Virtual Stylist: A Recommendation Based Virtual Try-On System

Sukanya Saha

I. INTRODUCTION

Have you ever looked at an Instagram model and wondered - "This summer look is a so gorgeous! how would it look on me?"(see fig 1). This problem requires an application that would recommend similar products and let customers try on these items digitally. This paper proposes a novel computer vision-based technique called Virtual Stylist to tackle two emerging problems of fashion e-commerce platforms- 1. Recommend similar fashion products, 2. Virtual Try-On based on those products. This can be beneficial for e-commerce giants like Amazon, Macy's, Walmart, Gap, Nordstrom, etc. to provide customers with a platform to try out different fashion looks, outfits based on a customer image and a style image.

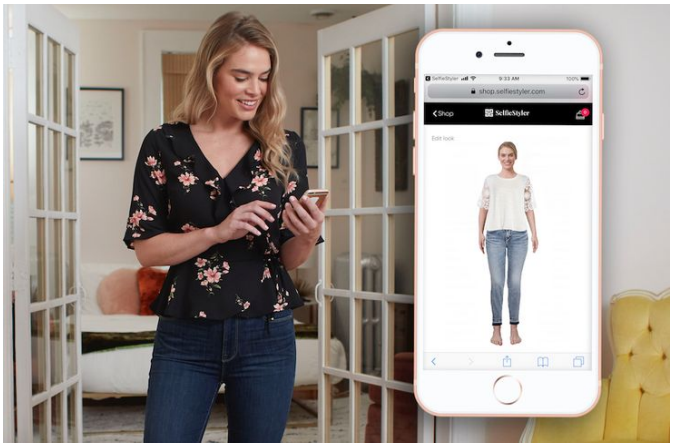

Fig. 1. Virtual Stylist [2]

Existing work on fashion products offerings focus either on recommending similar products or synthesizing images from a customer and product image. However, little has been done on an end-to-end system on recommending products as well as providing a virtual Try-on platform. Additionally, while an increasing number of studies have been conducted, the resolution of synthesized images is still limited to low. This acts as the critical barrier against satisfying online consumers. While recent works in virtual clothing try-on feature a plethora of possible architectural and data representation choices, they hardly improve the quality of generated images. This paper will combine a recommendation task and a virtual try-on task and generate high-resolution images to satisfy customer needs.

This paper introduces a baseline approach for building a virtual stylist system incorporating deep learning, generative adversarial network, and recommender systems. This architecture consists of two cascaded networks i) Fashion recommender network and ii) Image Synthesizer Network. The fashion product recommender network is based on human keypoint detection and object recognition. Whereas the image synthesizer network is a self-supervised model that given a reference image, I $\in R^{3\times H\times W}$ of a person and a clothing image, $c \in R^{3\times H\times W}$ (3, H and W denote the rgb channels, image height and width, respectively). The goal of this network is to generate a synthetic image, $\hat{I} \in R^{3\times H\times W}$ of the same person wearing the recommended target clothes c, where the pose and body shape of I and the details of c are preserved. In this model clothing-agnostic person representation is used to retrieve the pose map and the segmentation map of the person to eliminate the clothing information in I and finally generate $\hat{I}$). For experiments and evaluation of photo-realism, this paper utilizes the Deep Fashion dataset [8] and the High-Resolution Virtual Try-On (VITON-HD) dataset [7].

Fig. 2. Virtual Try On [3]

Finally, one application of this project can be social media advertisement eg. an Instagram fashion influencer want to

recommend a look for their followers to try and redirect the customers to the recommended product pages. This project can also open many doors for future research on recommendation based virtual try-on. For this project, both of these images should have overlapping body regions where the product needs to be fitted. Here, the challenge is synthesizing occluded body parts after fitting the recommended clothes on the customer image. Also, it is important to recognize a person's pose, body shape, and identity and deform the clothing product based on the person's posture without losing product details.

