Virtual Cloth Warping using Deep Learning

By

Surya Madhavan
Electronics and Telecommunications Engineering, SIES Graduate School of Navi Mumbai, India
Email: surya.m18@siesgst.ac.in

Dr. Preeti Hemnani
Electronics and Telecommunications Engineering, SIES Graduate School of Navi Mumbai, India
Email: preetih@sies.edu.in

Anjana Ashokkumar
Electronics and Telecommunications Engineering, SIES Graduate School of Navi Mumbai, India
Email: anjana.ashok18@siesgst.ac.in

Manasi Deshpande
Electronics and Telecommunications Engineering, SIES Graduate School of Navi Mumbai, India
Email: manasi.rajendra18@siesgst.ac.in

Shamika Aslekar
Electronics and Telecommunications Engineering, SIES Graduate School of Navi Mumbai, India
Email: shamika.prasad18@siesgst.ac.in

Abstract

Virtual Try-On is a technology that can realistically clothe an individual virtually, it transfers a clothing image onto a target person's image. This is attracting attention from the industrial and research centers and can make the in-store experience achievable. The 2D image-based and 3D model-based methods developed recently have their own benefits and limitations. This paper describes the development of a 2D image-based Virtual Try-On Clothing system. Our solution comprises major modules: Human representation which is pose estimation using OpenCV and human parsing using Self Supervised Joint Body Parsing and Pose Estimation Network (SS-JPPNet) and the Try-On module which utilizes Cloth Warping Module (CWM) and Cloth Fusion Module (CFM) to generate the final try-on output. The technologies that are used for CWM is Convolutional Neural Network and for CFM is Generative Adversarial Network. Our application as of now supports only upper body clothing (tops, t-shirts, etc.). A graphic user interface is created where one can virtually try on clothes by uploading their picture and selecting the clothing item to try on, bringing the shopping experience to one's doorstep.

Keywords– virtual try-on, convolutional neural network, generative adversarial network, cloth-warping

Introduction

As predicted by retail consultants and industry experts, the pandemic has changed the way people shop. Stores are reopening but are reoriented to avoid interaction by closing fitting rooms, sample counters, and testers. Choosing the appropriate designs, sizes, and fits may make or break a purchase. This brings the issue of trying on garments, as dressing rooms have the main role in real-life shopping trips. Customers will think twice about the products they purchase, and vast inventories of stock on the shop floor will make way for new, tech-driven experiences with Virtual Try-On technology at the forefront. Using our proposed model solution retailers and fashion brands can blend the physical and virtual systems to create a customer experience that is safe, easy, convenient, and efficient for customers, whether online or in-store.

Motivation

Global lockdowns meant shoppers were unable to buy products in-store. Even when those retail locations opened up again, people were more hesitant to try on clothing items, worried it increased their risk of exposure to the virus. Virtual fitting rooms empower customers to make more informed purchasing decisions, it can alleviate many of the challenges faced by the fashion industry today: struggles with the fit, sky-high returns, and the resulting impact on our environment that could help brands to achieve their conversion goals.

Research Objectives

The Primary objective of this study is to understand the working of Virtual Cloth Warping Systems and develop one.

- To develop an approach with Gender Independent, Cloth Warping in mind.
- To improvise on the Warping / Fusion issues of the precedent research works.
- To try Generative Adversarial Networks to be utilized to its fullest.

Proposed System

Virtual Cloth Warping can be implemented using various technologies like Convolutional Neural Network(CNN), Generative Adversarial Network(GAN), and Cloth Interactive Transformer(CIT). Our proposed system is based on two stages, the first stage is a Convolutional Neural Network called a Clothes Warping Module and the second stage is a Generative Adversarial Network called Clothes Fusion Module.

Related Work

Cloth Interactive Transformer for Virtual Try-On

A virtual try-on proposed consists of a novel two-stage Cloth Interactive Transformer (CIT) for image-based virtual try-on tasks. Their work is the first to utilize a Transformer for this task. In the first stage, a transformer-based matching block can model global long-range relations when warping a cloth via learnable thin-plate spline transformation and as a result, the warped cloth can be more natural [2]. The reasoning block can strengthen the important regions within the input data, on the other hand, the mutual interactive relations established via the reasoning block further improve the rendering process to make the final Tryon results more realistic.

