Virtual Try-On

Presented by Subhajit Ghosh

Department of Computer Science and Engineering
Indian Institute of Information Technology Guwahati

Introduction

- Given a reference person image and a clothing image, the goal of the virtual try-on model is to synthesize a new image of the same person wearing the target clothing such that the shape and pose of the person, as well as the details of the clothing, are preserved.
- It helps consumers in making superior choices. Consumers get to explore the item at their possess pace until they discover the proper choice.
- Virtual try-on represents a practical application of Generative Adversarial Networks (GANs).

Motivation

- This technology gives a huge boost to the fashion industry, as it is a convenient solution for those who would rather not or cannot visit the stores in person. A person can go through a variety of stuff and check in real-time how they look after wearing an outfit.
- Customers return clothing due to neglected desires, and each return of the thing encompasses a considerable natural effect amid fabricating, bundling, and transportation.

Objectives

- To make a new image virtual try-on model, which aims at transferring a target clothes image onto a reference person.
- Focus on preserving the character of a clothes image (e.g. texture, logo, embroidery) when wrapping it in an arbitrary human pose.
- To generate a photo-realistic try-on image when large occlusion and human pose are presented in the reference person.

Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning. It is a combination of two models-

- Generator (G): It generates fake data as much as possible so that it is like an actual data sample.
- Discriminator (D): This model distinguishes real data from fake data.

Generative Adversarial Networks

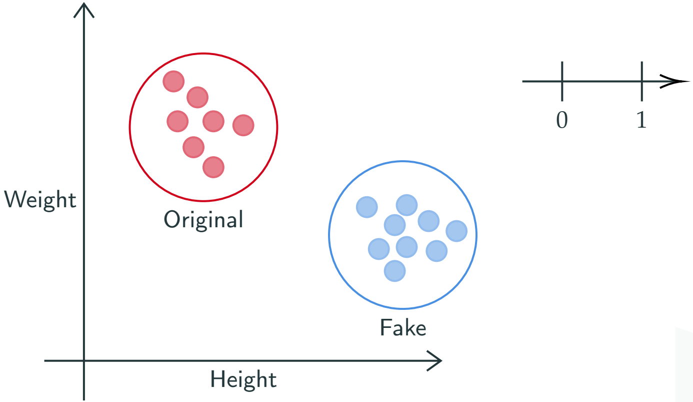

Figure 1: Human height and weight distribution

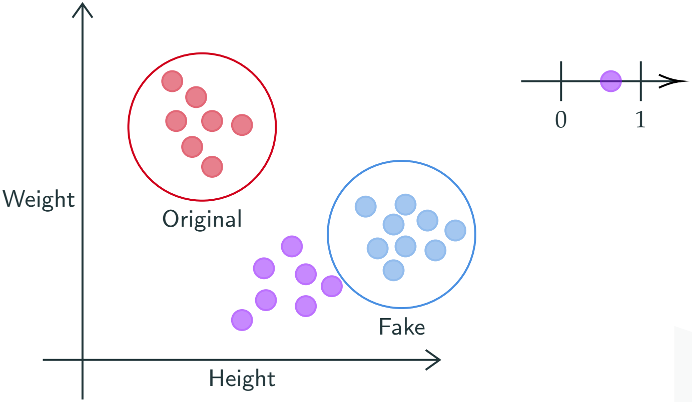

Figure 2: Human height and weight distribution after few iterations

The value function of Generative Adversarial Networks (GANs) is given as follows:

$$\min_{G} \max_{D} \left\{ \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}} \left[\log(1 - D(G(\mathbf{z})))\right] \right\}.$$

- D is the discriminator.
- G is the generator.
- x is the real sample.
- z is the fake sample generated by the generator.
- p_x is the distribution of real data.
- p_z is the distribution of noise.

Adaptively Generating Preserving Image Content[1] (ACGPN) model, it settles the issue of the semantic and geometric distinction between the target dress and referenced pictures beside the occlusions between the torso and limbs. It consists of three major modules:

- Semantic Generation Module.
- Clothes Warping Module.
- Content Fusion Module.

[1] H. Yang, R. Zhang, X. Guo, *et al.*, "Towards photo-realistic virtual try-on by adaptively generating↔preserving image content," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7847–7856. DOI: 10.1109/CVPR42600.2020.00787.

Fill in Fabrics[2] (FIFA) model, it is a self-supervised conditional generative adversarial network model that can handle the complex pose of a reference person while preserving the target clothing details. It consists of three modules.

- Fabricator.
- Segmenter.
- Wraper.

[2] H. Zunair, Y. Gobeil, S. Mercier, *et al.*, "Fill in fabrics: Body-aware self-supervised inpainting for image-based virtual try-on,", 2022. DOI: 10.48550/ARXIV.2210.00918. [Online]. Available: https://arxiv.org/abs/2210.00918.

