Picture-based Virtual Try-On

KHYATI PATEL[1] and DHRUVA SHAH[2]
[1]Master of Computer Science, Lakehead University, Thunder Bay, ON, Canada (e-mail: kpate101@lakeheadu.ca)
[2]Master of Computer Science, Lakehead University, Thunder Bay, ON, Canada (e-mail: dshah33@lakeheadu.ca)

ABSTRACT The popularity of internet shopping is increasing by the day. Although it has several conveniences, the online shopping marketplace does not allow buyers to try-on purchases. Therefore developing a methodology in this fashion for virtually trying would be a better option. Customers and the internet shopping business both would be benefited. Image-based virtual try-on is among the most recent alternatives to virtual fitting that attempts on target clothes in a customer's picture and has obtained considerable attention in recent years. Fitting an in-shop cloth picture to a target model image is required for this assignment. Two phases comprise an appropriate infrastructure for this: (1) warping the try-on fabric to align with the target model's body form and posture, and (2) an image composition module to flawlessly blend the warped try-on cloth onto the target model picture. Moreover, there are several obstacles to overcome associated with developing virtual try-on that make it extremely difficult to accomplish an inevitably looking virtual outfit, such as shape, pose, occlusion, illumination, cloth texture, logo and text. We tried to work with the Adaptive Content Generative and Preserving Network (ACGPN) model to predict the imposed virtual clothes on the user picture. ACGPN model is tested on three types of images depending on different levels of poses such as Easy, Medium and Hard. After inferencing the ACGPN model, the results turned out that the model works accurately on easy pose images, is moderate on medium pose images and is below par on hard pose images.ACGPN can create photo-realistic pictures with substantially higher perceptual quality and more fine details than state-of-the-art approaches.

INDEX TERMS Virtual Try-On, Image Alignment, Fashion Industry, Garment Simulation, Image Segmentation, Semantic Segmentation, Pose Detection

I. INTRODUCTION

In 2012, Converse [16] first used a virtual iPhone try-on for the first time allowing consumers to use their phones' cameras to see how shoes looked on them, share images on social media, and purchase shoes online. Online cloth purchasing has been a regular activity among millions of individuals throughout the world in the last several years. It gives greater choice, and the customer may evaluate other comparable commodities to get the lowest and cheapest one. People increasingly buy the majority of their products from the internet and spend a fortune doing so, particularly in the fashion industry, while examining a wide range of styles and types of clothes is simple with only a few clicks. Along with that, manufacturers and distributors make every effort to improve their customers' internet shopping experiences.

During the Covid-19 pandemic lockdown, most of the businesses went into kind of a crisis mode and not only big names but also small retailers are thinking about how they can survive. Taking our time in shops will be difficult in a post-Covid-19 world as a result, online shopping is ingrained significantly in our daily routine as trade becomes more and more like shopping in person, thanks to the efforts of businesses to add new features and services with the intent of providing their customers with the same support and comfort that they would have during an in-person shopping experience. This goal has been achieved by using computer technology to develop virtual try-on applications that assist the fit of garment products to make consumers know how clothes look on themselves. Therefore, online shopping would give more information and availability of all kinds of products to encourage fashion trailers to invest in the way to explore new sales methods and optimization of technological process of purchasing clothes like virtual fitting system. These solutions draw a new picture of the online shopping experience and bring it to a high level of reality and comfort.

Online shopping offers a variety of services, including a return policy and rewards programs that rank items based on price, reputation, and importance. This interface allows the user to order products to be delivered to the designated location while on the go. Even though all of these advantages are acquired, the internet retail business does not enable customers to try on the item before placing an order. Despite the convenience of online shopping, customers are frequently anxious about how a certain fashion item featured on the Internet will appear on them. As a result, a quick and simple virtual try-on solution is required. Also, since there was a complete lockdown during the Covid-19 pandemic, individuals who used to undertake window shopping switched

to internet shopping, but the objective of their assessing the look of the garment on them before purchasing was not accomplished. Online clothing procurement has a wide variety of business advantages (for instance, time, choice, and price). This strategy helps to avoid concerns like fit issues and environmental and financial return costs that come with traditional e-commerce sales platforms. It also invites users to try on a variety of clothing without physically changing them. Customers will be able to quickly determine whether or not they like the clothing.

