AIST4010 Project Milestone Report

YUEN Yu Ching 1155143580

March 18, 2023

1 Introduction

The problem of pose-based image synthesis has been studied in past works due to its potential in a wide variety of applications such as face editing and image retrieval. However, there has been less focus on pose-based image synthesis of fictional characters.
In this project, we attempt to investigat the problem of anime posture transfer. Specifically, we are aiming to produce a generative model such that given reference image of an anime character A and an image with an arbiturary character in pose B, the model will generate an image of character A in pose B.

2 Related Works

Pose transfer can be splitted in two components: pose extraction and pose-based image generation. In our reviewed papers, the task of pose-based image generation seems to be more thoroughly studied. Older methods are mostly based on Variational Autoencoders or Generative Adversarial Networks [1], while recent methods focuses more on transformer-based approaches.
Ren et al. [2] proposes a transformer-based approach of extracting input features and reconstruting in a semantically coherent method. Specifically, an extraction would extract neural textures by gathering features in the reference image, then a distribution operation would generate an output image for each target (e.g. head, hand, torso) according to the learned semantic distribution.
Liu et al. [3] proposes an multi-layer sparse attention module which wraps the exemplar image using attention maps to produce a finer matching result, especially for local details.
Zhou et al [4] proposes another method, in which a cross attention based style distribution module calculates between the target pose and semantic style represented by each semantic, and distributes them in a one-to-one matching basis.
Style transfer is a task in which an input image is tuned to minimize the content loss and style loss between a content image and style image respectively.
In this project, we focused on neural style transfer, proposed by Gatys et al. [5], in which a single convolutional neural network is used to obtain the style representation and content repesentation from a style image and a content image respectively, with reconstructions of the feature spaces using the output from different subsets of the CNN.
Image generation refers to the task of generating an image given prompts. A recently popular examples is the Stable Diffusion, introduced by Rombach et al. [6], which is trained with the objective of removing successive layers of gaussian noise similar to a reversal of diffusion in the physical world. It is noted that diffusion models perform the best out of its predecessors in this task: Kim et al. [7] is listed as the top performer with the lowest Frèchet inception distance FID in the CIFAR-10 image generation benchmark.

3 Dataset

We will be using a subset of the Danbooru2020 [8] dataset, provided by [9]. It is a small 14GB anime image dataset collected from the anime image aggregator Danbooru.

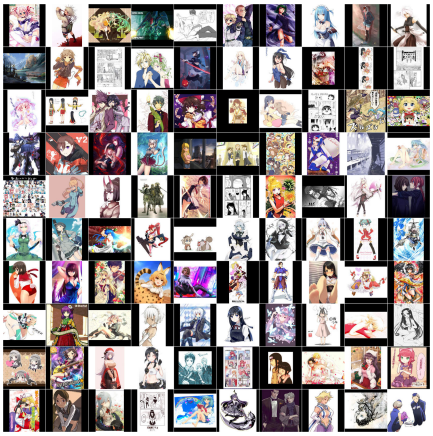

Figure 1: 100 sample images from the danbooru2020 dataset

4 Proposed methods

We approach this problem by splitting it in three components: character pose extraction, character feature extraction, and image generation. To this end, there are multiple possible methods to fufill the above components.
First is fine-tuning a stable diffusion model using ControlNet proposed by Zhang et al. [10]. By "locking" certain layers of the diffusion model, the generative feature of the model would be secured, while the features perceived by the model can be finely controled.
Another approach is the neural style transfer [5] mentioned above. We note that the output image has similar colors as the style image, while the outline of the output image is similar to that of the input image. Since characters usually are aligned similarly (e.g. head on top, legs on bottom), we hoped that this would translate to a coherent result.
Last but not least, the human-focused pose transfer could be adopted for character drawings too. However, due to difference in photos of actual humans and drawings of characters, another method would need to be developed or fine-tuned for our task.

5 Preliminary results

Given the nature of the task, we initially attempted to fine-tune the `waifu-diffusion` diffusion model provided by hakurei on huggingface [11] using ControlNet. However, due to hardware limitations, this proved to be infeasible.
We then attempted to approach it on a painting-oriented approach of style transfer, given the drawn nature of our dataset. While the outline of the output image resembles the content image, and colors are distributed roughly according to the style image (pink-ish color on the top, blue at the "torso" and "leg" parts), it does not resemble any of the two characters.
We plan to explore the human-based style transfer models in the next phase of the project. Should

Figure 2: From left to right: content image, style image, output

transformer-based models be unavailable due to hardware limitations, we will attempt the task with earlier GAN-based approaches.

References

[1] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," *CoRR*, vol. abs/1801.00055, 2018. [Online]. Available: http://arxiv.org/abs/1801.00055

[2] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," 2022.

[3] S. Liu, J. Ye, S. Ren, and X. Wang, "Dynast: Dynamic sparse transformer for exemplar-guided image generation," 2022.

[4] X. Zhou, M. Yin, X. Chen, L. Sun, C. Gao, and Q. Li, "Cross attention based style distribution for controllable person image synthesis," 2022.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [Online]. Available: https://github.com/CompVis/latent-diffusionhttps://arxiv.org/abs/2112.10752

[7] D. Kim, Y. Kim, S. J. Kwon, W. Kang, and I.-C. Moon, "Refining generative process with discriminator guidance in score-based diffusion models," 2023.

[8] Anonymous, D. community, and G. Branwen, "Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset," https://gwern.net/danbooru2021, January 2022, accessed: 2023/02/20. [Online]. Available: https://gwern.net/danbooru2021

[9] T. Mcnally, "Anime art [danbooru 2020 small]," 2020. [Online]. Available: https://www.kaggle.com/datasets/muoncollider/danbooru2020small

[10] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[11] hakurei, "waifu-diffusion v1.4 - diffusion for weebs," Huggingface, accessed 2023/02/20. [Online]. Available: https://huggingface.co/hakurei/waifu-diffusion