A Review on Virtual Try on

Tripti Agrawal[1], Shitanshu Goyal[2], Shivam Patil[3]
Poornima College of Engineering,302022, India
2020pcecstripti187@poornima.org,
2020pcecsshitanshu173@poornima.org,
2020pcecsshivam176@poornima.org

Mrs. Barkha Narang
Assistant Professor
Poornima College of Engineering, 302022, India
barkha.narang@poornima.org

***Abstract*—**
Now days the popularity of online shopping is increasing day by day Despite that growth stores have retained their main advantage – customers being able to try out products before buying them. But as shopping increasingly moves online, brands are striving to offer a holistic user experience on their digital channels too. Luckily, the advancing virtual try-on technology makes it possible to try on almost anything in just a few seconds – from makeup, jewellery and glasses to clothes and shoes. Prior arts usually focus on preserving the character of a clothing image (e.g. texture, logo, embroidery) when warping it to arbitrary human pose. However, it remains a big challenge to generate photo-realistic Tryon images when large occlusions and human poses are presented in the reference person. To address this issue, this work was done when Han Yang was a Research Intern at Sense Time Research. Adaptive Content Generating and Preserving Network (ACGPN) a visual try on network which first predicts semantic layout of the reference image which will be changed after try-on (e.g. long sleeve shirt→arm, arm→jacket), and then image content needs to be either generated or preserved will be determined By using the predicted semantic layout, it leads to photo-realistic try-on and rich clothing details. ACGPN generally involves three major modules.

I. Introduction

Nowadays electronic devices like smartphones, computers support Augmented Reality and the popularity keeps growing every day, mostly thanks to social networks. This makes virtual try-on solutions easily accessible and highly desirable to end users. It's a big opportunity for all the companies or brands on favourite platforms to get closer to the customers. To stimulate the visualisation of dressing the demand of developing virtual dressing rooms is increasing.

Therefore, most researchers in previous works are taking the approach to map a 2D texture to the user's body, and build an Avatar (model). In this work, we introduce a virtual dressing room application using Microsoft Kinect sensor. As the user stands in the front of the Kinect, his sizes measuring in real time, image mapping occurs. Further in this paper we have introduced VDRS i.e. Virtual Dressing Room system (VDRS), the main objective is develop an application that realistically reflects the look and feel of the clothes as it is supposed to. The clothes should adapt to certain bodies of different people.

It tough task to build up the photo-realistic virtual try-on system to apply in real world as to overcome such limitations Adaptive Content Generation and Preservation Network (ACGPN) is introduced which first predicts the semantic layout of the reference image and then adaptively determines the content generation or preservation according to the predicted semantic layout. The system incorporates a novel recurrent generation module to produce different looks depending on the order of putting on garments because of that it is named as DiOr, for Dressing in Order. In this research, we offer LA-VITON, an image-based virtual try-on network that outperforms earlier approaches in terms of providing pleasing-looking outputs without damage or distortion. Because GAN-based approaches have limits in terms of creating precise and clear pictures, an operation like GMM must be used to transmit data straight from the source to the target. The try-on photographs are smoothly and convincingly blended with the provided in-store apparel images and target human images.

1.1 Microsoft Kinect Sensor -

The application of virtual dressing room is using Microsoft Kinect sensor's as proposed approach is mainly based on extraction of the user from the video stream, alignment of models and skin colour detection. Modules are being used for locations of the joints for positioning, scaling and rotation in order to align the 2D cloth models with the user. Then, skin colour detection on video to handle the unwanted occlusions of the user and the model was applied. At the end the model is superimposed person who is using in real time. The problem is simply the alignment of the user and the cloth models with accurate position, scale, rotation and ordering. First, detection of the user and the body parts is one of the main steps of the problem. In literature, several approaches are proposed for body part detection, skeletal tracking and posture estimation, and superimposing it onto a virtual environment in the user interface. Kinect driver's middleware are used for the tracking process in combination with Microsoft and for various fundamental functions Kinect. As importance of Microsoft Kinect image sensor in the market is increasing so for that it is being applied and WFP to capture the user physical measurements which are as follows: -