Related Works: Virtual Try-On II

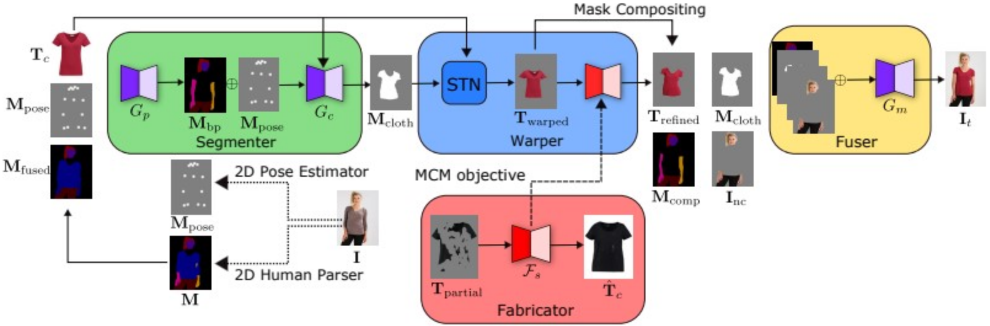

Figure 3: Schematic layout of FIFA framework[3]

[3]H. Zunair, Y. Gobeil, S. Mercier, *et al.*, "Fill in fabrics: Body-aware self-supervised inpainting for image-based virtual try-on,", 2022. DOI: 10.48550/ARXIV.2210.00918. [Online]. Available: https://arxiv.org/abs/2210.00918.

Experimental Results: Self Trained Model

(a) Input image **(b)** Input cloth **(c)** Output image

Figure 4: Applying self trained FIFA try-on model when trained with 100 image dataset.

Experimental Results: Self Trained Model

(a) Input image **(b)** Input cloth **(c)** Output image

Figure 5: Applying self trained FIFA try-on model when trained with 14,221 image dataset.

(a) Easy level image

(b) Medium level image

(c) Hard level image

Figure 6: Different level of images

Experimental Results: Models Output On Easy Level Image

(a) ACGPN model **(b)** FIFA model **(c)** Self trained model

Figure 7: Easy level output images

(a) ACGPN model **(b)** FIFA model **(c)** Self trained model

Figure 8: Medium level output images

(a) ACGPN model

(b) FIFA model

(c) Self trained model

Figure 9: Hard level output images

Structural similarity index measure (SSIM) is used to measure the similaity between synthesized images and ground-truth.

Method	SSIM
ACGPN	0.845
FIFA	0.886
Self-trained FIFA	0.818

Table 1: Performance comparison of ACGPN, FIFA and Self-trained FIFA on VITON Dataset

(a) Input image1

(b) Input cloth

(c) Output image1

Figure 10: Generating output image1 using ACGPN model with input cloth and input image1

Experimental Results: Identity Loss

(a) Input image2 (b) Input cloth (c) Output image2

Figure 11: Generating output image2 using ACGPN model with input cloth and output image1

Consider we have m number of models, represented as I_1, I_2, ..., I_m and n number of target dresses T_1, T_2, ..., T_n. Let ϕ denotes the trained pipeline of fill in fabrics model composed of segmenter, wrapper and fuser.

$$\hat{I}_{ij} = \Phi\left(I_i, T_j\right); \quad i = 1, 2, \ldots, \quad m; j = 1, 2, \ldots, n$$

$$\mathcal{L} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} \left\| \Phi\left(\hat{I}_{ij}, T_k\right) - \hat{I}_{ik} \right\|_2^2$$

Experimental Results: Identity Loss

Input1	Input2	MSE
Output image 1	Output image 2	14.23
Output image 2	Output image 3	09.33
Output image 3	Output image 4	12.09
Output image 4	Output image 5	11.16
Output image 5	Output image 6	14.44
Output image 6	Output image 7	08.11
Output image 7	Output image 8	17.22
Output image 8	Output image 9	12.84

Table 2: Mean Squared Error (MSE) of ACGPN models on output image generated in a sequence

Conclusions

- We examine a pre-trained Adaptive Content Generating and Preserving Network (ACGPN) model.
- We train the Fill In Fabrics (FIFA) model with a small set of 100 images and the entire set of $14,221$ training images.
- We have observed the structural similarity index measure using 2032 test images when we train the model with the whole dataset.
- Examined the abilities of the ACGPN and the FIFA models to preserve identities using an identity loss.

Future Plans

- Comparing the images generated by the model, we confirm that there is some identity loss. So we will implement the loss measure as defined in to train the model.
- Train the model on Zalando-Dataset, it contains $34,928$ frontal-view human (including man and woman) and clothing.

References

[1] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating↔preserving image content," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7847–7856. DOI: 10.1109/CVPR42600.2020.00787.

[2] H. Zunair, Y. Gobeil, S. Mercier, and A. B. Hamza, "Fill in fabrics: Body-aware self-supervised inpainting for image-based virtual try-on,", 2022. DOI: 10.48550/ARXIV.2210.00918. [Online]. Available: https://arxiv.org/abs/2210.00918.