Virtual try-on for fashion has recently received a lot of attention. Fitting an in-shop fabric image to a target model image is required for this assignment. Two phases make up an efficient framework for this: (1) warping the try-on fabric to align with the target model's body form and posture, and (2) an image composition module to smoothly merge the warped try-on cloth onto the target model picture. Existing approaches for try-on include many artifacts and distortions.

II. LITERATURE REVIEW

A renowned deal of time and work is taken to create a virtual try-on system. Here, we'd like to discuss a handful of ways that have been taken into consideration due to its huge profit potential mainly in fashion parsing. Our approach is also more difficult than recent interactive search work that focuses on more minute features and merely adjusts the properties (e.g., colour and textures) of a clothing item. We have taken the references of the other four to five papers in comparison to the preliminary review that had been done in our proposal report.

From a commercial standpoint, if a technology existed that captured a full-body photograph and imposed garments on the user's image according to the user's preferences, it would undoubtedly be profitable. If the garments are imposed on the user's body image, they will be more satisfied. They'd try on the clothing to see how it looked on them. This concept would minimize the number of returns and refunds for online clothing shopping applications since customers will be able to digitally try on goods before purchasing them.

A. IMAGE SYNTHESIS

Generative adversarial networks (GANs) [1] aim to imitate the real picture distribution by driving the produced samples to be indistinguishable from the real ones. Image-to-picture translation, whose purpose is to translate an input image from one domain to another, has provoked outstanding results using conditional generative adversarial networks (cGANs). Moreover, when dealing with massive spatial deformations between the conditioned image and the target image, most of these techniques have certain issues. Chen and Kolton [3] developed a CNN without adversarial training using a regression loss as an alternative to GANs for this job. These techniques may have created photo-realistic pictures, but they struggle when geometric changes occur. So, we decided to go with a refinement network for virtual try-on that focuses on garment areas and deals with clothing deformations. Between the mask of in-shop garments and the expected foreground mask, WEI et al. [19] developed a computer architecture,in which Convolution Neural Networks produces image-dependent space models and photographic characteristics. The relationships between the variables for articulated pose estimation are the basis for this work. VITON [1] computes shape context Thin-plate spline (TPS) transformation. Form context is a labour-intensive property of shape, therefore matching two shapes takes time.

B. PERSON IMAGE GENERATION

Lassner et al. [2] developed a generative model which can build human parsing maps and convert them into apparel people. However, it stood unclear how to manage the fashion products that are created. Zhao et al. [4] tackled the challenge of creating multi-view clothing pictures from a single-view clothing image. PG2 synthesizes human photos in any posture, with the goal pose explicitly used as a condition. Siarohit et al. [5] tackled the same problem as PG2, but this time with the help of correspondences between the goal pose and the conditional picture pose. The created fashion items in the conditional photographs were maintained constant. Chen et al. [17] contrasted the network with two discriminators and a multi-position generator is employed to estimate pose. The two discriminators' defining illogical stances are fair poses. This discriminant is used by the multitasking generator as an expert who recognises true and false positions and teaches them to produce a position that fools the expert into believing it is true. This approach gives a more accurate assessment of posture since it may cross, obstruct, and twist human bodies. This method is also suitable for other issues involving shape measurement, such as facial mark detection using DCNNs.

FashionGAN [6] modified a person's clothing and created new ensembles according to text descriptions. The purpose of a virtual try-on is to create a photo-realistic new picture with a fresh item of clothing while removing the previous one's effects. Yoo et al. [7] found that apparel in stores is conditioned on the person wearing it, not the other way around.

We have identified certain research gaps after a thorough literature review such as the limited amount of images in the dataset which aren't sufficient for the training model. Only the outcome of the most common and easy poses is detected. The clothing pixels often leak into the skin pixels, and in the case of self-occlusion, the skin pixels may be completely replaced. A huge mismatch in the current and target clothing shapes.

III. RESEARCH QUESTIONS

Our contribution consists in responding to the following research questions:

RQ1. How can we figure out unique poses?
RQ2. How can we prevent the scattering of the pixels?
RQ3. How to get the same target image as the input image?
RQ4. How to prevent overlapping of the input image on the target cloth?

