Supplement: Improving Text-to-Image Synthesis Using Contrastive Learning

Hui Ye[1]
hye2@student.gsu.edu
Xiulong Yang[1]
xyang22@student.gsu.edu
Martin Takáč[2]
takac.MT@gmail.com
Rajshekhar Sunderraman[1]
rsunderraman@gsu.edu
Shihao Ji[1]
sji@gsu.edu

[1] Department of Computer Science, Georgia State University Atlanta, GA, USA
[2] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) Masdar City, Abu Dhabi, UAE

1 Visual Quality

To further compare our proposed approach with the baselines, we visualize the synthetic images generated from the typical example captions. As shown in Figure 1, compared with the baseline AttnGAN, the images generated from our approach are more realistic and better match with the text descriptions in most cases. In the 8th column, the bird in the image from AttnGAN [1] fails seriously with two heads, while the one from our approach has the reasonable appearance. In the 2nd column, we can see the vivid green crown in the bird from our approach, which matches the description “green crown" well, while the image from AttnGAN does not show this feature of the bird. As shown in Figure 1, the comparison between our approach and the baseline DM-GAN [2] is similar to previous comparison. In the 3rd column, the image from our approach has the correct white belly to match the text description “while belly", while the image from DM-GAN has the additional incorrect red color in the belly. Figure 2 shows the example images on COCO from our approach and the baselines AttnGAN and DM-GAN. It is challenging to generate photo-realistic images for the models showed in the figure. However, compared with the baselines, the images generated from our approach are more realistic and better match with the text descriptions in some cases. In the 4th column, the boat in the image from our approach has the red and while color, which aligns with the caption, while the image from AttnGAN does not show the red boat. In the 5th column, the image from our approach has better shape of cows than the one from AttnGAN. As shown in the 3rd column, the image from our approach has the basic shape of a girl, while it can not be observed in the image from DM-GAN at all.

Figure 1: Comparison of example images between our approach and baselines on the CUB dataset.

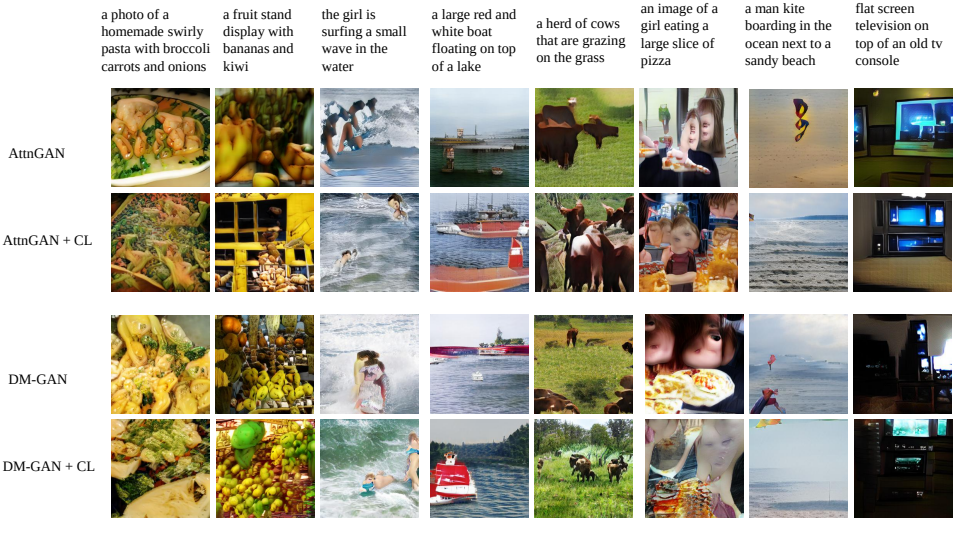

Figure 2: Comparison of example images between our approach and baselines on the COCO dataset.

References

[1] 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.

[2] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.