CS 229 Project Milestone: Multiview Human Synthesis From a Singleview

Si Wen (06246679), Tiancong Zhou (06247022), Honghao Qiu (06246258)

Abstract – This project targets at training a model to synthesize multiview human images given a singleview image. To achieve this, we firstly generate a large image dataset of 360-degree views of human, then combines variants of GAN/VAE models for model training, and try to improve the result in researching into better loss functions and parameters setups. We finally apply the model to front-view human images to synthesize 360-degrees views for full human body. One potential application of this project is in E-commerce cloth model multiview image generation.

I. Introduction and Related Work

Our project aims to use machine learning techniques to generate multi-view images of a person given a single view RGB image. This could potentially enable many useful applications in fashion/E-Commerce websites and in the field of photo/video edition and content generation.

In recent years, there have been numerous attempts to solve subsets of this problem. On one hand, Zhao et al [2] proposed a model to synthesize the side and back view of a person given the frontal view. However, his work focused on the person's clothing, and the resulting image contains poorly synthesized face. On the other hand, Huang et al [3] proposed a method to provide photorealistic synthesis of the frontal views of a person's face given a side view. There are also some other related works for object rotation [6,8] and 3d model generation based on multiview images [3,5,6]. We will combine and improve upon these methods in order to generate multiview images of an entire person and try to achieve a better result.

II. Proposed Approach

A. Dataset

Solving this problem requires large amount of training data for the network to learn the latent representation. We use a combination of real world datasets and self-generated datasets. For real world datasets, we use the publicly available DeepFashion [13] and MVC datasets [12] that contain multiview images of a person.

Since the above mentioned real world views are rather limited and often only have frontal and side views, we try to overcome this limitation by using synthetic images generated from 3D modeling softwares. We also notice that some past usage of synthetic data on vision problems has shown promising results [2,5,6].

Our 360-degree human view dataset (about 100,000 full-body human images in 360-degree views) is generated by the following approach:

i Create 3D character models in Adobe Fuse. Adobe Fuse provides plenty of hairstyles, faces, cloth and shoes to combine and create 3D characters;
ii Export the model as .obj file;
iii Import the obj file into Blender. (Blender is a 3D graphics software that could be programmed and rendered to generate 360-degree views);
iv Render and generate front view;
v Rotate the 3D character model by 1 degree clockwise, render and generate new view from that degree;
vi Repeat step 5 and 6 until generating views for all 360 degrees.

Here are the samples of a generated human body viewed from two different angles:

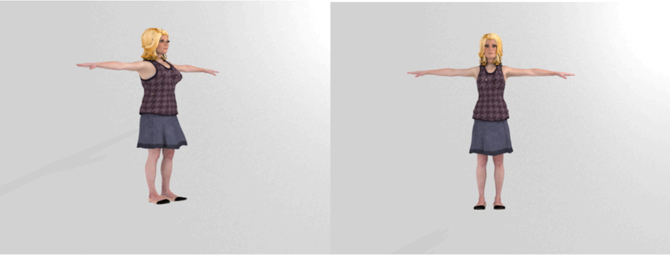

Figure 1: Generated Multiview Images Samples

We use standard cross-validation approach (80/10/10) for getting training, dev and testing datasets, and try to keep the dataset distribution the same across train/dev/test sets by following the same sampling percentage from real world and self-generated image datasets.

B. Model

We combines variants of the Generative Adversarial Network (GAN) [1] to solve this problem. GAN is a generative approach that learns a distribution to represent the training data. We can then generate new images sampled from that distribution, conditioned on the input image and the desired angles. To ensure quality of the synthesized images, we use multiple GANs to synthesize different parts of the an image (TP-GAN[3] for face and VariGAN [2] for body) and combine the results together.

As is shown in Zhao et al paper [VariGan, 2], GAN is good at filling rich details of the synthesized image but less capable of capturing global appearance and rough outlines for human/clothings, and on the other hand, VAE [7] is more capable of finding global appearance with less detailed details, therefore, GAN and VAE are good complements of each other for image synthesis. Our model follows the following two-step approach to achieve image synthesis: 1) firstly use VAE to model low resolution global appearance (outline/structure of the image), 2) then use variants of GAN to model refined details of the person and get high resolution synthesized images of full human body.

Additionally, We noticed that in previous work [2], the head of human body are poorly synthesized when combined with body/cloth synthesis, we try to resolve this problem by combining TPGAN with the above model, which is, using TPGAN that has a relatively good performance on head/face synthesis, and combine it with the above mentioned VAE plus GAN approach to achieve better performance for synthesizing full human body (head, body and clothes) at the same time. Another modeling improvement we are looking into is to get rid of the help of conditional images used in VariGAN in the GAN training procedure. With more data for human body and from all angles, we try to experiment the approach of generating high resolution human synthesis for a given view (e.g. frontal view) without using conditional images from a different view (e.g. side view in [2]).