II. Related Work

A. Recommending Similar Fashion Products

Researchers over the years have proposed different methods that recommend similar fashion items given a user query or/and Product Display Image. This approach aims to recommend similar products for all fashion items and can boost customer engagement. I will talk about two major topics in recommending fashion products-

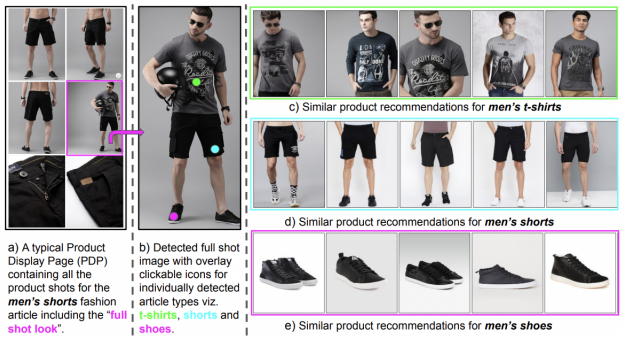

Fig. 3. Buy Me That Look: products recommendation [18]

1) Human Keypoint Estimation: Some of the noticeable works in keypoint estimation are Cascaded Pyramid Network (CPN) [6] , which has been dominant on the COCO 2017 key-point challenge. The Hourglass method [16] played a vital role in the MPII benchmark [4]. The CMU-Pose method [5] used a bottom-up approach that makes use of Part Affinity Fields. However, the Simple baseline pose estimation paper [24] outperformed the above methods by using optical flow-based pose propagation and similarity measurement. This method has been used in the Buy Me That Look [18] model whose architecture I am using in this paper.

2) Object Detection: The object detection task involves not only recognizing and classifying every object in an image but also localizing each one by determining the bounding box around it. Deep learning has been widely used for object detection. Researchers have been keen to use darknet architecture-based single-stage detectors like YOLO [19] for the task of real-time object detection. However, this project's goal of object detection does not require real-time output hence, this component of my pipeline can be done offline. So, the best choice would be to pick a model with a better mean Average Precision (mAP) score, while disregarding the run-time latency. The Mask RCNN [10] method has been chosen for this purpose as proposed by the Buy Me That Look [18] model.

3) Embeddings Learning: Embedding learning [25] [21] aims to learn representations of raw images so that similar images are grouped while moving away from dissimilar ones. My work follows the proposed method by the Buy Me That Look model that makes use of embedding learning to obtain representations, and compute image similarity.

My paper makes use of the proposed method mentioned above for more sophisticated application of product recommendation based on any Product model image given by the user as opposed to querying from existing database images. Additionally, this paper also synthesises the customer's look based on the recommended product which previous models did not take into account.

B. Virtual Try-On via Generative Adversarial Networks

There have been a large number of publications since the inception of virtual try-on methods in 2017. The focus areas of the methods are to parse human, segment products and body, estimate pose. Pose information has been embedded through DensePose model [20] and there exists impressive architectures for body and product segmentation.

1) Conditional Image Synthesis: In recent works of Conditional generative adversarial networks (cGANs) [15] researchers have utilized class labels [17] and attributes [22] to improve the image generation process. Some recent cGAN papers have tried to generate high resolution images. But, it causes blurry image generation when trying to handle large spatial deformation from input to output image. Moreover, methods using conditional GANs for try-on models were iteratively improved by removing adversarial loss, introducing perceptual loss, and experimenting with different extended network architectures to synthesize more detailed virtual try-on outputs. Most of the proposed approaches for these methods follow a two-stage architecture that involves cloth warping and person rendering. A two-stage approach has enough expressiveness for the model to learn how to do virtual try-on. In this paper, I propose a method that can address the spatial deformation of input images and properly generate customer images with the recommended products.

