Transfer learning for Person Image with Attribute-Decomposed GAN

Puda Zhao, Hao Wang, Yinhong Qin
pz2078@nyu.edu,hw2671@nyu.edu, yq2021@nyu.edu
New York University

November, 2022

0.1 Introduction

Generative adversarial network is a famous and widely used backbone for generative modeling. Recent years, there is a new category of GAN model named Attribute-Decomposed GAN (A-D GAN), which is effective to synthesize high-quality person images with user controlled human attributes, such as pose, head, clothes and pants. The corresponding keypoint-based can be automatically extracted via an existing pose estimation method[3]. Typically, processing an intact human picture is a tough work, since directly encodes the entire image might be tedious. The A-D GAN model tries to improve this by adopting automatic and unsupervised component attributes generator[1] into the frame of GAN[2]. Besides, to improve the generalization ability of texture encoding, A-D GAN model introduce an architecture of global texture encoding by concatenating the VGG features in corresponding layers to its original encoder, which is inspired by a style transfer method[4] which directly extracts the image code via a pretrained VGG network.

In our project, we will use a mordern discriminator method[5] for our model. It will adapt two discriminators D_p and D_t, and their specific attributes and responsibilities are explained in the discriminator sections. We would like to also introduce a new metric called contextual (CX) score, which is first proposed for image transformation[6] and use the cosine distance between deep features to measure the similarity of two nonaligned images while ignoring the spatial position of the features.

0.2 Hypotheses & Research Questions

Since Attribute Decomposed GAN works well on human poses transporting, we assume that it is likely to represent attributes of human poses in other datasets. Therefore, we will strive to solve the following research questions:

1. Does it still feasible on other kind of datasets, such as unfashionable data sets, or non-3D characters pictures, for example anime characters?
2. How well it would work with three dimensional characters pictures?
3. Are there any statistical or information-theoretic intuitions behind these methods and experimental results? If so, what are they?
4. Are there any possible solutions to reduce the training cost of this network?

0.3 Dataset

We conduct experiments on the Inshop Clothes Retrieval Benchmark in the Deep Fashion database. The dataset containing various poses and clothes, which is of large scale, diversities and quantities. It also has rich annotations, including 7982 number of clothing items.

In the original experiment design, the author randomly pick 101,996 pairs of images for training and 52,712 pairs for testing. We are going to reduce the scale of training to 1000 pairs for training and 75 pairs, because we make transfer learning from the pretrained model, which could take advantage of the original largescale model. We will crop images into 172 x 256 resolution which is $I \in R^{3\times172\times256}$, then we generate Component Transfer by human parsing model Look Into Person. Each image is segmented by 8 categories(i.e., background, hair, face, upper clothes, pants, skirt, arm and leg). As for Pose Transfer, we will generate key points of body joints by using OpenPose, the output will be a 18 channel heatmap representing human pose $P \in R^{18\times172\times256}$ of I. Finally synthesis them together. Starting from the original dataset, the path of data processing is as shown in the following figure 1.

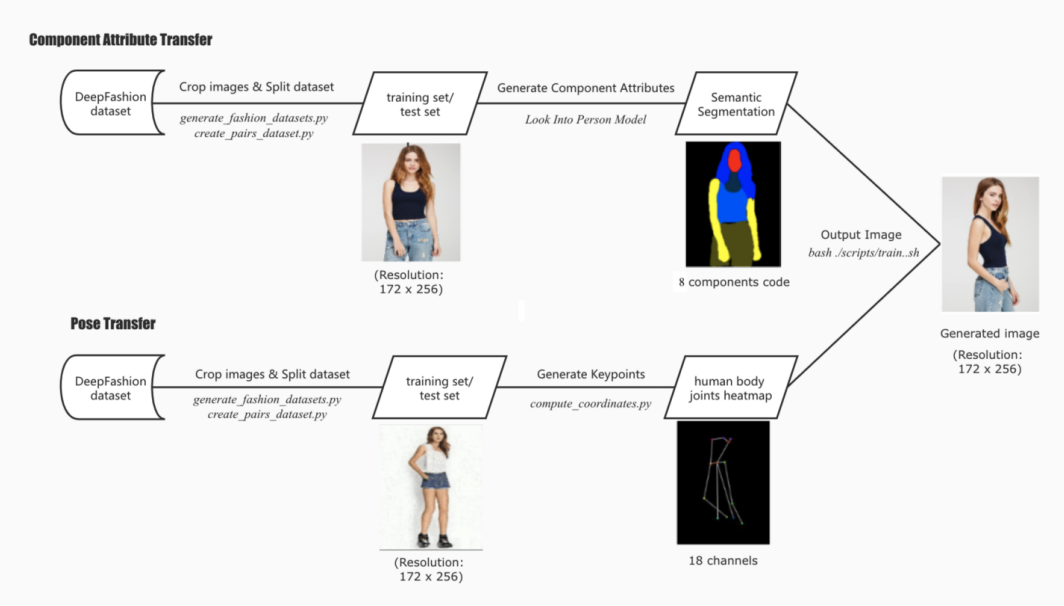

Figure 1.

And when do the transfer learning, we use 14 3D anime character pictures as the transfer learning dataset, pair by pair which means 7×7 data. Since this model is a really huge model and we only have limited resources, so the dataset could not set too big.

0.4 Deliverables

By the end of this project, we are expected to:

1. Achieve component attributes transform model to deepfashion dataset.
2. Transform this model to anime dataset and evaluate the result.
3. Finally create a model which could transform fashion style and pose between different anime characters.
4. Collected all these results in a detailed report and a GitHub repository.

0.5 References

1 Yifang Men et al. "Controllable Person Image Synthesis with Attribute Decomposed GAN". In: Computer Vision and Pattern Recognition (CVPR), 2020 IEEE Conference on. 2020.

2 Ian J. Goodfellow et al. Generative Adversarial Networks. 2014. arXiv: 1406.2661 [stat.ML].

3 Zhe Cao et al. OpenPose: Realtime MultiPerson 2D Pose Estimation using Part Affinity Fields. 2019. arXiv: 1812.08008 [cs.CV].

4 Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. 2017. arXiv: 1703.06868 [cs.CV].

5 Zhen Zhu et al. "Progressive Pose Attention Transfer for Person Image Generation". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2019.

6 Roey Mechrez, Itamar Talmi, and Lihi ZelnikManor. The Contextual Loss for Image Transformation with Non-Aligned Data. 2018. arXiv: 1803.02077 [cs.CV].