Finally, we find that the reason why VariGAN approach has a poor performance on head synthesis is due to the fact that only adversarial loss is being used in GAN's objective, inspired by Huang et al's work [3], we try to model face generation better by using pixel to pixel comparison loss, in this way face features would be better captured since incorrectly predicted faces are penalized with additional loss. As for body modeling, we also try to introduce a new symmetry loss for front/back-view human synthesis, this is based on the inherent fact that human body are symmetry in general (More details are discussed in the following Loss Function section). We hope by extensive experiments (in Section IV) we can show that this modeling approach can help us achieve better performance compared with the state-of-art synthesis results of human body/head in multiviews (front and side), and can further extend this to all 360 degree angles.

C. Loss Function

We mainly use the following loss functions (with penalty factor assigned to each loss and calculating the combined loss as objective loss function):

- Adversarial loss:
  Without considering the conditional image (as is mentioned in model section), we define adversarial loss as:
  $$E_{I_v\sim p_{data(I_v)}}[logD(I_v)]+E_{z\sim p(z)}[log(1-D(I_v,G(z,I_v)))],$$
  here I_v stands for Image given a Viewpoint, D is the discriminator and G is the generator to be trained simultaneously.
- VAE loss:
  As is shown in Bowman et al's work [11], we are also using the standard VAE loss term, which is defined as the sum of KL-divergence loss and another posterior Expectation loss term:
  $$L(\theta;x)=-KL(q_\theta(z|x)||p(z))+E_{q_\theta(z|x)}[log(p_\theta(x|z))]$$
- Pixel to pixel loss:
  The pixel-wise loss can be expressed as [3]:
  $$\frac{1}{L*H}\sum_{x=1}^{L}\sum_{y=1}^{H}|\hat{I}_{x,y}-I_{x,y}|,$$
  where L stands for Length (number of pixels in row-wise), and H stands for Height (number of pixels in column-wise). $\hat{I}_{x,y}$ is the predicted pixel of the image and $I_{x,y}$ is the corresponding pixel in GT.
- Symmetry loss (only in front/back-view synthesis):
  Using the fact that human body is (largely) symmetric (left to right) in front and back view, introducing symmetry loss could be helpful for better human body synthesis performance in front/end view cases. This term can be defined as:
  $$\frac{1}{L/2*H}\sum_{x=1}^{L/2}\sum_{y=1}^{H}|\hat{I}_{x,y}-\hat{I}_{W-x+1,y}|,$$
  As is shown in Huang's paper [TPGAN 3], the introduction of Symmetry loss can improve the synthesis performance of human face. Since human body has similar characteristic, this term should also help improve our model's synthesis performance of full human body. We would show this in experimentation section under algorithmic analysis part.
- Regularization term in loss function:
  We use adaptive L2 regularization on the discriminator to penalize large weights and mitigate over-fitting. As is shown in [1], this is helpful when discriminator is too strong for the generator and the regularization term can be decreased when generator catches up with the discriminator.

Apart from the similarities of the above loss functions defined in previous related works [1,2,3], the difference of our loss function is that we introduce Symmetry loss trying to improve full body synthesis, while using pixel to pixel loss for improving human face synthesis.

III. Network Architecture

We consider the following network architectures:

- i simply applying GAN to training dataset and make synthesis prediction
- ii using VAE to generate low resolution image (with global structure and outline of the person and cloth), and combine the generated low resolution image with condition images to generate high resolution multiview image synthesis based on variants of GAN methods
- iii using VAE to generate low resolution image (with global structure and outline of the person and cloth), and apply variants of GAN to generated high resolution multiview images without using any conditional images
- iv using TP-GAN [3] to make image synthesis, especially for face synthesis

IV. Experiment and Analysis

We use both quantitative and qualitative measures to assess the result of our work. We start our project by generating side-view images from a frontal synthetic image, and then move to real images. We then try to generate views from all angles, including the ones where one cannot easily infer the person's look from the frontal image (e.g. the back of the head, hair, cloth, etc). Given time, we will also try solving the problem under different lighting and background conditions. The performance of the model output would be judged by both loss comparison with previous models, and by comparison results provided by a set of human judges. We would also conduct algorithmic analysis to find out the impact of each component of our networks (GAN, VariGAN, TPGAN, VAE, etc.) and the contribution of each type of training datasets (real images, synthesized images in front, back, and side views, etc.).