VITON: An Image-based Virtual Try-on network

The Network proposes to help in visualizing the person in target clothing using 2D resources. It initially generates an image with the person having the same pose and in the target clothing. and it is trained to improvise on the blurry portions formed during the generation process [4], [14]. The Encoder-Decoder generator and Refinement network both work on Convolutional ***neural networks.***

VTNFP: An 2D-based Try-on Approach

In this paper, a three-module approach is proposed wherein the target clothing is warped and a segmentation map is predicted based on the previous results, and a synthesis module is used to fuse the target cloth & input human image. [20]

Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔ Preserving Image Content (ACGPN)

The last approach [4] produces a semantic layout of a rough image that needs to be modified after try-on and then decides if the information of the image needs to be produced or preserved giving rise to fine results with details included. The first module uses semantic segmentation to collectively predict the needed semantic layout after try-on. Second, the warping of clothes is done according to the produced layout. In the third module, all information is gathered to do the clothes fusion and produce the output.

It consists of U-Net generators and all the discriminators are from pix2pixHD. Second-order spatial transformation is used in this approach to prevent Logo distortion and retain the characters making the model more fine-tuned. The first module specifies and preserves the unchanged portions of the image directly. It is made for fitting the clothes into the shape of clothing items with perceptible natural distortion according to the pose estimation. For the second module, a coarse body shape is created and used as a reference to produce the final output.

Some problems which occurred in [2] are overcome in this approach which are high distortion and misaligned warped clothes, and networks responsible for blending cannot retain the remaining clothes due to improper human representation. This approach has two stages which are the Geometric Matching Module and Try On Module. In the first stage, the target clothing is warped around the target human and in the second stage, the wrapped clothing is blended with the target person's image. The first stage is very important to get the target body silhouette from the target person's image.

Toward characteristic preserving image-based virtual try-on network

This approach [20] uses 2D image synthesis methods and 3D model deformation methods to target person pose. This approach can be applied to various clothing categories. The Skinned Multi-Person Linear (SMPL) model is used for the reconstruction of clothes with various poses. It is a fusion method in which 3D warped clothes are blended with 2D human poses to generate accurate outputs while preserving the original pixel quality of the image.

Methodology

Preprocessing

Human Parsing

Human parsing refers to the partitioning of the person or people captured in an image into multiple semantically consistent regions, e.g., body parts and clothing items. To implement Human Parsing we compared SCHP (self-correction for human parsing) and SS-JPPNet (Self Supervised Joint Body Parsing). The results obtained in SS-JPPNet were found to be better and more reliable than in SCHP. The recommended model to use for the generation of parsed human images is Body Parsing using SS-JPPNet (Self Supervised Joint Body Parsing and Pose Estimation Network) which is a deep learning method for human parsing built on Tensorflow. The model used is known as Self Supervised since the model is capable of generating “Structure-Sensitive” losses on its own without any additional information, these losses are used to improve the accuracy of parsing. DeepLab Model, which is a widely used Machine

Learning library has been used here with ResNet-101 and Attention as the primary network. In the ResNet model, the input sent to a layer is also sent as input to its subsequent layer, thus they help in solving the problem of vanishing gradient. The model is trained on the LIP dataset and tested on our custom dataset. Consider an image I, we define joint configurations $C_I^P = \{c_i^p \mid i \in [1, N]\}$, c_i^p is the heatmap of the i-th joint. C_I^{GT} I = $\{c_i^{gt} \mid i \in [1, N]\}$, obtained from the parsing ground truth respectively. N is a variable that defines the number of joints in the input images of human bodies. If there are joints missing in the image, the heatmaps are replaced with maps filled with zeros. The joint structure loss is the Euclidean (L2) loss, which is calculated as

$$L_{\text{Joint}} = \frac{1}{2N} \sum_{i=1}^{N} \| c_i^p - c_i^{gt} \|_2^2 \quad (1)$$

Here, c_i^p is the heatmap of the i-th joint and c_i^{gt} refers to the heatmap of ground truth.

The final structure-sensitive loss, denoted as $L_{\text{Structure}}$, is the combination of the joint structure loss and the parsing segmentation loss, L_{Parsing} is the pixel-wise softmax loss calculated based on the parsing annotations.

$L_{\text{Structure}} = L_{\text{Joint}} \cdot L_{\text{Parsing}}$ (2)

The proposed SS-JPPNet significantly improves the performance of the labels such as arms, legs, and shoes, which demonstrates its ability to refine the ambiguity of left and right. The overall accuracy of SS-JPPNet is 84.53%. The mean IoU score is 44.59.