Picture: *These people are not real – they were produced by our generator that allows control over different aspects of the image.*

Fig. 4. StyleGAN [12]

2) Image-to-Image Translation and StyleGAN: Deep Generative Adversarial Networks and StyleGANs [12] (Figure 4) have shown great potential for synthesizing high-quality photo-realistic images. It is very efficient and accurate in generating high resolution realistic images and can be done using the pre-trained model. My work focuses on designing a pose-conditioned GAN with precise control on the localized appearance (for virtual try-on) and pose (for reposing). Image-to-Image Translation [11] provides a general framework for mapping an image from one visual domain to another. Recent advances include learning from unpaired dataset [26], extension to videos. Similar to many existing human reposing methods, my work can be thought of as an image-to-image translation problem that maps an input target pose to an RGB image with the appearance from a source image. This paper uses spatial modulation in StyleGAN for detail transfer.

III. Methods

This architecture consists of two cascaded networks i) Fashion Image recommender network and ii) Image Synthesizer Network. First I will talk about the Fashion image recommender network.

A. Fashion Image recommender network-

This network architecture can be derived in four modules as stated below. In this part of the architecture, I followed the architecture that is mentioned in the Buy me that look [18] paper(Figure 5).

Front-facing Full-shot Image Detection- In this module, all full shot images are detected using a pose detection classifier then I look for all front-facing images.

Fashion article Detection and Localisation- Images from stage-1 are passed in stage-2 a CNN network with active learning detects the fashion objects in the image and also does localization.

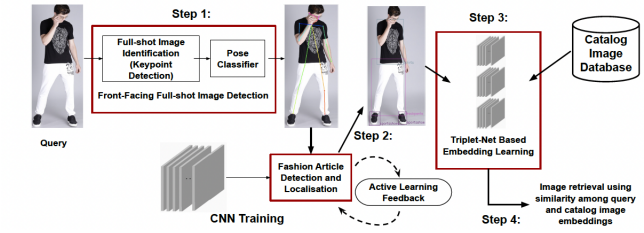

Fig. 5. Recommender Network Architechture [18]

Embedding generation for article types- In this module, embeddings are made of all images available in the catalog and stored in the database. Then, I am retrieving similar images from the database to the query image.

Since the steps/modules are loosely connected so there is no need not to stick to algorithms that are used in the paper. Hence, these steps can be easily reproduced by experimenting with a new set of algorithms for the new dataset. One can use this as a base architecture and try various sets of algorithms that fit this or a new dataset.

B. Image Synthesizer Network: Pose-conditioned StyleGAN2-

I am using the model architecture proposed by the TryOnGAN [13] paper. Generative Adversarial Networks (GANs)[9] have been shown to synthesize impressive images from latent codes. As discussed in related work section, the idea to combine progressive growing and adaptive instance normalization (AdaIN) with a novel mapping network between the latent space, Z, and an intermediate latent space, W, encouraged disentanglement of the latent space. Transforming intermediate latent vectors wW into style vectors s further allowed different styles at different resolutions. Furthermore, recently developed StyleGAN2 inversion methods enable the projection of real images into the extended StyleGAN2 latent spaces, Z+ and W+ [23]. Motivated by those advances they choose StyleGAN2 as the base architecture. Hence, I am using this pre-trained StyleGAN2 model on fashion images, with key modifications on pre-condition(Figure 6).

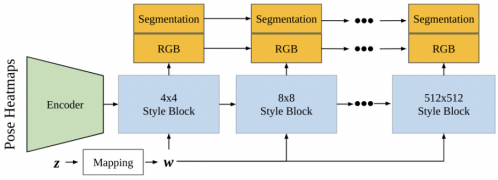

Fig. 6. Modified StyleGAN2 Architechture [13]

IV. Experimental Design-

A. Dataset

I am using the DeepFashion2 [8] for training the recommendation system and the VITON-HD [7] dataset for training the try-on module. Both the dataset has high-resolution images with VITON-HD having 1024×768 virtual try-on data. DeepFashion2 is a benchmark dataset curated for clothes detection, pose estimation, segmentation, and retrieval. It has 801K clothing items where each item has rich annotations such as style, scale, viewpoint, occlusion, bounding box, dense landmarks. There are also 873K Commercial-Consumer clothes pairs. The VITON-HD dataset has been generated by crawling 13,679 frontal-view women and top clothing image pairs on an online shopping mall website. Since this datasets are huge and it will take weeks to train models on these, I am using 10% sample of these datasets.

