and re-initialize the image until the requirements are met. Compared with current works, we are the first one who supervises the generation process.

See Figure 2, our generation process is defined as follows:

$$
\begin{aligned}
&h_0 = re\left(F_0\left(z, F^{CA}(\bar{e})\right)\right) \ \ if\ CE\left(D_0^{fe}\left(\hat{I}, \bar{e}\right), 1\right) > \alpha \\
&h_i^{temp} = F_{DU}\left(h_i, F_i^{att_1}\left(e, h_i\right)\right) \\
&h_i^{new} = \phi\left(h_i^{temp}, F_i^{att_2}\left(h_i^{temp}, e\right)\right) \\
&h_{i+1} = F_{i+1}\left(h_i^{new}\right) \\
&I_i = G_i\left(h_i\right),\ for\ \ i = 0,1,2,...,m-1
\end{aligned}
\tag{3}
$$

where F^{CA} means CA operation. $re(\cdot)$ donates re-sampling the noise and re-initializing the image. When the cross entropy $CE(\cdot)$ between the output of final epoch discriminator $D_0^{fe}(\hat{I}_0, \bar{e})$ and real label is larger than α, the model will re-initialize the image.

4.2. Comprehensive Guiding Discriminator

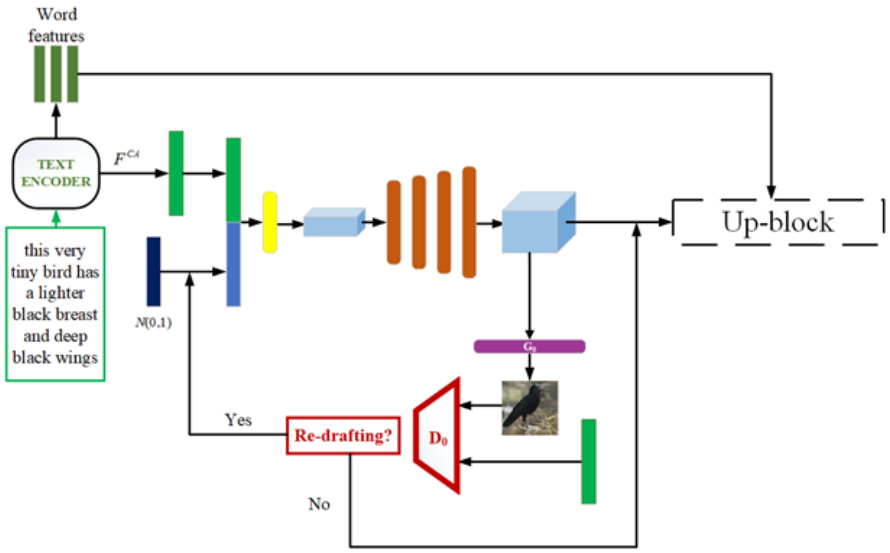

Figure 2: The generation process of ALG-GAN. We apply the final epoch discriminator to conduct the re-drafting judgment.

As mentioned in Preliminaries, the traditional text-to-image discriminators only compute the probability of reality and matching, which cannot guide the generator comprehensively. To address this, we propose a comprehensive guiding discriminator to guarantee the semantic consistency on the generated image during the generation process through the additional weak-supervisions such as category label, the fine-grained attribute label etc. The comprehensive guiding discriminator is shown in Figure 1. It has an auxiliary task of predicting the weak-supervision information on the synthesized and real data. Thus, discriminators give comprehensive surveillance rather than a biased one on the generator. The outputs of discriminators are defined as

$$\begin{cases} p_u = D_i^{GAN}(I) \\ p_c = D_i^{GAN}(I, \bar{e}) \\ p_{ws} = D_i^{ws}(I, \bar{e}) \end{cases} \tag{4}$$

where D is implemented as a discriminator. p_u means predicting the image whether it is real or synthesized. p_c measures the matching of image and sentence. p_{ws} discriminates the additional weak-supervision information on images.