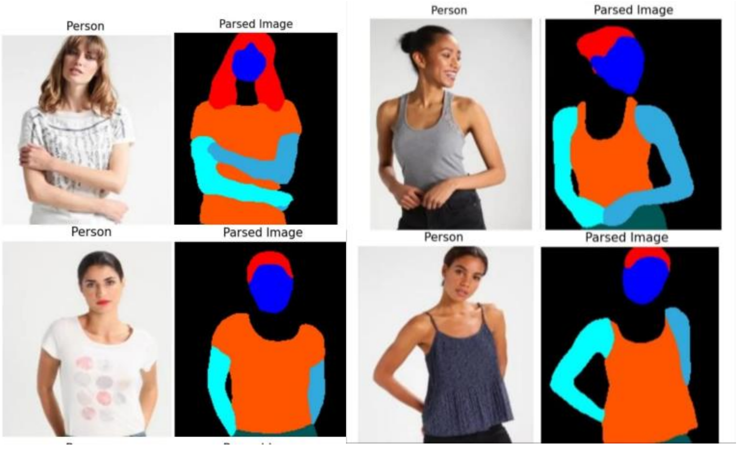


Figure 1: *Human Parsing Results*

Fig.1 shows the input image selected from the dataset and its parsed image. The images in the dataset are of women wearing different sizes, colors, and patterns of clothes in various poses. The image on the left-hand side is the person image and on the right side is the parsed image.

Pose estimation

It is a technique used to track various movements of human beings as well as objects. It is usually performed by finding the location of critical points for a given image. Based on these critical points various poses can be compared to draw conclusions.

For the execution of the pose estimation model, MediaPipe is used. The MediaPipe Body landmark model gives high-fidelity body pose tracking. When an image is passed through the pose estimator model, pose landmarks are obtained, and an array of key points is marked on the image using various draw utiles. Pose estimation deals with labeling each image pixel-wise semantically along with joint-wise structure prediction. MediaPipe uses the BlazePose model in which 33 key points can be marked on the input image [21] but some modifications have been made to get 18 specific key points that are required in the output.

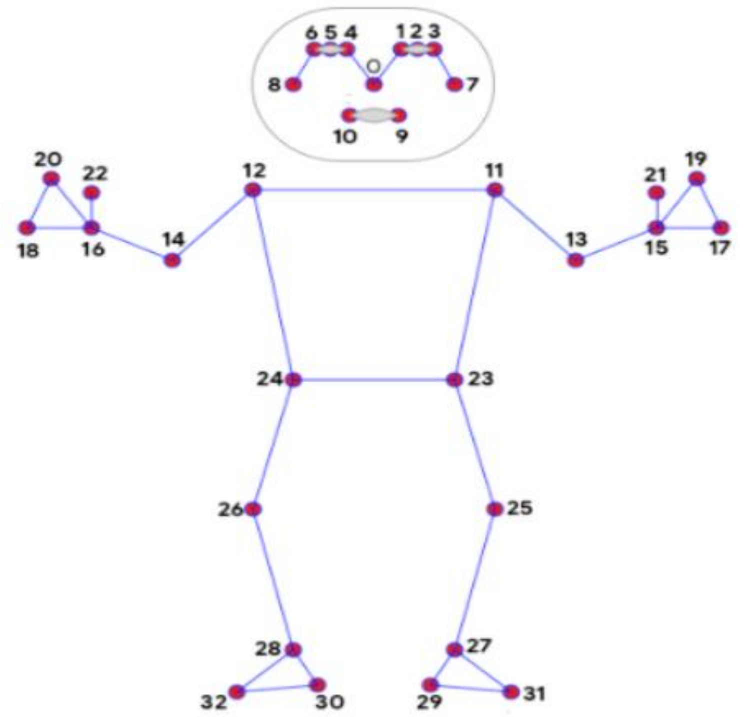


Figure 2: *BlazePose topology*

Fig.2 shows the BlazePose topology which is a pose tracking solution with 33 key points and is used by MediaPipe for pose detection purposes.