1.1.1 ***Kinect General components*****:** The components of Kinect for Windows are mainly the following:

- Kinect hardware: including the Kinect sensor
- USB hub, through which the sensor is connected to the computer
- Microsoft Kinect drivers: Windows 8 drivers for the Kinect sensor

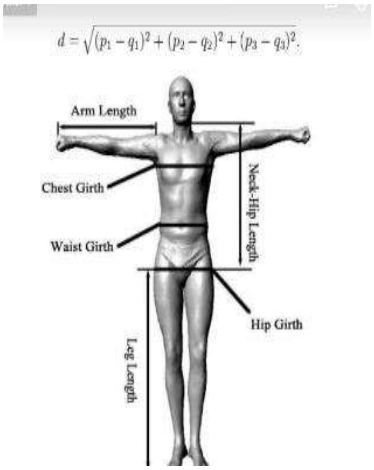
1.1.2 ***WPF Application*****:** WPF stands for Windows Presentation Foundation which is UI framework to create applications with a rich user experience. It is part of the .NET framework 3.0 and higher. Its vector-based rendering engine uses hardware acceleration of modern graphic cards. This Makes the UI faster, scalable and resolution independent. In this section we will introduce the windows presentation foundation and discover its components and the features of each component as well.

In addition, the application keeps calculating distances to continue body movement:

The key points are as follows: -

1. Distances between joints positions

Using a general distance formula in 3D, we formulate an algorithm to all body distances. Each user enters the application in real time.



2. Convert Data:

Units of distance calculated using joint -joint coordinates results in metres so we have to convert it to pixel Data.

Using m – pixel converter

1 m = 3779.527559055 pixels.

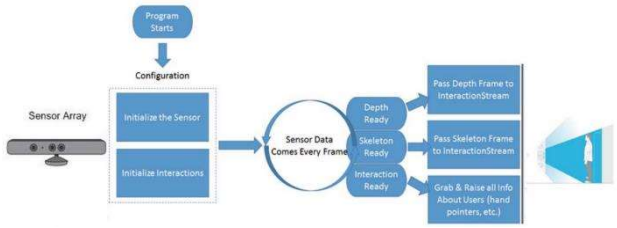
3. Create new joint positions:

Sometimes, key points that the Sensor found it, we face many problems in hip center, hip left, hip right points and it imprecise and inappropriate to what we need, in order to resolve this issue, we decided that we could make new points away from the original points left or right or down.

4. 2D cloth: We divide the clothes into parts of pixels data to control the movement of the body and the cloth in upper, lower frame. Using Photoshop CS6: we cut, divide, coloured the cloth to simulate it to a real body.

5. Interaction between human and 2D cloth:
• First, we take skeleton Data.
• Second, we take Depth Data.
• Third, we take the RGB data
• Fourth, we measure the upper Frame and mapping the
2D cloth.
• Fifth, we measure the Lower Frame and mapping the
2D cloth.
• Sixth, we measure the Lower, upper Frame and mapping the 2D cloth for both.

1.1.3 ***Microsoft Kinect SDK version 1.0***: This version supports up to four Kinect sensors on a single computer, skeletal tracking, a Near Mode feature can make camera able to recognize objects just 40cm away, stability and audio, Drivers and runtime. The following diagram shows the components of the Kinect for Windows developers Toolkit.



1.2 ADAPTIVE CONTENT GENERATING AND PRESERVING NETWORK(ACGPN)

Adaptive Content Generating and Preserving Network which is used to preserve the character of clothes as well as the posture of person as to show the précised image. ACGPN generally involves three major modules.