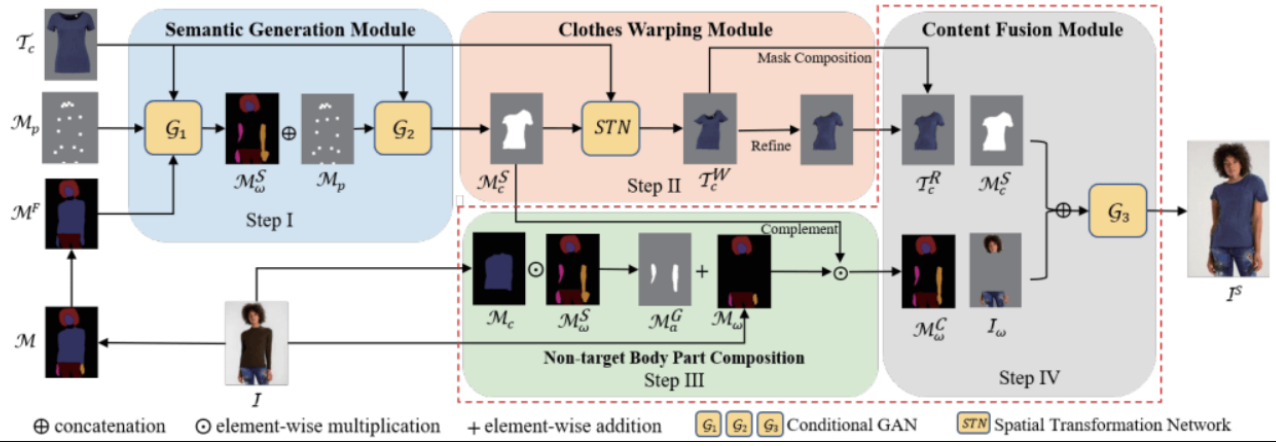

FIGURE 1: Architecture of ACGPN [14]

RQ5. Which dataset should we utilize to get the desired results?
RQ6. Which model gives the best results beneath the existing issues?

IV. DATASET

The dataset that we are utilizing in our project is a combination of the two datasets [18]. We are utilizing the datasets produced by Han et al. [1] which is the VITION dataset that contains around 19,000 front-view ladies and top garment picture pairings are included. There are 16253 cleaned pairings, divided into a training set of 14221 pairs and a validation set of 2032 pairs, respectively. This dataset contains a variety of picture kinds at various levels of difficulty. It includes a simple position with the model's arms straight and towards the ground. Models' arms are in the pocket in a medium position, and their arms are folded or in a similar pose in a hard pose. As the testing set, we rearrange the photos in the validation set into unpaired pairs. We have also extracted images from another dataset, Fashion-MNIST [8] which includes a considerable variety of women images to improve the accuracy results.

To generate the output, we provided a test dataset which consists of different variables such as-
Test_img: Images of User
Test_color: garment image (to be imposed on the user's image)
Test_edge: consists of garment image edges
Test_label: human parsing representation of test_img
Test_pose: body keypoints of test_img
Test_colormask: consists of strokes
Test_mask: consists of a random mask

Test_colormask and test_mask are used to shade the image.

V. METHODOLOGY

One of the important technological challenges of virtual try-on generation is disintegrating the desired garment picture to accommodate a person's position. For that purpose, we provide an apparel human representation, which includes a collection of attributes including pose, body parts, face, and hair as a prior constraint on the synthesis process. Figure 2 shows the fundamental flowchart for our approach.

To impose desired clothes on the user's image, we employed a pre-trained model called Adaptively Generating Preserving Image Content (ACGPN) [14]. The ACGPN predicts the critical aspects that may be altered in the future after a virtual attempt from the user image model. When the user selects a long-sleeved shirt, the user's current sample shirt (long-sleeve shirt -> arm, arm -> jacket) from the standard arm image is added. This determines how the contents of the image will be stored for future analysis so that the model can acquire important information. This preservation can improve the creation of meaningful garment descriptions and virtual photography trials.