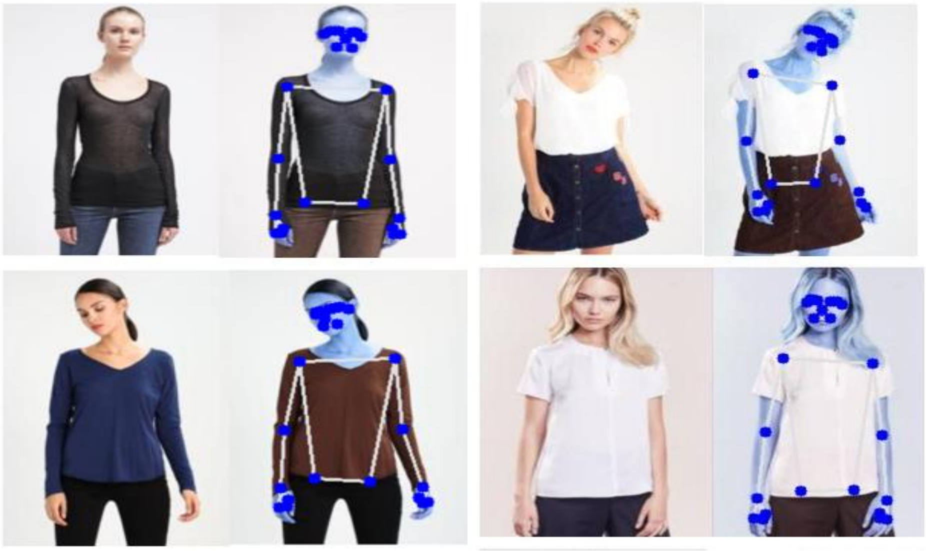


Figure 3: *Pose Estimation Results*

Fig.3 shows the image from the dataset.The image on the left side is the input image and on the right side is the final output after performing the pose estimation. The images are converted to RGB for ease of processing. The blue dots that are seen on the image are the required key points.

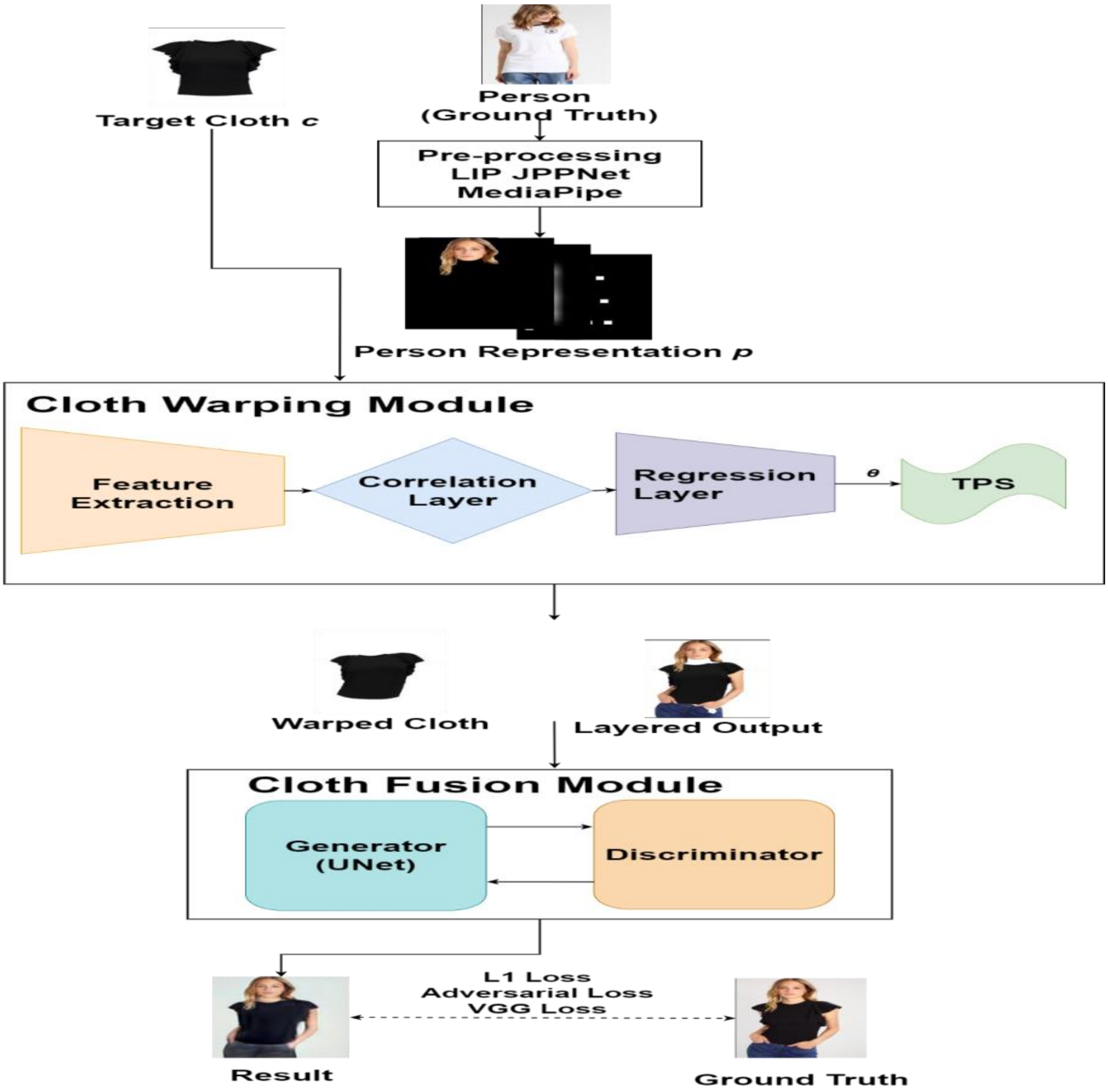


Fig 4: *Flowchart*

Fig.4 shows the flow chart of the proposed model.

