IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12102

A Survey on virtual try-on clothing system

Girish Mantha[1], Praveen G Shet[2], Rahul Athreya K M[3], Sathwik P Bhat[4], Shamanth G.P[5]

Assistant Professor, Department of Information Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[1]

Student, Department of Information Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[2-5]

Abstract: The Virtual cloth Try-on is one of the biggest inventions took place in fashion industry which contributes to enhance user experience by allowing them to try out garments virtually without wearing it. Prior arts usually focus on preserving the character of a clothing image (e.g., texture, logo, embroidery) when warping it to arbitrary human pose. This project suggests a better solution using Artificial Intelligence (AI) and Augmented Reality (AR) for non-tech-savvy customers who aims at transferring a target clothing image onto a himself. It remains a big challenge to generate photo-realistic try-on images when large occlusions and human poses are presented in the reference person then determines whether its image content needs to be generated or preserved according to the predicted semantic layout, leading to photo-realistic try-on and rich clothing details.

Keywords: Augmented Reality (AR), Artificial Reality (AI), Virtual Try-On (VTON), Photo-Realistic Image

I. INTRODUCTION

Image virtual try-on (VTON) aims at transferring a target clothing image onto a reference person and has become a hot topic in recent years. 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 try on images when large occlusions and human poses are presented in the reference person.it generally involves three major modules. First, a semantic layout generation module utilizes semantic segmentation of the reference image to progressively predict the desired semantic layout after try-on. Second, a cloth warping module warps clothing images according to the generated semantic layout, where a second-order difference constraint is introduced to stabilize the warping process during training.

Third, an inpainting module for content fusion integrates all information (e.g, reference image, semantic layout, warped clothes) to adaptively produce each semantic part of human body. In comparison to the state-of-the-art methods, this method can generate photo-realistic images with much better perceptual quality and richer fine-details. Motivated by the rapid development of image synthesis [6], image-based visual try-on [9] aiming to transfer the target clothing item onto reference person has achieved much attention in recent years. Although considerable progress has been made [2], it remains a challenging task to build up the photo-realistic virtual try-on system for real-world scenario, partially ascribing to the semantic and geometric difference between the target clothes and reference images, as well as the interaction occlusions between the torso and limbs.

To illustrate the limitations of existing visual try-on methods, we divide the VITON dataset [9] into three subsets of difficulty levels according to the human pose in 2D reference images. As shown in Fig. 1, the first row gives an easy sample from VITON dataset, where the person in the image is represented with a standard posture, i.e., face forward and hands down. In such case, the methods only need to align the semantic regions between the reference and target images. Some pioneering synthesized based methods [6] belong to this category. From the second row, the image with medium-level difficulty is generally with torso posture changes. And several models [2] have been suggested to preserve the character of the clothes such as texture, logo, embroidery and so on. Such goal is usually attained by developing advanced warping algorithms to match the reference image with clothes deformation. The last row of Fig. 1 presents a hard example, where postural changes occur on both the torso and the limbs, leading to the spatial interactions between the clothing regions and human body parts, e.g., occlusions, disturbances, and deformation. Therefore, the algorithm is required to understand the spatial layout of the foreground and background objects in the reference image, and adaptively preserve such occlusion relationship in the try-on process.

