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.

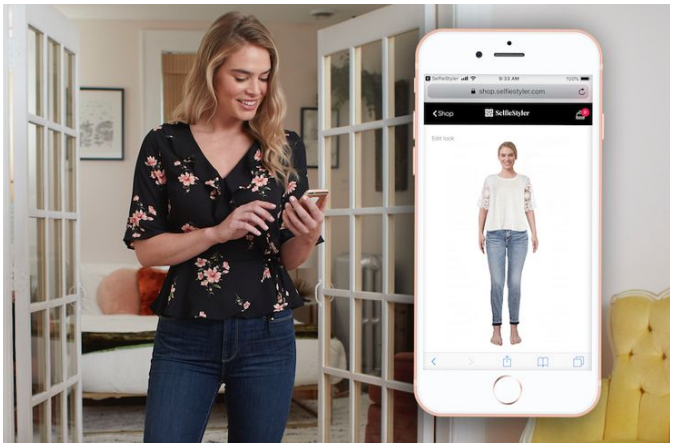

Fig. 1. Virtual Stylist [1]

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 [6] and the High-Resolution Virtual Try-On (VITON-HD) dataset [5].

Fig. 2. Virtual Try On [2]

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 a few major topics in recommending fashion products-

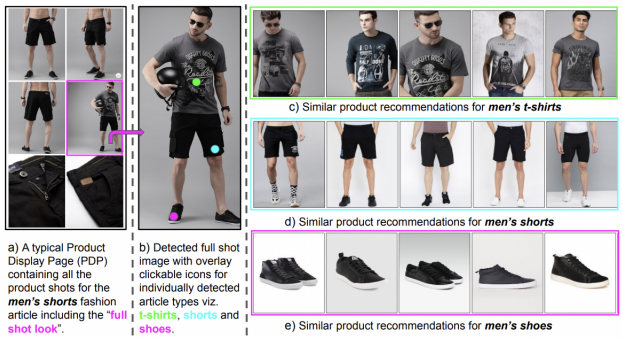

Fig. 3. Buy Me That Look: products recommendation [14]

1) Human Keypoint Estimation
   a) Cascaded Pyramid Network for Multi-Person Pose Estimation [4]
   b) Pose Recognition with Cascade Transformers [3]
   c) Simple Baselines for Human Pose Estimation and Tracking [20]

   Out of these four methods, the Simple baseline pose estimation paper [20] outperformed the other methods by using optical flow-based pose propagation and similarity measurement. This method has been used in the Buy Me That Look [14] model whose pre-trained model has been utilized in my paper.

2) Object Detection
   a) You Only Look Once: Unified, Real-Time Object Detection [15]
   b) Mask R-CNN [8]
   c) Fast R-CNN [7]

   This project's goal of object detection omits the requirement of real-time output hence, this component of pipeline can be done offline. The best choice would be to pick a model with a better mean Average Precision (mAP) score, while disregarding the runtime latency. The Mask RCNN [8] method has been chosen for this purpose instead of other methods.

3) Embeddings Learning
   a) Towards a Flexible Embedding Learning Framework [22]
   b) FaceNet: A unified embedding for face recognition and clustering [17]
   c) Buy Me That Look: An Approach for Recommending Similar Fashion Products [14]

   Embedding learning [22] [17] 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 [14] that makes use of embedding learning to obtain representations, and compute image similarity.

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 [16] and there exists impressive architectures for body and product segmentation. Here are few related topics -

1) Conditional Image Synthesis
   a) Conditional generative adversarial networks (cGANs) [12]
   b) Conditional Image Synthesis with Auxiliary Classifier [13]
   c) Learning Residual Images for Face Attribute Manipulation [18]

   In recent works of cGANs[12], researchers have utilized class labels [13] and attributes [18] 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. 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.

2) Image-to-Image Translation and StyleGAN
   a) Deep Generative Adversarial Networks and StyleGANs [10]
   b) Image-to-Image Translation with Conditional Adversarial Networks [9]
   c) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [23]

Deep GANs and StyleGANs [10] have shown great potential for synthesizing high-quality photo-realistic images. This and models like Image-to-Image Translation [9] provides a general framework for mapping an image from one visual domain to another. Recent advances include learning from unpaired dataset [23]. Similarly this 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.

3) Attention and Activations Functions
   a) Attention Is All You Need [19]
   b) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [21]
   c) ShineOn: Illuminating Design Choices for Practical Video-based Virtual Clothing Try-on [11]

   Recent works in Attention and transformers Attention Is All You Need [19] [21] have been thriving in outperforming state-of-the-arts with extended and improved architectures especially in the machine translation task. But only a handful of papers have been published in the try-on spectrum utilizing self-attention techniques. I propose to use a self attention model as proposed by the ShineOn paper [11] to utilize its ability to attend to visual and spatial regions of importance and it has been placed where the feature map depth is greatest.

REFERENCES

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

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

[3] 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.

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

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

[6] 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.

[7] Ross Girshick. Fast r-cnn, 2015.

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

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

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

[11] Gaurav Kuppa, Andrew Jong, Vera Liu, Ziwei Liu, and Teng-Sheng Moh. Shineon: Illuminating design choices for practical video-based virtual clothing try-on, 2021.

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

[13] 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.

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

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

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

[17] 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.

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

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

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

[21] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

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

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