Cloth warping module

The Fusion of Cloth onto the person's image is research in itself, since it requires an understanding of human bodies and categories of shapes and how they can be utilized to ensure a piece of cloth, a two-dimensional image can be fused well. Since the traditional use case would involve two-dimensional images, the model developed here utilizes the two-dimensional properties and tends to build the logic over it.

Cloth Images can not be fused simply by overlapping the image over the person's image. There is a need for segmenting the picture of the person on the basis of regions (skin, hands, head, etc.) and additionally estimating the pose, in order to form a clear understanding as to how the dress needs to fit over the person. This part of segmentation and pose estimation is done with the help of preprocessing the images.

Further, the model is divided into two major steps - a model trained to warp the two-dimensional cloth image based on the features of the person (arm size, body shape, pose, etc), and another model which is trained to develop the nearest possible result that would look similar to the person actually “trying on” a cloth.

The Cloth Warping stage needs the information of the person image, i.e. its feature details, and thus, a network is trained to extract features from the person and cloth images. Following the features, a layer is used to combine the tensors onto a single feature. This Feature is used as an input, based on which, the level of warping/warping parameters would be determined. This conversion of feature tensor to a parameter rating for warping is performed with the help of a regression neural network.

The Cloth Warping Module takes Person representation *p*, target cloth *c* as input and it comprises 3 stages: (1) Feature Extraction Networks for *p* & *c*, (2) Correlation layer to combine the results onto one tensor to pass it as input to regressor network, (3) Regression network to predict the transformation parameters θ. The parameters are passed onto the Thin Plate Spline Transformation Module, which warps the cloth accordingly.

The above-mentioned pipeline is learnable end-to-end and the sample triplets (p, c, c_t) have been trained, under the pixel-wise L1 loss between the warped result $\hat{c}$ and ground truth c_t, where c_t is the clothes worn on the target person in I_t.

The parameter θ is calculated using the following formula. The Human Representation H_t and the Mask M_{Ci} is used here, instead of the colored mask C_i , to compute

$$\theta = f_\theta(f_H(H_t), f_C (M_{Ci})) \qquad (3)$$

The Approach for Cloth Warping Module of the Virtual Cloth Warping using Deep Learning (VCWDL-CWM) is improvised on certain aspects: the loss function of CWM includes the *L1* distance between the warped (C_{Warped}) and real images (C_t) of cloth on the body. Hence getting the warped cloth & mask of warped cloth as output. Our experiments with existing methods showed that the current warped cloth is distorted and thus we chose to keep Regularization for the estimation of TPS parameters. The pixel-wise L1 loss is formulated as follows.

$$L_{VCWDL\text{-}CWM} = \lambda_1 \cdot L_1(C_{warped}, I_{Ct}) + \lambda_{reg} \cdot L_{reg} \ (4)$$

Cloth Warping stage is crucial since, during the training the model learns how to warp a two dimensional cloth image based on the person and cloth features. This Warped cloth, i.e. Output of the first stage plays a huge role in further processing of the model, without the need for using the previously used features.

Being a retrainable model, it allows a lot of flexibility based on the market fit, datasets available or the target audience. The Output, i.e. the warped cloth image is not directly multiplied over the person, instead, it is taken as a reference and developed over it, based on the body shape of person and the fit of cloth onto the individual.

Cloth Fusion Module

There have been several approaches to fuse the cloth over the person. Primarily, a Network using UNet Architecture with additional loss functions to generate the output with comparison over ground truth to improvise it further. Failure of these approaches has been its ability to learn from the losses calculated. To solve this, a generative adversarial network trained to improvise on the “level” of fusion of cloth onto the human is proposed.

For the Cloth Fusion Module, in-shop clothing image *c*, person representation *p* & CFM output image I_o are taken as inputs & model is put to judge whether the output is real/fake. This module contains Unet Generator, VGGLoss, NLayerDiscriminator and L1 loss.

The Unet generator has a u-shaped architecture that propagates context information to higher resolution layers. Supplementing a usual contracting network by successive layers is the main idea and pooling operations are replaced by non-sampling operators, due to which these layers increase the resolution of the output. The Generator network is trained to produce the “try-on” result using the cat function provided by pytorch.

