Proposal

Yue Yang

December 6, 2022

1 Introduction

Image synthesis has been receiving significant attention recently owing to its potential for boosting many fields such as artworks [Tan et al., 2017], medical imaging [Nie et al., 2018], etc. Generative models are one mainstream for synthesizing images. Generative adversarial network (GAN) [Goodfellow et al., 2014] gained popularity in this field due to its efficient sampling of high-resolution images, but it's unstable for optimization. Variational autoencoders (VAE) [Kingma and Welling, 2013] are likelihood-based methods but are limited to the sample quality. Diffusion Probabilistic Models (DM) [Sohl-Dickstein et al., 2015], which emerges as a promising image generation method by leveraging a guidance technique [Dhariwal and Nichol, 2021], are another line of work. However, the diffusion models could be extremely computationally expensive because they oftentimes directly operate in pixel space. To mitigate this issue, Rombach et al. [2022] propose the latent diffusion models (LDMs) that apply DM in latent space. This work has also achieved state-of-art results in image synthesis.

2 Proposed Method

Although LDMs have achieved fascinating results, their capacity is still limited when synthesizing some specific objects. One observation is that the generation of human-like hand images could be a big issue for LDMs. It's easy to find some unreasonable mistakes in the number or the position of fingers (As shown in Fig. 1) in synthesized images by LDMs. Therefore, we propose a novel method to mitigate this hand

(a) Hands with dislocated fingers

(b) Hands with wrong number of fingers

Figure 1: Synthesized hand images with mistakes

generation issue by leveraging additional information to input data. One possible information could be the pose of hands. The pose information can be taken as additional channels of the input images so that the embedded geometry features would be valuable for LDMs to learn. The collection of pose images could be obtained by using some existing pose estimation technologies [Wei et al., 2016, Simon et al., 2017]. Although LDMs save lots of computational resources because it works on latent space, they could still be time-consuming if we hope to test our method's feasibility on LDMs. Therefore, before evaluating our methods on LDMs, we propose to begin with GAN models that allow for efficient sampling. StyleGAN [Karras et al., 2019], which achieves state-of-art results among GAN methods, could be one choice for us. We could transfer our method to LDMs after seeing a good performance on GAN models.

By using the proposed methods, we believe that the human-like hand generation with LDMs could be improved, which will be helpful to many image synthesis applications.

3 Timeline

- Week 01 (08/29 $\rightarrow$ 09/02): Hand images collection.
- Week 02 $\sim$ 03 (09/05 $\rightarrow$ 09/16): Hand pose channel generation.
- Week 04 $\sim$ 07 (09/19 $\rightarrow$ 10/14): Training on the GAN model.
- Week 08 $\sim$ 14 (10/17 $\rightarrow$ 12/02): Training on the LDMs.
- Week 15 $\sim$ 16 (12/05 $\rightarrow$ 12/15): Write a report or paper.

References

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12): 2720–2730, 2018.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.