DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis Supplementary Document

Shulan Ruan[1†], Yong Zhang[2*], Kun Zhang[3], Yanbo Fan[2], Fan Tang[4], Qi Liu[1], Enhong Chen[1*]
[1]School of Computer Science and Technology, University of Science and Technology of China
[2]Tencent AI Lab, [3]Hefei University of Technology, [4]Jilin University
slruan@mail.ustc.edu.cn, {zhangyong201303, zhang1028kun, fanyanbo0124, tfan.108}@gmail.com, {qiliuql, cheneh}@ustc.edu.cn

For better demonstration of our approach, in Section 1, we first make more detailed analyses of the experiments presented in the manuscript. Next in Section 2, we present more visual results on CUB-200 [2] and COCO [1].

1. Detailed Experimental Analyses

In order to demonstrate the effectiveness and rationality of our proposed DAE-GAN, we make more detailed analyses for the qualitative results.

As shown in Figure 1, in the 1^{st} column, compared with AttnGAN [3] and DM-GAN [5], only our proposed DAE-GAN synthesizes the details of '`wild animals`'. Moreover, the image generated by DM-GAN also fails to contain '`a cloudy sky`'. We could address the problem well because DAE-GAN obtains comprehensive text representations, especially aspect-level features. Moreover, *ALR* is developed to dynamically enhance image details with aspect information. Thus, DAE-GAN could generate images with more details matching the text description.

In the 2^{nd}, 3^{th} and 4^{th} columns of Figure 1, it can be observed that AttnGAN and DM-GAN often generate one object multiple times (*e.g.*, '`blue sign`' and '`clock`') and the spatial distribution is also chaotic.

We investigate the reasons from the generation mechanism and some related researches. Current methods mainly first generate a low-resolution image at the initial stage, and then refine them by repeatedly employing the attention mechanism to select important words for image enhancement at the refinement stage. However, it may be stuck in one or two of the most important words due to the lack of supervisory information [4]. For example, in Table 1, when applying the attention mechanism to select one word each time from sentence '`a person in a red shirt and black pants hunched over`', the attended word sequence is '`red shirt red shirt red shirt`'. Obviously, another important aspect information '`black pants`' is overlooked.

Our proposed DAE-GAN can greatly alleviate these problems. By alternately applying *AGR* and *ALR*, DAE-GAN will not only enhance local details but also refine images from a global perspective. This mechanism allows DAE-GAN to avoid getting stuck in a few most important words like other methods.

Figure 1. Example results for text-to-image synthesis by AttnGAN [3], DM-GAN [5] and our proposed DAE-GAN COCO.

Table 1. Attended word sequence selected by repeatedly applying the attention mechanism as listed by Zhang *et al.* [4]. For each sentence, they repeatedly apply the attention mechanism for six times and select one word each time.

Sentence	Attended word sequence
a person in a red shirt and black pants hunched over.	red shirt red shirt red shirt
a woman paints a portrait of a person.	paints portrait portrait portrait portrait person

[†]Work done during an internship in Tencent AI Lab.
[*]Corresponding Authors.

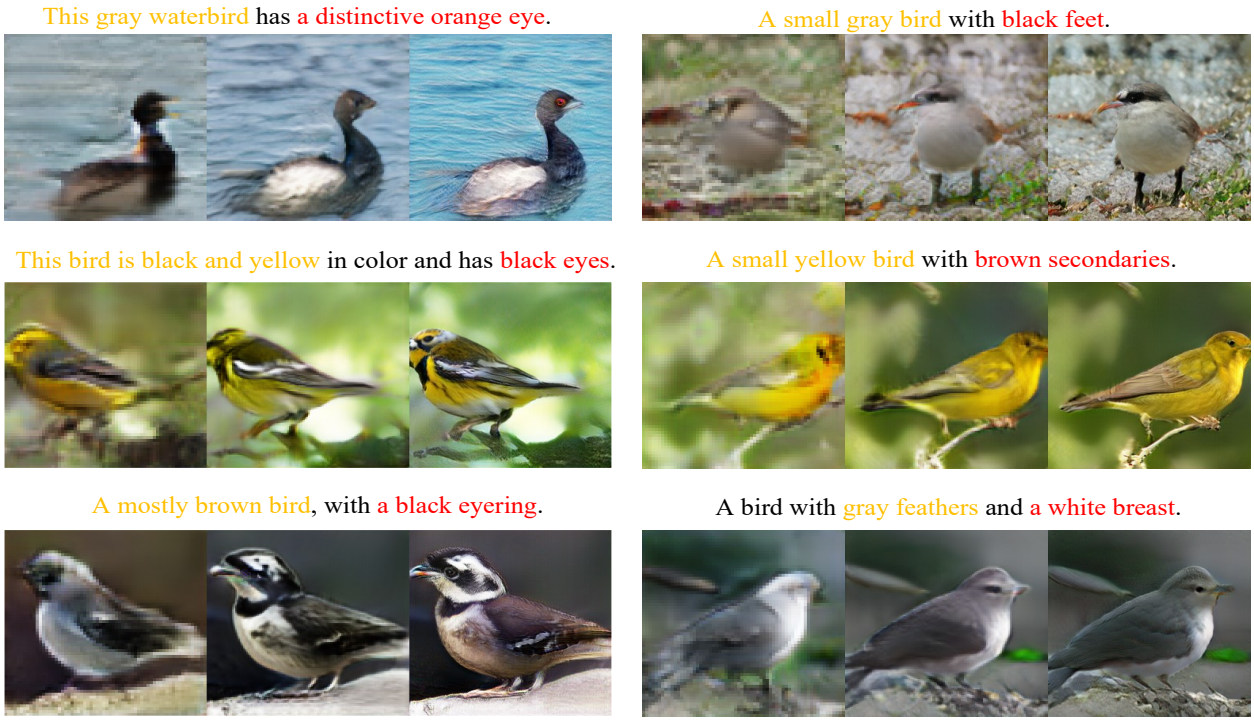

Figure 2. Text-to-image synthesis visualization of different generation steps. DAE-GAN initially generates a low-resolution image with the size of 64×64. Then, at the refinement stage, DAE-GAN refines the image with the size of 128×128 on the basis of the image generated initially. The final output image has the size of 256×256. We denote aspects with different colors (*i.e.*, gold and red). Each aspect is utilized in one refinement step.

Figure 3. Synthesized images by our proposed DAE-GAN on CUB-200.

Figure 4. Synthesized images by our proposed DAE-GAN on COCO.

2. More Visual Results

In this section, we present more visual results on CUB-200 and COCO to show the effectiveness of our proposed DAE-GAN.

In Figure 2, we provide more experimental examples to illustrate the synthesis process of DAE-GAN. In Figure 3 and Figure 4, more synthesized images are shown on both CUB-200 and COCO datasets.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[2] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[3] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong 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*, pages 1316–1324, 2018. 1

[4] Kun Zhang, Guangyi Lv, Linyuan Wang, Le Wu, Enhong Chen, Fangzhao Wu, and Xing Xie. Drr-net: Dynamic re-read network for sentence semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7442–7449, 2019. 1

[5] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 1