The Outputs are further segmented into the combination of *cloth* and composition mask (*comp_mask*) and the rendered person (*rendered)*. The Try On Person image is generated by:

try_on_person = cloth(comp_mask) + rendered*(1-comp_mask) (5)*

For the purpose of solving the adversarial networks, the real and fake variables are created. We compute the L1 loss, L1 mask loss and VGG loss for the results obtained from the generator.

Following which, the forward propagation of the discriminator network is processed and the loss is computed for the discriminator performance. The losses for generator and discriminator are calculated and these losses are rectified with the help of backpropagation of the generator and discriminator networks.

The Generator loss(*gen_loss)* is the summation of L1, VGG and Mask loss,

gen_loss = l_l1 + l_vgg + l_mask + l_adv/batch_size (6)

where, L1 loss (*l_l1*) , VGG loss (*l_vgg*) and the loss related to mask (*l_mask*) along with adversarial loss (*l_adv/batch_size*) is appended so as to get a good clarity on how the generator is performing.

The discriminator network is a deep convolutional network containing convolution blocks which are followed by dense fully connected layers, followed by a batch normalization layer. The two dense layers at the end of the network work as a classification block. The last layer is used for prediction to predict the probability of an image belonging to the real dataset or to the fake dataset.

Graphical User Interface

The Model, which comprises of Preprocessing Network (SS-JPPNet[17] and Mediapipe [18]), Cloth Warping Model, and Cloth Fusion Module, is integrated into a pipeline-like structure with a Graphic User Interface (GUI) such that user could access and interact with the model and get results which could visually give them an idea on how the cloth looks on a particular human.

The Graphical User Interface is created with the help of Tkinter library in python. Tkinter is an open source, standard Graphical User Interface library for Python language. User Interface could be built with the help of an inbuilt Tk function.

Results and Discussions

CP-VTON clothing-human pair dataset has been used for all experiments. Dataset split contains 13221 image pairs in the training and 1000 image pairs in testing and 2032 pairs in validation. The Dataset has been procured with the help of different previous approaches and also customized for the suitable use case of our paper and target audience. The Dataset contains images of target cloth and person wearing the same target cloth in the training set, so as to help the model learn how well the warping needs to be done compared to the original image.

For the Cloth Fusion Module, UNet Architecture is used, and a custom N Layer Discriminator with 6 layers. In the Unet generator, parameters used are input channels as 25, output channels as 4 with batch normalization. In NLayerDiscriminator, input channels used are 28, and the norm layer is InstanceNorm2d. The cloth Warping Module was trained for 100 epochs and Cloth Fusion Module for 50 epochs with batch size 16, with the Adam optimizer with β1 = 0.5 and β2 = 0.999. The learning rate was fixed at 0.0001. The Proposed Model was trained using NVIDIA 1050Ti & 1060 Max Q Mobile GPUs.

In the previous approaches of VITON and CP-VTON-PLUS, there were occlusions like blurry arms and misaligned images due to reconstruction loss. To address that, the proposed model uses CNN in the first stage and GAN in the second stage. Adversarial loss is considered and the model has trained adversarially against the discriminator. For evaluation of the model, the metrics chosen are SSIM, and Inception Score as performance metrics to compare other approaches like VITON, CP-VTON, and ACGPN over the same clothing try-on cases. The Proposed Model outperforms VITON, CP-VTON, and ACGPN in all measures.

Table I: *Performance comparison*

	IS	**SSIM**
VITON	2.514 ± 0.13	0.77
CPVTON	2.78 ± 0.05	0.745
ACGPN	2.798 ± 0.12	0.81
Ours	**3.0 ± 0.04**	**0.79**

Table I shows a performance comparison between our model and existing models

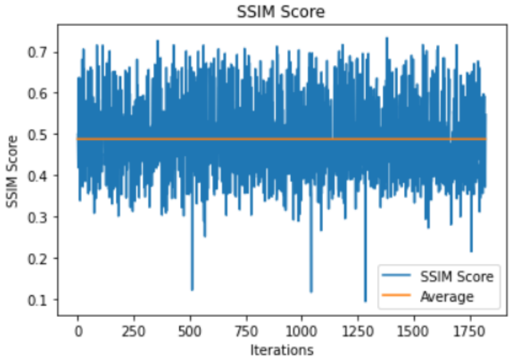


Figure 5: *SSIM Score vs Iterations*

Fig. 5 shows the SSIM score for each iteration along with the average SSIM score.
The Qualitative Results are as follows.