The Adaptive Content Generating and Preserving Network consists of three modules as given in it's architecture in figure 1. First, the Semantic Generation Module (SGM) uses semantic segmentation to gradually construct the masks of body parts and distorted garment areas, resulting in semantic alignment of the spatial arrangement. It receives the image of the target clothing and its mask, data on the person's pose, a segmentation map with all the body parts, and clothing items specified. The person's segmentation map is modified by the first generative model (G1) in the Semantic Generation module, which indicates the region on the person's body that should be covered with the target garments. The second generative model (G2) warps the garment mask to match the region it should occupy after receiving this information. Second, the Clothes Warping Module (CWM) is designed to warp the target clothes picture according to the warped clothing mask, using a second-order difference constraint on Thin-Plate Spline (TPS) [13] to create geometrically matching yet character-retentive clothing images. And finally, the Content Fusion Module (CFM) gets the warped clothing image, the modified segmentation map from Semantic Generation Mod-

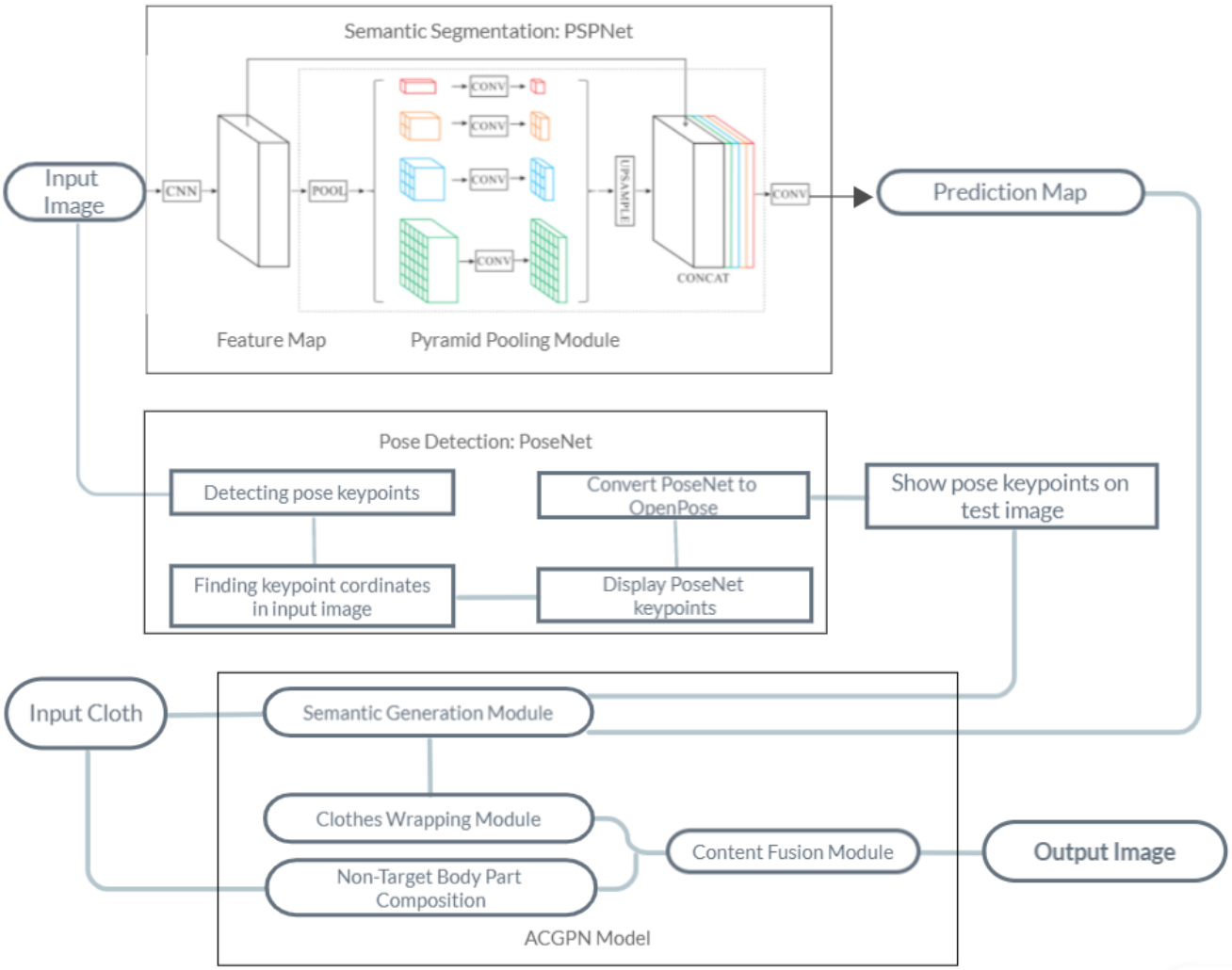

FIGURE 2: FlowChart

ule, and a person's image that are fed into the third generative module (G3). It combines past element information to identify the preservation and growth of certain human body components in a combined picture flexibly and the final result is produced.

The models which were used to construct person segmentation labels and detect key points on a human body were not mentioned by the authors of the original publication [14] that exclusively focused on Adaptively Generating Preserving Images. As a result, we selected the best models ourselves and verified that the ACGPN model's results were comparable to those given in the research.

To get the semantic segmentation of the image as shown in figure 3, we have used PSPNet-MobileNet-v2 which is one of the most well-recognized image segmentation algorithms that utilize a pyramid parsing module that exploits global context information by different-region-based context aggregation.

MobileNetV2 is considerably better than MobilNetV1 and supports state-of-the-art for various mobile visuals identification tasks such as image classification, object detection, and semantic segmentation. This mechanism was provided for the Tensorflow Slim Image Recognition library. MobileNetV2 is based on the principle of MobileNetV1 and employs depthwise separable convolutional as a building element. Especially given the fact that MovbilNetV2 offers two new capabilities. 