4.3. Objective Function

We define our objective function as following.

$$L = \sum L_i^{GAN} + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM} + \lambda_3 \sum L_i^{ws} \tag{5}$$

In Eq.(5), L_i^{GAN} is an adversarial loss. It is related to GAN [1]. λ are the corresponding weights of CA loss L_{CA}, DAMSM loss LDAMSM and weak-supervision loss L_{ws}.The definition of L_{CA} follows StackGAN++ [12]. The definition of L_{DAMSM} follows AttnGAN [13]. L^{ws} is defined as Eq.(6), which guarantees the comprehensive surveillance of discriminator.

$$L_i^{ws} = CE\left(p_{ws}, l_{re}\right) \tag{6}$$

In Eq.(6), l_{re} is real label of weak-supervision information. *CE*() is implemented as cross entropy. The discriminators minimize the training objective function L_D, which is defined as follows. The subscript denotes the index of the discriminator. Table 1 shows the inputs to the discriminators.

$$L_{D_i} = L_{D_i}^{GAN} + \lambda_3 L_i^{ws} \tag{7}$$

$$\begin{aligned} L_{D_i}^{GAN} = & -\frac{1}{2}\mathrm{E}_{I_i \sim p_{data}}\left[\log D_i\left(I_i\right)\right] - \frac{1}{2}\mathrm{E}_{\hat{I}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{I}_i\right)\right)\right] \\ & -\frac{1}{3}\mathrm{E}_{I_i \sim p_{data}}\left[\log D_i\left(I_i, \bar{e}\right)\right] - \frac{1}{3}\mathrm{E}_{\hat{I}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{I}_i, \bar{e}\right)\right)\right] \\ & -\frac{1}{3}\mathrm{E}_{I_i \sim p_{data}}\left[\log\left(1 - D\left(I_i, \hat{\bar{e}}\right)\right)\right] \end{aligned} \tag{8}$$

$$L_i^{ws} = CE\left(D_i^{ws}\left(I, \bar{e}\right), l_{re}\right) + \frac{1}{2}CE\left(D_i^{ws}\left(\hat{I}, \bar{e}\right), l_{re}\right) + \frac{1}{2}CE\left(D_i^{ws}\left(I, \hat{\bar{e}}\right), l_{re}\right) \tag{9}$$

The first row of $L_{D_i}^{GAN}$ means the unconditional loss, which distinguishes the real images from synthetic images. Another row is the conditional loss, which measures whether the text and image are matching. Notice that discriminators should minimize the matching probability of real image with mismatching text pair.

Table 1. The meaning of inputs for the discriminators.

Input	Meaning
I	Real image
$\hat{I}$	Generated image
$(I, \bar{e})$	Real image with matching text
$(\hat{I}, \bar{e})$	Generated image with corresponding text
$(I, \hat{\bar{e}})$	Real image with mismatching text

When we optimize the generators, there are no real or mismatched images for discriminators to be fed as input. The generators should minimize the unconditional loss and conditional loss $L_{G_i}^{GAN}$

to synthesize the real and matching images. The subscript denotes the index of the generator. The training objective L_G is defined as follows.

$$
\begin{aligned}
L_{G_i} &= L_{G_i}^{GAN} + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM} + \lambda_3 L_i^{ws} \\
L_{G_i}^{GAN} &= -\frac{1}{2}\mathrm{E}_{\hat{I}_i \sim p_{G_i}}\left[\log\left(D_i(\hat{I}_i)\right)\right] - \frac{1}{2}\mathrm{E}_{\hat{I}_i \sim p_{G_i}}\left[\log\left(D_i(\hat{I}_i, \bar{e})\right)\right] \\
L_i^{ws} &= CE\left(D_i^{ws}\left(\hat{I}_i, \bar{e}\right), l_{re}\right)
\end{aligned}
\tag{10}
$$

5. EXPERIMENTS

We conduct extensive experiments to validate ALG-GAN. First, we compare our ALG-GAN with the state-of-the-art GAN models [10], [11], [12], [13], [18], [20], [21]. Then, we validate the effectiveness of each new module proposed by our method, including FM, DM and MTD.

5.1. Experimental Setup

Dataset We use CUB [14] and Oxford [15] datasets to verify the text description based image generation. We preprocess and split the images into two disjoint sets following the same pipeline as GAN-INT-CLS [10]. CUB contains 11,788 bird images belonging to 200 categories, where 150 categories with 8,855 images are employed for training while the remaining 50 categories with 2,933 images are used for testing. Besides, CUB contains category and fine-grained attribute annotations. We choose them as the weak-supervision information and figure out the effectiveness for each of them. Oxford contains 8,189 images of flowers from 102 different categories, where 82 categories with 7,034 images are employed for training while the remaining 20 categories with 1,155 images for testing. Oxford only has category annotation, which is employed as the weak-supervision information. Each image in both CUB and Oxford has 10 text descriptions.

Parameter and model setting In our experiments, same as the setting [13], we define λ_1 as 1.0 and λ_2 as 5.0 following. We define λ_3 as 1.0 empirically. For text embedding, we employ a pretrained text encoder on CUB. For Oxford, we train the text encoder. During training, we fix the parameters of encoder to get the word features and sentence features. Then, we train AttnGAN [13] for Oxford as the baseline.