B. Fashion article Detection and Localisation

Inside the Fashion image recommender network, I am training the Mask RCNN on a custom training on the DeepFashion2 dataset to generate the bounding box location and classification from across different apparel types. I am able to obtain around an average mAP(Mean Average Precision)

of 65% for all the classes while reaching as high as 80% for some of the top-wear classes (shirts and t-shirts). These values are in line with this model's performance on the Microsoft Common Objects in Context (MS COCO) data set [14], which is around 60.3% at IOU 0.5. This is justified because the COCO dataset has more classes with natural real-world images, whereas images in my case are from a lesser number of article categories.

C. Discriminator-

In the Image Synthesizer Network, I have two discriminators, one for pose and one for segmentation. The pose discriminator receives as input either a real RGB image/pose heatmap pair or a conditional pose heatmap with the corresponding generated RGB image. The segmentation discriminator receives real/generated RGB image/segmentation pairs. The two discriminators are weighted equally during training. To prevent overfitting of pose to style, I am using the following data augmentations on the pose input to the discriminator: 1) add gaussian noise to the normalized keypoint locations before creating the heatmap, and 2) drop keypoints with probability less than 0.4.

D. Retraining StyleGAN2-

TryOnGAN paper's conditional StyleGAN2 network was implemented in TensorFlow with 25 million iterations, on 8 Tesla v100 GPUs, for 12 days. Once the network was trained, they performed a hyperparameter search for the optimization loss weights. Since I am using a smaller sample of the dataset, I experimented with these parameters- batch size, gamma, augmentation. After trying different values for these parameters I am choosing these values which gave the best results- batch size = 32, gamma=5, augmentation= ada [12].

E. Evaluation Metrics-

In the Fashion image recommender network, for Front-facing Full-shot Image Detection Results, I make use of a ResNet18 network as the pose classifier to classify an image into one of the following categories: front, back, left, right, or detailed shot. For this part, I am using precision, recall for each of the categories and mAP(Mean Average Precision) since these work best for classification model evaluation.

For the Fashion article Detection and Localisation results, I am treating it as a classification task of bounding box with evaluation metrics as IOU of threshold 0.5 for measuring ground truth as IOU works best for object recognition related tasks. Results on embedding generation for article types are judged using cosine similarity among embeddings.

For Image Synthesizer Network, I am using two metrics to compare the methods and types of images: FID(Frechet Inception Distance) [1] to evaluate photorealism and ES (Embedding similarity) to evaluate the quality of try-on or how similar is the result to the input in the garment part. These two metrics efficiently evaluate the photorealism and garment quality of the Image Synthesizer Network as proposed in the TyOnGan paper.

REFERENCES

[1] frechet-inception-distance. https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/. Accessed: 2021-11-17.

[2] Virtual fitting room. https://in.pinterest.com/pin/345369865172616862/. Accessed: 2021-11-09.

[3] Virtual try-on. https://venturebeat.com/2020/06/05/amazons-new-ai-technique-lets-users-virtually-try-on-outfits/. Accessed: 2021-11-09.

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields, 2017.

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation, 2018.

[7] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021.

[8] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[13] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[17] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 06–11 Aug 2017.

[18] Abhinav Ravi, Sandeep Repakula, Ujjal Kr Dutta, and Maulik Parmar. Buy me that look: An approach for recommending similar fashion products, 2021.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[20] Iasonas Kokkinos R{iza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. 2018.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[22] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation, 2017.

[23] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space, 2020.

[24] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking, 2018.

[25] Chin-Chia Michael Yeh, Dhruv Gelda, Zhongfang Zhuang, Yan Zheng, Liang Gou, and Wei Zhang. Towards a flexible embedding learning framework, 2020.

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.