1) There is a linear bottleneck between layers, and 2) there is a shortcut between the bottlenecks. The supposition is that the bottleneck reflects the algorithm's input and output. The internal layer, on the other hand, reflects the algorithm's ability to shift from lower-level definitions like pixels to higher-level descriptors like image categories. Eventually, shortcuts, like the remaining traditional links, provide for a faster training process and more accuracy. As a result, while estimating local level predictions, the PSPNet architecture analyses the image's global context, resulting in enhanced performance as compared to U-Net and FCN [15]. Furthermore, for pose detection as figure 4, we have used PoseNet. PoseNet is a real-time posture recognition

system that allows us to recognize human poses in images or videos in real-time. It operates in single-mode (single human posture detection) and multi-mode (multi-position detection) modes (Multiple humans pose detection). We even picked the OpenPose model as a body keypoint detector as it provides the correct sequence of keypoints (COCO keypoint dataset) and has previously been utilized in virtual try-on for clothing replacement research.

For implementation, the python library is used for Open pose to extract key points. To bring all the variables together, Google Colab has been used. To install OpenPose into the system, the system needs the proper configuration of CUDA and CudNN 7.5 and windows 10. The system should have a higher level of graphic cards. Since this criterion is not met for our system, Open pose and Nvidia-smi have been installed in Google Colab. A model for COCO data and Pretrained weights is downloaded and uploaded using the library. The open pose gives 14 key points for input images; this key point is generated and stored in the .json file for further implementation. The key points generated are in 2d form, which is converted into regular key points. This key point is stored in the ACGPN model as test_pose.

ACGPM delivers realistic simulated pictures while keeping the basic features of clothing (texture, logo, and fraternity) as well as components of human identification (pose, body parts, and base clothes). Three modules (GMM), (CWM), and (CFM) has been created specifically for this purpose. Depending on the posture, the VITION dataset contains distinct images. We inferred the model and obtained the results after adding test labels and test pose. While inferencing the model, the index number in the aligned dataset file has been modified to reflect the fact that only one cloth is required.