DOI: 10.17148/IJARCCE.2023.12102

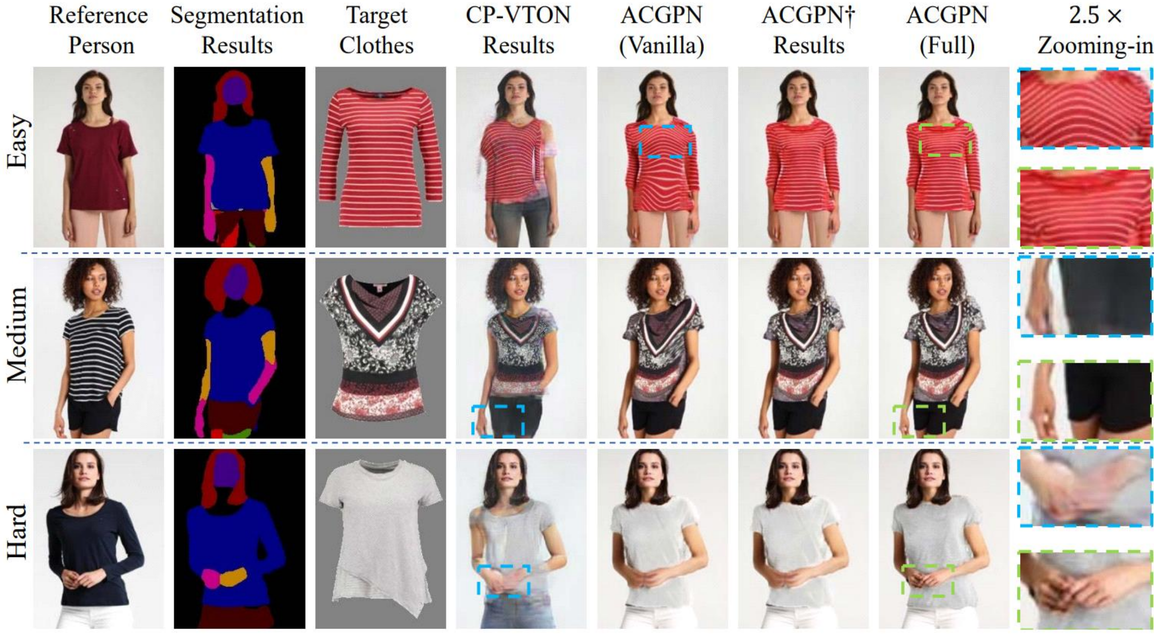

Fig.1 Cloths wrapping images

However, content generation and preservation remain an un-investigated issue in virtual try-on. To address the above limitations, this report presents Adaptive Content Generation and Preservation Network (ACGPN), 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.

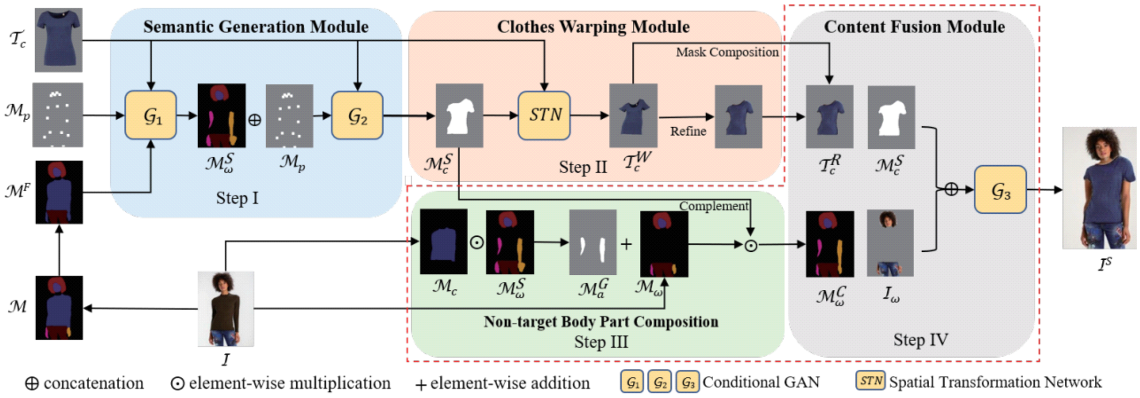

Fig.2 Modules

This Method consists of three major modules as shown in Fig. 2. The first one is the Semantic Generation Module (SGM), which uses the semantic segmentation of body parts and clothes to progressively generate the mask of exposed body parts (i.e., synthesized body part mask) and the mask of warped clothing regions. As opposed to prior arts, the proposed SGM generate semantic masks in a two-stage fashion to generate the body parts first and synthesize clothing mask progressively, which makes the original clothes shape in reference image completely agnostic to the network.

The second part is the Clothes Warping Module (CWM), which is designed to warp clothes according to the generated semantic layout. Going beyond the Thin-Plate Spline based methods a second-order difference constraint is also introduced to the Warping loss to make the warping process more stable, especially for the clothes with the complex texture. Finally, the Content Fusion Module (CFM) integrates the information from the synthesized body part mask, the warped clothing image, and the original body part image to adaptively determine the generation or preservation of the distinct human parts in the synthesized image. With the above modules, It adopts a split transform-merge strategy to generate a spatial configuration aware try-on image. Experiments on the VITON datasets show that our model not only

DOI: 10.17148/IJARCCE.2023.12102

promotes the visual quality of generated images for the easy and medium difficulty levels (Fig 2), but also is effective in handling the hard try-on case with the semantic region intersections in an elegant way and produces photo-realistic results.

II. VIRTUAL TRY-ON APPROACHES

A literature survey is an overview of the previously published works on a specific topic. The term can refer to a full scholarly paper or a section of a scholarly work such as a book, or an article. Either way, a literature survey is supposed to provide the researcher/author and the audiences with a general image of the existing knowledge on the topic under question. A literature survey serves to situate the current study within the body of the relevant literature and to provide context for the reader. The following are the different works carried out in the given area. Each paper is discussed with methods, advantages, drawbacks and brief methodology.