1. A semantic layout generation
2. Clothes warping module
3. For content fusion module named as inpainting

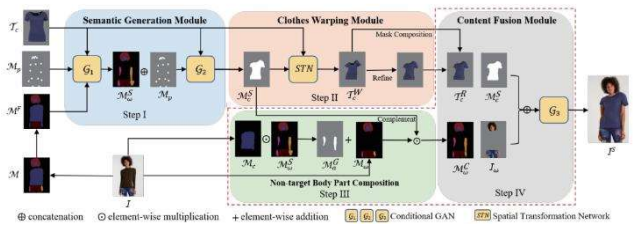


Fig: The overall process of ACGPN involving its three models

1.2.1 ***Semantic Generation Module (SGM)***;- It is basically constructs the mask of various body parts as well as the clothes which are to be wrapped.

1.2.2 ***Clothes Warping Module (CWM)***:- This module is made to warp the targeted clothing image according to the mask generated for wrapping the clothes and author also introduce an updated version or a second-order difference constraint on Thin-Plate Spline (TPS) to produce a geometric matching yet preserving the character of clothing images.

1.2.3 ***Content Fusion Module (CFM)***:- If go further in deep or beyond the semantic alignment and character retention , it remains challenging task to realize layout adaptation on the task. Both the target clothing region is mandatory to get clearly rendered, and fine-scale details of various body parts (for example finger gaps) are needed to be adaptively Target Clothes

To preserve the character of clothes as well as the posture of person as to get the fine details of the clear and précised image adaptive content generating and preserving Network is introduced, i.e., it consists three carefully designed modules, i.e., Mask Generation Module, Clothes Warping Module and Content Fusion Module (CFM). The proposed approach is being evaluate their ACGPN on the VITON dataset by categorising them into three difficulty levels. The results proved that the ACGPN dominates or we can say had a great superiority over the state-of-the-art methods in terms of quantitative metrics, visual quality and user study.

Fig: Categorising the try on task into three difficulty levels

By using ACGPN it greatly changed the quality as it improves a lot in various segments such as semantic alignment, character retention and layout adaptation. The higher order difference increases the stability of the training process of wrapping module and it can handle complex textures on clothes as it increases the ability of the model to handle such complex situations.

ACGPN, the current existing try-on methods mainly concentrate on keeping the posture, identity and character of clothes. There are some methods such as VITON, CP-VTON which by using coarse human shape and pose map as an input to generate a image of a person who is wrapped with clothes in precise manner. Whereas some methods such as SwapGAN, SwapNet and VTNFP adopt semantic segmentation as input to synthesize clothed person.

The results proves that the ACGPN show great advantage or more beneficial over other methods in terms of quantitative metrics, visual quality and user study. ACGPN as compared to the other methods can generate photo-realistic images with much better perceptual quality and richer fine-details.

1.3 DRESSING IN ORDER(DiOr)

Dressing in Order (DiOr) is a flexible person generation framework, which supports 2D pose transfer, virtual try-on, and several fashion editing tasks. The benefit or advantage of DiOr over other methods is that garments can be put on a person in a sequence so that if person try same garments in different order or poses it will result in different looks. DiOr very finely detects the shape and texture of each garment, so that it enables various elements to change separately. Joint training on pose transfer and inpainting helps with detail preservation and coherence of generated garments. On evaluating it extensively it clearly shows that DiOr shows great superiority over other methods which are came recently like ADGAN in terms of output quality, and handles a wide range of editing functions for which there is no direct supervision

DiOr is a flexible person generation pipeline trained on pose transfer and inpainting but capable to layer a wide range or diverse garments and the tasks for which there is no direct supervisions also edited. In general, the shading, garment detail preservation and texture warping of our method, are far better than those of other recent methods and still it is not realistic.