FIGURE 3: Semantic Segmentation

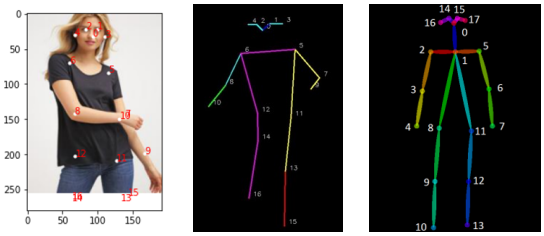

FIGURE 4: Pose Detection

VI. RESULTS

Our project mainly revolves around potential use cases for generative networks in fashion. Broadly speaking, we accomplished the goal of our project by implementing a virtual try-on network — essentially taking in-shop clothing. A person's image is taken as an input to give the output of a person wearing those clothes. We tried overcoming flaws and improved the accuracy by integrating the two different datasets. ACGPN does a far better job of maintaining both the character of the garments and the information about the body parts. We utilized the PoseNet and OpenPose for resolving the pixel distortion issue up to a certain level during the image pre-processing. Also, we were successful in identifying even the most unusual poses. Resolved one of the main limitations where it predicted different target images in comparison to the input image. The model implemented has been done with a focus on tops, with complete apparel transfer being potential future work.

The ramifications of our effort are displayed in the figure 5, which indicates that we were able to get a better favourable outcome. It takes the photo of the model and the image of the t-shirt as inputs and outputs the mapped cloth on that model with the same posing.

FIGURE 5: Result

VII. DISCUSSION

After understanding the data and structure, we discovered that the test label could be constructed using Self-Correction Human Parsing methods, and the ACGPN segmentation labels could be utilised in this model. Similarly, the test pose is the input image body key points, which may be produced using OpenPose. The ACGPN model is then used to test all of these data variables. We made sure that the input pictures were in a different position to evaluate ACGPN's ability to impose garments on a new attitude. We realised that the test label and the test pose may be created using the OpenPose in order to use ACGPN on our own dataset.

The results are saved in the test label folder. We used PoseNet and OpenPose on input photos to produce body key points (test pose), and those key points were saved as.json files in the test pose folder. The photos of the clothing that required to be superimposed on the user's photograph are in the test color folder. The garment edge of the test color is test edge. Shade the image with test colormask and test mask to make it unfinished so that the network can learn to paint it. The supplied picture is Tets img. For testing, all of these test files are loaded into the ACGPN model.

We can see that the model performed well on the Easy pose picture and well on the Medium pose image. Hard posture photographs, on the other hand, did not turn out that well. We aimed to overcome flaws and improve accuracy by integrating the two datasets. ACGPN preserves both the nature of the garments and the information about the body components better. We utilized PoseNet and OpenPose to fix the pixel distortion issue to a certain extent during the image pre-processing. We were also capable of recognizing even the strangest of postures. One of the challenges we overcame was a significant mismatch between the present and goal apparel shapes. One of the primary issues was that the picture of the target and the input were overlapping.

VIII. CONCLUSION AND FUTURE WORK

We used the Adaptive Content Generating and Preserving Network, or ACGPN, in this project, which attempts to provide photo-realistic try-on results while conserving both clothing character and human identifying information (posture, body parts, bottom clothes). We integrated the two datasets to improve the precision of our results. We overcame the issue of the overlapping of cloth pixels on the input image to a certain limit with the help of PoseNet and OpenPose.For the time being, Virtual Try-On worked only for straight poses and moderate poses. We tried to determine the output for the most unusual poses. At all difficulty levels, VITON's graphics exhibit a range of visual irregularities, including colour mixing, boundary-blurring, and cluttered texture.

The ACGPN model has certain restrictions, such as the requirement that the training data include matched photographs of target clothing and persons wearing those items. Considering the facts, a virtual try-on wardrobe may be impossible to implement, however, this is not the scenario. While it is now a difficult undertaking, it also provides a potential window of opportunity for AI-based advancements. And there are already new techniques in the works to address such problems. Another crucial factor to consider when selecting a suitable use case scenario is the technical capabilities. Other efforts can be undertaken to improve the quality of human body parts generated and adapt to the topological changes in clothes. The model may be further translated into Torchscript by utilizing ImageNet. Pytorch Mobile may be used to import that Torchscript into Android or iOS apps. If the deployment goes well, online buyers will be able to try on garments digitally on their smartphones from anywhere around the globe. This will help the enterprise as well as the individual with return and refund hassles.