Evaluation metric We use IS [16] and FID [17] as the quantitative evaluation measures. IS measures both quality and diversity of generated images. It computes KL-divergence between the generated class distribution and the real class distribution, which uses the pre-trained Inception v3 network. A higher score means a better performance. FID computes the Fréchet distance between generated images and real images using the extracted features from a pre-trained network. A lower FID means a closer distribution between generated images and real ones.

5.2. Comparative Results

We compare our results with the state-of-the-art text-to-image methods on CUB [14] and Oxford [15] datasets. We report the results of IS in Table 2. ALG-GAN outperforms other methods with a higher IS. It indicates that ALG-GAN generates images with better quality and diversity.

Table 2. The comparison of IS by our ALG-GAN and the state-of-the-art GAN models on CUB and Oxford datasets

Methods	CUB	Oxford
GAN-INT-CLS	2.88±0.04	2.66±0.03
TAC-GAN	-	3.45±0.05
StackGAN	3.70±0.04	3.20±0.01
StackGAN++	4.04±0.06	-
AttnGAN	4.36±0.03	3.74±0.09
HDGAN	4.15±0.05	3.45±0.07
MirrorGAN	4.56±0.05	-
Our (ALG-GAN)	**4.62±0.07**	**4.10±0.08**

Table 3. FID between AttnGAN and ours, lower is better

Methods	CUB	Oxford
Baseline (AttnGAN)	16.898	46.459
Our (ALG-GAN)	**16.500**	**44.307**

Table 3 compares the performance between AttnGAN and ALG-GAN with respect to FID on CUB and Oxford. We measure FID by the officially pre-trained model. After resizing the real test images and the synthesized images in the same size, we compute FID between them. Compared with AttnGAN, our ALG-GAN decreases FID from 16.898 to 16.500 on CUB and from 46.459 to 44.307 on Oxford, which demonstrate that ALG-GAN can learn a better data distribution on objects. Representative examples generated from text descriptions by different methods are shown in Figure 3.

Figure 3. Qualitative examples of the proposed ALG-GAN comparing with HDGAN [20] and AttnGAN [13] on CUB and Oxford dataset.

5.3. Ablation study and discussion

To further demonstrate the effectiveness of each component, we perform some ablation experiments. Table 4 shows the results. These results demonstrate that each component in ALG-GAN is indispensable.

Table 4. IS produced by combining different components of ALG-GAN.

Methods	CUB	Oxford-102
Baseline	4.36±0.03	3.74±0.09
Baseline + FM	4.46±0.04	3.86±0.08
Baseline + FM + MTD(CCT)	4.56±0.04	4.08±0.08
Our (Baseline + FM + MTD(CCT)+DM)	4.58±0.04	**4.10±0.08**
Baseline + FM + MTD(ACT)	4.59±0.05	-
Our (Baseline + FM + MTD(ACT)+DM)	**4.62±0.04**	-

FM Baseline + FM improves IS of2.3% over the baseline on CUB. Meanwhile, It results in 3.2% improvement on Oxford. The results confirm that forgetting the redundancy and incorrect information benefits the generation. In addition, in order to prove that the improvement of model performance is not caused by the increase in computing power after the introduction of the new structure, we set up a comparative experiment to replace the down-up sampling dual structure in ALG-GAN with naive Conv3×3. IS is shown in Table 5. It can be seen from the results that the introduction of additional convolutional layers will cause the model learn knowledge more redundant, resulting in a decrease performance.

CGD Based on FM, we further evaluate the effectiveness of CGD by validate Baseline + FM + CGD. When the discriminators conduct the category classification task (CCT), it improves IS from 4.46 to 4.56 on CUB and 3.86 to 4.08 on Oxford. When the discriminators conduct the attribute classification task(ACT), it improves IS from 4.46 to 4.59 on CUB. Those improvements prove that the weak-supervision of discriminator guides generators more comprehensively. Moreover, more fine-grained guidance results in better quality of the generated images. Because of the lack of attribute annotation in Oxford, we do not verify the validity of attribute classification on the flower dataset.

Table 5. The comparison of IS by FM and naive Conv3×3 on CUB and Oxford datasets

Method	CUB(ACT)	Oxford(CCT)
Baseline	4.36 ± 0.03	3.74 ± 0.09
Baseline + FM + CGD	4.59 ± 0.05	4.08 ± 0.03
Baseline + Conv3×3 + CGD	4.33 ± 0.05	3.64 ± 0.07

DM We discuss the effect of hyper parameter α in DM to IS through Baseline + FM + CGD. When α is set as 5.1, the model has best performance on ACT. When the weak-supervision information is category label, α =5.6 achieves best results. On Oxford, IS is stable when α are set from 5.5 to 5.7.The results of IS are shown in Table 6 and 7, which show that guaranteeing the initialization quality through supervision on the generation process benefits the subsequent image refinement. However, although DM works well, it is difficult to find one α to apply to all models. The reason is that α is influenced by the model initialization and other hyper parameters such as learning rate and batch size etc.