2.1 "Deep Virtual Try-on with Clothes Transform"

They developed a system that uses images for virtual try on, which allows trying on clothes without limiting the view direction of people and target clothes. Their proposed method contains four steps. Firstly, given a person image and a target clothes image, CAGAN[1] is used to generate a preliminary result and a binary mask of where to change. Secondly, a transform network is used to extract the clothes only. Meanwhile, the mask and output from previous step are used in segmentation step for a better mask to indicate where the clothes should be transformed to. Next, the mask is used to transform the target clothes. Lastly, the transformed clothes and the output from CAGAN are combined which becomes our final output. During training image from the generator is used to fool the discriminator while discriminator try to discriminate it. Which can result in an improvement of the generated image. It requires accurate poses and segmentations, which need manually fine tuning and are difficult to be obtained. Sometimes the output can be blurry and preserves less details than the target clothes.

2.2 "Towards Multi-pose Guided Virtual Try-on Network"

They proposed a multi stage method to synthesize the image of person conditioned on both clothes and pose. MG-VTON is constructed as a desired human parsing map of the target image is synthesized to match both the desired pose and the desired clothes shape, a deep Warping Generative Adversarial Network (Warp-GAN)[2] warps the desired clothes appearance into the synthesized human parsing map and alleviates the misalignment problem between the input hu man pose and desired human pose, a refinement render utilizing multi-pose composition masks recovers the texture details of clothes and removes some artifacts. It adopts clothes and pose guided network to generate the target human parsing, which is helpful to alleviate the problem that lower-body clothing and hair cannot be preserved. This process involves alleviation of miss alignments. The quality of human parsing significantly affects the quality of the synthesized image in the virtual try-on task.

2.3 "DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images"

Their work presents DeepFashion2, a large-scale benchmark with comprehensive tasks and annotations of fashion image understanding. DeepFashion2 contains 491K images of 13 popular clothing categories. A full spectrum of tasks are defined on them including clothes detection and recognition, landmark and pose estimation, segmentation, as well as verification and retrieval. A full spectrum of tasks is carefully defined on the proposed dataset. With DeepFashion2, They extensively evaluate Mask R-CNN[3] that is a recent advanced framework for visual perception. This establishes benchmarks covering multiple tasks in fashion understanding, including clothes detection, landmark and pose estimation, clothes segmentation, consumer-to-shop verification and retrieval. It is also interesting to explore multi-domain learning for clothing images, because fashion trends of clothes may change frequently. making variations of clothing images changed.

2.4 "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to Image Translation"

The task of image-to-image translation is to change a particular aspect of a given image to another. They proposed StarGAN[4], a novel generative adversarial network that learns the mappings among multiple do mains using only a single generator and a discriminator, training effectively from images of all domains. They demonstrate how one can successfully learn multi domain image translation between multiple datasets by utilizing a mask vector method that enables StarGAN to control all available domain labels. They provide both qualitative and quantitative results on facial attribute transfer and facial expression synthesis tasks using StarGAN, showing its superiority over baseline models. This model is the scalability in terms of the number of parameters required. StarGAN generated images of higher visual quality compared to existing methods. When the wrong mask vector was used, it fails to synthesize facial expressions.

2.5 "Poly-GAN: Multi-Conditioned GAN for Fashion Synthesis"

Poly-GAN[5] allows conditioning on multiple inputs and is suitable for many tasks, including image alignment, image stitching and inpainting. They proposed a new conditional GAN architecture, which can operate on multiple conditions that manipulate the generated image. They demonstrate that their architecture can perform many tasks, including shape manipulation conditioned on human pose for affine transformations, image stitching of a garment on the model, and image inpainting. Their method is able to preserve the desired pose of human arms and hands without color spill, even in cases of self occlusion, while performing Fashion Synthesis. They proposed a new conditional GAN architecture, which can operate on multiple conditions that manipulate the generated image. This method suffers from slight color shift in some samples, which is a common problem with GANs. Limitation with this method is texture preservation when generating letters, graphics or patterns that are present in the reference garment.

2.6 "SwapNet: Image Based Garment Transfer"

They presented SwapNet[6], a framework to transfer garments across images of people with arbitrary body pose, shape, and clothing. Garment transfer is a challenging task that requires disentangling the features of the clothing from the body pose and shape and realistic synthesis of the garment texture on the new body. They presented the first method that operates in image-space to transfer garment information across images with arbitrary clothing, body poses, and shapes. With the absence of ideal training data for supervision, they introduced a weakly supervised learning approach to accomplish this task. They used weakly supervised training procedure to train the warping and texturization modules in the absence of supervised data for same clothing in different poses. It has difficulty handling large pose changes between source and target images. If one of the images contain a truncated body and the other contains a full body, their model is not able to hallucinate appropriate details for the missing lower limbs.