IX. CONTRIBUTION

By combining the two datasets, we attempted to overcome weaknesses and enhance accuracy. Both the nature of the clothes and the information about the body parts are significantly better preserved by ACGPN. During the image pre-processing, we used PoseNet and OpenPose to resolve the pixel distortion issue up to a specific degree. We were also able to recognise even the most bizarre positions. One of the thing we were able to overcome is a huge mismatch in the current and target clothing shapes. One of the major flaws was that it also overlapped the image of target and the input.

X. ACKNOWLEDGEMENT

We would like to recognize the writers from different associations we have perused papers from and the individuals who have helped us in directing this research. Likewise, explicitly we would like to say thanks to Dr. Garima Bajwa for directing us through her talks and for giving important inputs and feedback that helped us finish this project.

REFERENCES

[1] Han X, Wu Z, Wu Z, Yu R, Davis L. VITON: An Image-Based Virtual Try-on Network. https://arxiv.org/pdf/1711.08447.pdf
[2] Lassner C, Pons-Moll G, Gehler P. A Generative Model of People in Clothing. https://arxiv.org/pdf/1705.04098.pdf
[3] Chen Q, Koltun V. Photographic Image Synthesis with Cascaded Refinement Networks. https://arxiv.org/pdf/1707.09405.pdf
[4] Zhao B, Wu X, Cheng Z-Q, Liu H, Jie Z, Feng J. Multi-View Image Generation from a Single-View. https://arxiv.org/pdf/1704.04886.pdf
[5] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for Pose-based Human Image Generation.". [Online]. Available: https://arxiv.org/pdf/1801.00055.pdf
[6] Zhu S, Fidler S, Urtasun R, Lin D, Loy C. Be Your Own Prada: Fashion Synthesis with Structural Coherence *. https://arxiv.org/pdf/1710.07346.pdf
[7] Yoo D, Kim N, Park S, Paek A. Pixel-Level Domain Transfer. https://arxiv.org/pdf/1603.07442v1.pdf
[8] zalandoresearch, "zalandoresearch/fashion-mnist: A MNIST-like fashion product database. Benchmark," GitHub, https://github.com/zalandoresearch/fashion-mnist
[9] Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M. Toward Characteristic-Preserving Image-Based Virtual Try-on Network. https://arxiv.org/pdf/1807.07688.pdf
[10] S. G. Salve and K. Jondhale, "Shape matching and object recognition using shape contexts," 2010 3rd International Conference on Computer Science and Information Technology, Jul. 2010, doi: 10.1109/iccsit.2010.5565098.
[11] Jetchev N, Research Z. The Conditional Analogy GAN: Swapping Fashion Articles on People Images. https://arxiv.org/pdf/1709.04695.pdf
[12] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual Fitting by Single-Shot Body Shape Estimation," Semantic Scholar, 2014. https://www.semanticscholar.org/paper/Virtual-Fitting-by-Single-Shot-Body-Shape-Sekine-Sugita/09bf3ce51d404d7f014057057726af4d19a78446
[13] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," 2012. https://www.semanticscholar.org/paper/Splines-minimizing-rotation-invariant-semi-norms-in-Duchon/dd573acce51d862cfb576bf9588f7ccad07b7272
[14] Yang H, Zhang R, Guo X, Liu W, Zuo W, Luo P. Towards Photo-Realistic Virtual Try-on by Adaptively Generating Preserving Image https://arxiv.org/pdf/2003.05863.pdf
[15] "How PSPNet works? | ArcGIS Developer," Arcgis.com, 2016. https://developers.arcgis.com/python/guide/how-pspnet-works/
[16] H. Ghodhbani, A. Alimi, M. Neji, and I. Razzak, "You can Try without Visiting: A Comprehensive Survey on Virtually Try-on Outfits," Aug. 2021, doi: 10.36227/techrxiv.13904099.v2.
[17] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A Structure-aware Convolutional Network for Human Pose Estimation." [Online]. Available: https://arxiv.org/pdf/1705.00389.pdf
[18] "RM_Dataset - Google Drive," Google.com, 2013. https://drive.google.com/drive/folders/16oEp7TOF1wjMx0nbA6BWf7d2msY9dVAE
[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines." [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/papers/Wei_Convolutional_Pose_Machines_CVPR_2016_paper.pdf