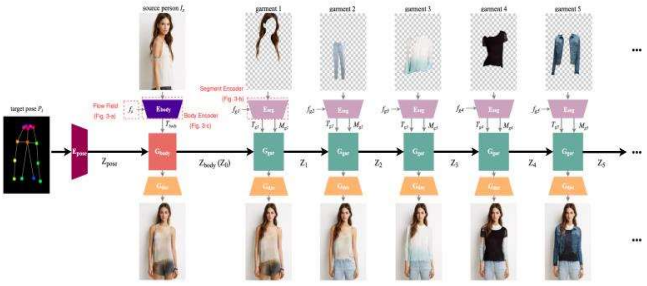


Fig: DiOr Generation pipeline

1.3.1 Generation Pipeline

In the main generation pipeline, firstly the “skeleton” P is to be encoded and after then generating the body from T 0 body and then the garments from texture which is encoded previously and shape masks (T 0 g1 , Mg1), ...,(T 0 gK , MgK) are to be taken in sequence. Pose and skin generation. To start generation, we encode the desired pose P by the help of pose encoder Epose, is to be implemented as three convolutional layers, each followed by instance normalization and leaky ReLU . This results in hidden pose map Zpose ∈ R L×H/4×W/4 , where L denotes the latent channel size and after that we make or generate the hidden body map Zbody given Zpose and the body texture map T 0 body by a body generator . by using two style blocks in ADGAN we can implement Gbody . Because our body texture map T 0 body is in 2D, SPADE. Replace the ADGAN’s adaptive instance normalization in the style block. DiOr was very much inspired by ADGAN. Same as ADGAN, we each garment can be encoded separately, condition the generation on 2D pose, and train on pose transfer. ADGAN encodes a garment into single 1D vector, but in terms of texture and shape garment is encoded separately in 2D. DiOr allows shape and texture of individual garments to be

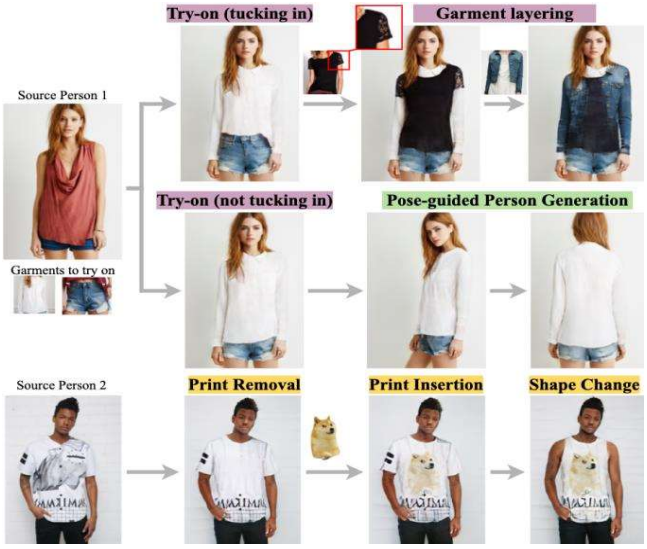


Fig: Applications supported by DiOr system

edited separately where as in ADGAN it is not possible always filled in properly. For capturing complex spatial patterns Our 2D encoding is better than ADGAN's 1D encoding which giving us superior results on virtual try-on, as shown in next section. Besides, In ADGAN, after garments are separately encoded, all the embeddings are fused into a single vector, so the number and type of garments are fixed, and garment order is not preserved. By contrast, in our recurrent pipeline, garments are injected one at a time, and 14641 their number, type, and ordering can vary.

1.4 GEOMETRIC MATCHING MODULE

We use the bottom-up technique suggested by Rocco et al. for implementing the end-to-end trainable geometric matching (GMM). To improve GMM's performance, a matching grid constraint, an occlusion management strategy, and GAN loss are used.

1.4.1 ***Grid Interval Consistency*** - TPS transformations have a good track record, but their high flexibility can lead to pattern and print distortion. TPS with a high degree of freedom causes undesirable warping even when the coarse-to-fine strategy is used. We introduce grid interval consistency (GIC loss), which helps clothing preserve their qualities after warping.

1.4.2 ***Occlusion Handling*** **-** Hair and arms readily obscure a person's attire in a real-life setting. To solve the problem, we remove occlusion zones from the Lwarp computation. This technique helps the network to be taught to estimate transforming parameters with greater precision and reasonableness.