Table 6. Results on CUB from different α in DM

Method	CUB
Baseline + FM	4.46±0.04
Baseline + FM +MTD(ACT)	4.59±0.05
Baseline + FM +MTD(ACT) +DM (α = 4.5)	4.57±0.05
Baseline + FM +MTD(ACT) +DM (α = 5.0)	4.60±0.05
Baseline + FM +MTD(ACT) +DM (α = 5.1)	**4.62±0.04**
Baseline + FM +MTD(ACT) +DM (α = 5.2)	4.60±0.06
Baseline + FM +MTD(ACT) +DM (α = 5.5)	4.59±0.05
Baseline + FM + MTD(CCT)	4.56±0.04
Baseline + FM +MTD(CCT) +DM (α = 5.0)	4.57±0.05
Baseline + FM +MTD(CCT) +DM (α = 5.5)	**4.58±0.03**
Baseline + FM +MTD(CCT) +DM (α = 5.6)	**4.58±0.04**
Baseline + FM +MTD(CCT) +DM (α = 5.7)	**4.58±0.04**
Baseline + FM +MTD(CCT) +DM (α = 6.0)	4.57±0.04

Table 7. Results on Oxford from different α in DMs

Method	Oxford
Baseline +FM	3.86±0.08
Baseline +FM +MTD(CCT)	4.08±0.08
Baseline +FM +MTD(CCT) +DM(α=5.5)	4.09±0.07
Baseline +FM +MTD(CCT) +DM(α=5.6)	**4.10±0.08**
Baseline +FM +MTD(CCT) +DM(α=5.7)	4.08±0.09
Baseline +FM +MTD(CCT) +DM(α=6.0)	4.07±0.09

6. CONCLUSION

In this paper, we propose a novel ALG-GAN method for efficient text-to-image synthesis. Compared with previous models, our ALG-GAN performs better in generating consistent and high-quality images because the generator learns more adaptively with the forgetting mechanism and the drafting mechanism. Besides that, the comprehensive guiding fiscriminators reduces the mode collapse.

As future work, using efficient language model to process text description can obtain more informative condition vector, and using this vector to generate text to image can obtain higher quality images.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China under Grant No.62173265, the Fundamental Research Funds for the Central Universities, the Innovation Fund of Xidian University.

REFERENCE

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-

consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.
[3] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," arXiv preprint arXiv:1805.08318, 2018.
[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in Advances in neural information processing systems, 2016, pp. 2172-2180.
[5] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.
[7] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 2642- 2651.
[8] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," arXiv preprint arXiv:1511.02793, 2015.
[9] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," arXiv preprint arXiv:1502.04623, 2015.
[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," arXiv preprint arXiv:1605.05396, 2016.
[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photorealistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907-5915.
[12] Zhang H, Xu T, Li H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.
[13] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.1316-1324.
[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltechucsd birds-200-2011 dataset," 2011.
[15] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, 2008, pp. 722-729.
[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in Advances in neural information processing systems, 2016, pp. 2234-2242.
[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Neural Information Processing Systems, 2017, pp. 6626-6637.
[18] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "Tac-gan-text conditioned auxiliary classifier generative adversarial network," arXiv preprint arXiv:1703.06412, 2017.
[19] M. Cha, Y. L. Gwon, and H. Kung, "Adversarial learning of semantic relevance in text to image synthesis," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3272-3279.
[20] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.6199-6208.
[21] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505-1514.

AUTHORS

Yuting Xue was born in 1996. He received the M.S. degree in Electronic Science and Technology from Xidian University, Xi'an, China, in 2021. His main research interests include image and video processing, data mining.

Heng Zhou now is working toward the Ph.D. degree in Electronic Science and Technology with Xidian University, Xi'an, China. His current research interests include image processing, pattern recognition, and their applications in infrared target detection and segmentation.

Yuxuan Ding was born in 1995. He received the B.S. degree in Intelligent Science and Technology from Xidian University, Xi'an, China, in 2018. He is currently a Ph. D. Candidate at School of Electronic Engineering, Xidian University. His main research interest covers Machine Learning, Computer Vision, Vision-Language, and their applications.

Xiao Shan now is working toward the M.S. degree in Electronic Science and Technology with Xidian University, Xi'an, China. Her main research interests include image processing, machine learning, and image generation.