Graphical User Interface Output

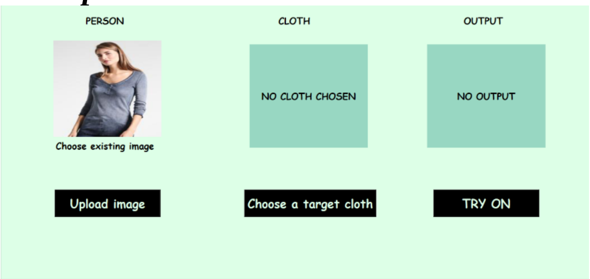


Figure 6: *Final Page of the GUI without output*

Fig.6 shows a Tkinter window letting the user choose the human image and the target clothing.

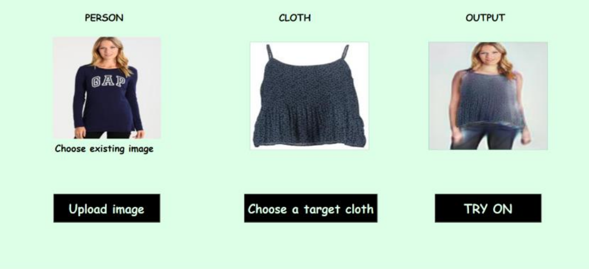


Figure 7: *Final Page of the GUI including result*

Fig.7 shows the Tkinter window where the human image, target clothing item, and the final output are displayed.

First Stage Output

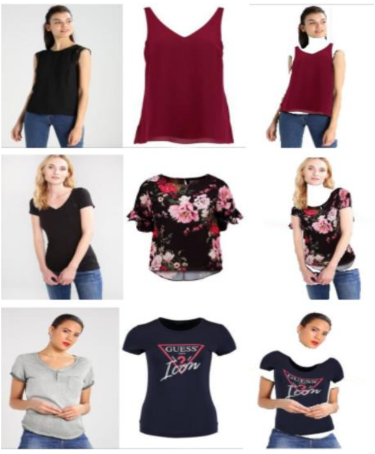

Input Target Cloth Output

Figure 8: First Stage Output

Fig.8 shows the images in following order (L to R): person image, cloth image and cloth overlapped over person image. This is the output of the first stage of the model.

SECOND STAGE OUTPUT

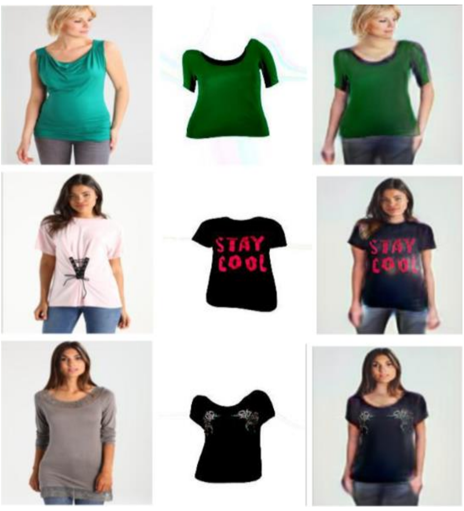
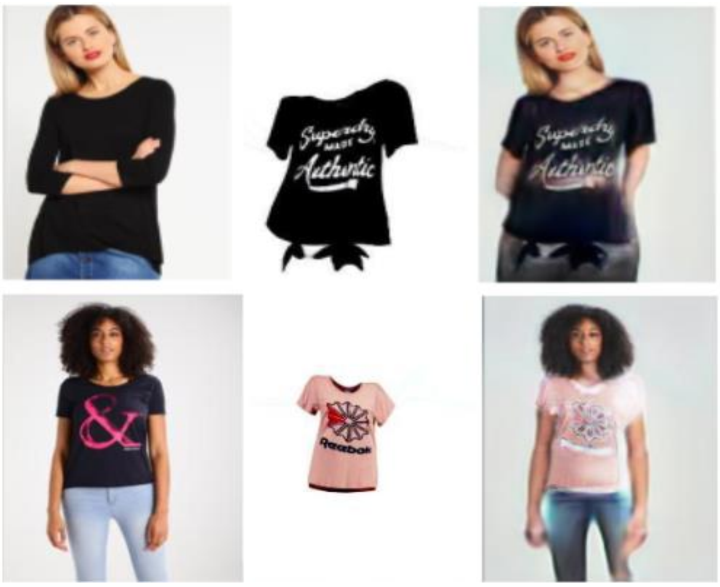

Input Target Cloth Output

Figure 9: *Second Stage Output*

Fig.9 shows the second stage output of the model. The images are in the following order (Left to Right): person image, warped cloth image, and output generated by the second stage of the model.