1.4.3 *Try on Module*

For the input person-image TOM synthesizes the final try-on image. To preserve the original details of in-shop clothing items, warped clothing should be exploited as much as possible. By blending the warped clothing from GMM and the intermediate person with a composition mask a seamless image is obtained. By following the generator, refinement layers improve the quality of the blended image higher. Dilated convolutions are used in the refinement layers to maintain the high-resolution feature map and to preserve picture details. for increasing the quality of pictures created using try on module SNGAN is used for training.

1.4.4 *Experiments*

Han et al. obtained a paired dataset, which we use. There are 14,221 training and 2,032 testing pairings in this dataset. The suggested GMM's alignment results are compared to those of the prior alignment methods. The suggested solution preserves pattern and print properties while also precisely aligning in-store clothing. For the first time, LA-VITON was compared to VITON and CP-VTON to evaluate if in-store garments had the same patterns and designs as those purchased in stores.

II. DISCUSSION

The results prove the various advantages of ACGPN, DiOr, Kinect sensor, VDRS over the state-of-the-art methods in terms of quantitative metrics, visual quality and user study. In comparison to the methods, ACGPN can generate photo-realistic images with clear and much better quality and richer fine-details. DiOr is a flexible person generation pipeline trained on pose transfer and inpainting but capable of diverse garment layering and editing tasks for which there is no

direct supervision those of other recent methods, are still not entirely realistic. In the future, we plan to work on increasing the quality of the image produces by wrapping the clothes over the person through more advanced warping and higher-resolution training and generation.

III. CONCLUSION

To preserve the character of clothes as well as the posture of person as to show the précised image Author propose a novel adaptive content generating and preserving Network, i.e. Author present three carefully designed modules, i.e. Mask Generation Module (GMM), Clothes Warping Module (CWM), and Content Fusion Module (CFM). ACGPN is evaluated on the VITON dataset by categorising them into three levels of try-on difficulty. The results proves that ACGPN has great superiority over the state-of-the-art methods in terms of quantitative metrics, visual quality and user study. We can expect to see many more innovative uses for both technology in the future and perhaps a fundamental way in which we communicate and work thanks to the possibilities

IV. REFRENCES

[1] Remi Brouet, Alla Sheffer, Laurence Boissieux, and Marie-´Paule Cani. Design preserving garment transfer. ACM Trans. Graph., 31(4):36:1–36:11, 2012.

[2] Szu-Ying Chen, Kin-Wa Tsoi, and Yung-Yu Chuang. Deep virtual try-on with clothes transforms. In ICS, volume 1013 of Communications in Computer and Information Science, pages 207–214. Springer, 2018.

[3] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In CVPR, pages 8789–8797. IEEE Computer So-ciety, 2018.

[4] Haoye Dong, Xiaodan Liang, Bochao Wang, Hanjiang Lai, Jia Zhu, and Jian Yin. Towards multi-pose guided virtual try-on network. CoRR, abs/1902.11026, 2019.

[5] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Zhenyu Xie, Bowen Wu, Ziqi Zhang, Xiaohui Shen, and Jian Yin. Fashion editing with multi-scale attention normaliza-tion. CoRR, abs/1906.00884, 2019.

[6] V. Scholz and M. Magnor, “Multi-view video capture of garment motion,” in Proc. IEEE Workshop on ContentGeneration and Coding for 3D-Television, 2006, pp. 1–4.

[7] A. Del Bue and A. Bartoli, “Multiview 3d warps,” inComputer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 675–682.

[8] J. Shen, X. Yan, L. Chen, H. Sun, and X. Li, “Re-texturing by intrinsic video,” Information Sciences, vol. 281, pp. 726–735, 2014.

[9] A. Hilsmann and P. Eisert, “Optical flow based tracking and retexturing of garments,” in Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. IEEE, 2008,pp. 845–848.

[10] http://glamstorm.com/en/fittingroom/clothes# .

[11] http://www.metail.com.