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

Shulan Ruan1†, Yong Zhang2*, Kun Zhang3, Yanbo Fan2, Fan Tang4, Qi Liu1, Enhong Chen1*
1School of Computer Science and Technology, University of Science and Technology of China
2Tencent AI Lab, 3Hefei University of Technology, 4Jilin 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 1st 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 2nd, 3rd and 4th 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.
This gray waterbird has a distinctive orange eye.
This bird is black and yellow in color and has black eyes.
A mostly brown bird, with a black eyering.
A small gray bird with black feet.
A small yellow bird with brown secondaries.
A bird with gray feathers and a white breast.

Figure 2. Text-to-image synthesis visualization of different generation steps. DAE-GAN initially generates a low-resolution image with the size of $64 \times 64$. Then, at the refinement stage, DAE-GAN refines the image with the size of $128 \times 128$ on the basis of the image generated initially. The final output image has the size of $256 \times 256$. 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.
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