Input Target Cloth Output
Figure 10: Occlusion Cases

Occlusion cases

Fig.10 shows the cases where our model failed to produce appropriate results. In the first figure, hands are not correctly identified in the output picture. In the second figure, the details of the target cloth are missing in the output and a blurry image is produced.

Statement of contribution

The approach proposed, Virtual Cloth Warping using Deep Learning (VCWDL) has several improvements over the previous approaches. Compared to CPVTON [14], our approach tends to work on a larger spectrum of image datasets and produces better warping and quantitative progress as shown in Table 1. Also, the use of Generative Adversarial Networks (GANs) for this idea is a novel concept and other approaches using GANs [23], use a rather complicated structure that tends to perform inferior as seen in Table 1.

The Idea of using a Convolutional Neural Networks (CNNs)-like model along with a GAN is new. Our team trials noted that, with better access to hardware, this approach can perform qualitatively and quantitatively. Since our Entire Approach ran for a mere 50 epochs, we could see the path it leads to, but as we train it for more epochs and test it with fresh data, one can attain standard results which can be brought to public usage and business prospects.

Conclusion

The proposed VCWDL model consists of two pre-processing stages: pose estimation and human representation and two processing models which are the Clothes Warping Module(CWM) and Clothes Fusion Module(CFM). CWM is based on CNN and CFM is based on GAN. For pose estimation, MediaPipe is used and for human parsing, JPP Net is used. In the processing stage, the CWM gives the warped cloth along with the coarse result, and these go as input to CFM which gives the end result. The training and testing of both stages are done. The accuracy matrix used is Inception Score and SSIM. The SSIM achieved is 0.78 which is higher than CP-VITON and CP-VITON+.

References

[1] Sasadara B. Adikari ,Naleen C. Ganegoda,Ravinda G. N. Meegama, Indika L. Wanniarachchi. Applicability of a Single Depth Sensor in Real-Time 3D Clothes Simulation: Augmented Reality Virtual Dressing Room Using Kinect Sensor. Advances in Human-Computer Interaction,2020.

[2] Bin Ren , Hao Tang , Fanyang Meng, Runwei Ding , Ling Shao , Philip H.S. Torr , Nicu Sebe. Cloth Interactive Transformer for Virtual Try-On, 2021.

[3] Frédéric Cordier, WonSook Lee, HyeWon Seo, Nadia,Magnenat-Thalmann.From 2D Photos of Yourself to Virtual Try-On Dress on the Web,2001.

[4] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis University of Maryland, College Park. VITON: An Image-based Virtual Try-on Network, IEEE Xplore, 2018.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Real-Time multi-person 2d pose estimation using part affinity fields. CVPR, 2017.

[6] Hajer Ghodhbani, Mohamed Neji,Imran Razzak, Adel M. Alimi. You can Try without Visiting,2021.

[7] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. CVPR, 2017.

[8] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han.Deep Fashion3D: A Dataset and Benchmark for 3D Garment Reconstruction from Single Images,2020.

[9] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-topshop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. CVPR, 2012.

[10] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. Virtual fitting by single-shot body shape estimation. 3D Body Scanning Technologies, 2014.

[11] N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people's images. ICCVW, 2017.

[12] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it simple: Automatic estimation of 3d human pose and shape from a single image. ECCV, 2016.

[13] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. arXiv preprint arXiv:1811.08599, 2018.

[14] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic preserving image-based virtual try-on network. Proceedings (ECCV), 2018.

[15] Gokhan Yildirim, Nikolay Jetchev, Roland Vollgraf,Urs Bergmann. Generating high-resolution fashion model images wearing custom outfits. arXiv preprint arXiv, 2019.

[16] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image-based garment transfer. European Conference on Computer Vision, 2018.

[17] Thibaut Issenhuth, Jer´ emie Mary, and Clement Calauzennes. End-to-end learning of geometric deformations of feature maps for virtual try-on. arXiv preprint arXiv, 2019

[18] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. IEEE International Conference, 2017.

[19] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose-guided person image generation. Neural Information Processing Systems, 2017

[20] Ruiyun Yi, Xiaoqi Wang and Xiaohui Xie. VTNFP: An image-based Try-on Network with Body and Clothing Feature Preservation, 2019

[21] Sankeerthana Rajan Karem, Sai Prathyusha Kanisetti, K.Soumya, J.Sri Gayathri Seelamanthula, and Madhurima Kalivarapu. AI Body Language Decoder using MediaPipe and Python, 2021

[22] Liang, Xiaodan & Gong, Ke & Shen, Xiaohui & Lin, Liang. (2018). Look into Person: Joint Body Parsing & Pose Estimation Network and A New Benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2018.2820063.

[23] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, Ping Luo. Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content. arXiv 2020.