Here is the current preliminary result of our model output for human synthesis in different angles (using VAE without conditional image input, 200 filters and 10000 training images for each layer, based on MVC and DeepFashion Dataset). We use this simple VAE model as the baseline model and compare other modeling results with the baseline performance to observe the improvement. As is shown in the picture, currently, our model is able to synthesize roughly the structure and color for human body in different angles, but it lacks the details to make it high resolution, and we are also having difficulty to ensure that the synthesized images are always rotated.

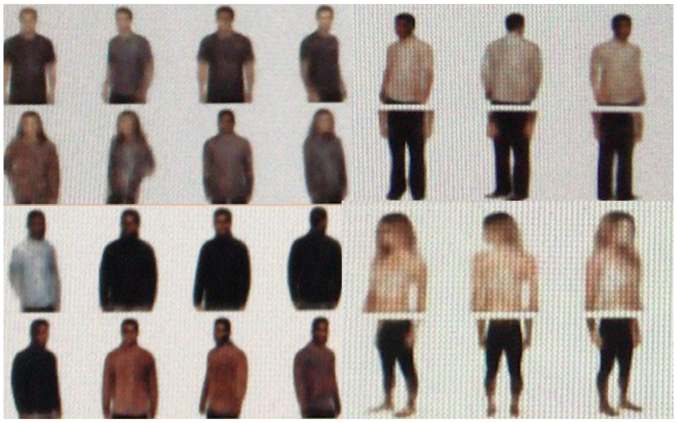

Figure 2: Current preliminary result

In the next step, we would use all images in our self-generated datasets and combine the current GAN model with VAE models, along with TPGAN for better face/head synthesis, to achieve a better synthesis model performance for human body. For the low resolution difficulty, we hope to solve this by introducing variants of GANs and longer training time with more training data to better learn the latent representation of the picture. We also plan to try introducing conditional image (i.e. both using target image and the conditional image in two different views) for encoder input, in order to achieve a better rotational result for our human body synthesis.

V. Conclusion

We aims at solving 360-degree full human body view synthesis problem by using our self-generated dataset and VAE+GAN modeling approach, while introducing some minor changes on modeling architecture and loss function compared with previous works [1,2,3]. Current preliminary results show some promising potential of our modeling approach. With the help of large volume self-generated human view data from 360 view angles, we hope we can solve the 360 degree full human body view synthesis problem and achieve state-of-art results.

Teammate Contributions: Our teamwork break down:

- Si (Vincent) Wen: Vicent is mainly working on variants of GAN modeling, he also contributed to dataset generation (Fuse and MVC) and writing reports (some of modeling section). He is also working on TPGAN to improve the model performance.
- Tiancong Zhou: Tiancong is mainly working on dataset generation (human viewed from 360 degrees in Blender), he also contributed to the report (some of data description section) and modeling part with regards to data processing.
- Honghao Qiu: Honghao is mainly working on modeling and reporting. He works on VAE+GAN modeling, and he is the main contributor to writing this report, he is also engaged in data collection (Deep Fashion).

Our team target at equal contributions from three team members across data, modeling, and reporting. In general, all our teammates have made equal contribution to this project so far. We hope to successfully finish the project with our collective efforts!

Appendix: References

[1] Goodfellow et al, Generative Adversarial Nets, $https://arxiv.org/pdf/1406.2661.pdf$

[2] Zhao et al, Multi-View Image Generation from a Single-View, $https://arxiv.org/pdf/1704.04886.pdf$

[3] Huang et al, Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis, $https://arxiv.org/pdf/1704.04086.pdf$

[4] Yim et al, Rotating Your Face Using Multi-task Deep Neural Network, $https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/-Yim_Rotating_Your_Face_2015_CVPR_paper.pdf$

[5] Ashish et al, Learning from Simulated and Unsupervised Images through Adversarial Training, $https://arxiv.org/abs/1612.07828$

[6] Rajpura et al, Object Detection Using Deep CNNs Trained on Synthetic Images, $https://arxiv.org/pdf/1706.06782.pdf$

[7] Diederik et al, Auto-Encoding Variational Bayes, $https://arxiv.org/pdf/1312.6114.pdf$

[8] Kihyuk et al, Learning Structured Output Representation using Deep Conditional Generative Models, $https://pdfs.semanticscholar.org/3f25/-e17eb717e5894e0404ea634451332f85d287.pdf$

[9] IJCAI 2017, NEU, Fashion Style Generator, $https://www.ijcai.org/proceedings/2017/0520.pdf$

[10] NIPS 2017, ETH/MPII, Pose Guided Person Image Generation, $https://arxiv.org/pdf/1705.09368.pdf$

[11] Bowman$*$ et al, Generating Sentences from a Continuous Space, $https://arxiv.org/pdf/1511.06349.pdf$

[12] MVC Fashion Dataset, $http://mvc-datasets.github.io/MVC/$

[13] Deep Fashion Dataset, $http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html$