2.7 "Image-to-Image Translation Using Generative Adversarial Network (GAN)"

They used Conditional GAN[7] for image-to-image translation which generated more realistic and high-quality data. Then we analyzed the performance of this framework by doing hyper-parameter tuning and have a comparison of the loss for both Generator and Discriminator. The input source image which is translated to a target image by applying a condition on input image. In image to image translation, there is a supposition that an image is combination of two types of attributes, domain independent and domain specific. The domain independent attributes are those which will not be changed during translation and domain specific are those which will be changed. It is used for image-to-image translation which generated more realistic and high quality data. It is used for video-to-video translation which generated more realistic and high-quality data. This model needs to be paired with training datasets of input and output images. Another way for image-to-image translation is without having paired training datasets i.e., unpaired domain translation.

2.8 "FiNet: Compatible and Diverse Fashion Image Inpainting"

They proposed to explicitly model visual compatibility through fashion image inpainting. They present Fashion Inpainting Networks (FiNet)[8], a two-stage image-to-image generation framework that is able to perform compatible and diverse inpainting. Disentangling the generation of shape and appearance to ensure photorealistic results, their framework consists of a shape generation network and an appearance generation network. FiNet suggests that it can be potentially used for compatibility-aware fashion design and new fashion item recommendation. By decomposition of shape and appearance generation, FiNet can in paint garments in a target region with diverse shapes and appearances. There is a trade-off between compatibility and diversity. When the diversity of a method increases, it often has a higher probability of generating fewer compatible results. FiNet achieves the highest human fooling rate by generating photorealistic images. FiNet w/o two-stage cannot properly determine the clothing boundaries.

2.9 "VITON: An Image-based Virtual Try-on Network"

They presented a virtual try-on network (VITON)[9], which can transfer a clothing item in a product image to a person relying only on RGB images. A coarse sample is first generated with a multi-task encoder-decoder conditioned on a detailed clothing-agnostic person representation. The coarse results are further enhanced with a refinement net work that learns the optimal composition. They conducted experiments on a newly collected dataset, and promising results are achieved both quantitatively and qualitatively. This indicates that our 2D image-based synthesis pipeline can be used as an alternative to expensive 3D based methods. Failure cases are due to rarely-seen poses or a huge mismatch in the current and target clothing shapes. Here for a person with a complicated pose, using body shape information alone is not

DOI: 10.17148/IJARCCE.2023.12102

sufficient to handle occlusion and pose ambiguity. Body shape information is also critical to adjust the target item to the right size.

2.10 "Generative Image Inpainting with Contextual Attention"

They proposed a coarse-to-fine generative image inpainting[10] framework and they showed that the contextual attention module significantly improves image inpainting results by learning feature representations for explicitly matching and attending to relevant background patches. The proposed inpainting framework and contextual attention module can also be applied on conditional image generation, image editing and computational photography tasks including image-based rendering, image super-resolution, guided editing and many others. They introduced several techniques including inpainting network enhancements, global and local WGANs and spatially discounted reconstruction loss to improve the training stability and speed based on the current the state-of-the-art generative image inpainting network. The reconstruction loss is helpful in capturing content structures and serves as a powerful regularization term for training GANs. The method is not suitable for very high-resolution inpainting applications.

III. Advanced Methodology

Multistage method can be used to synthesize the image of person conditioned on both clothes and pose. Given an image of a person, a desired cloth, and a desired pose, we generate the realistic image that preserves the appearance of both desired clothes and person

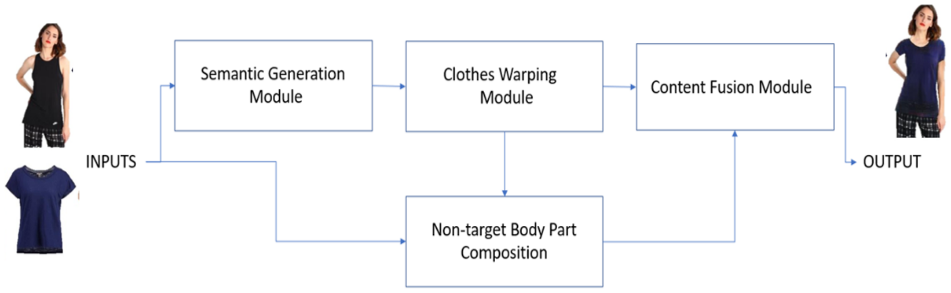

Fig 3 Multistage Synthesizing of Images

Semantic Generation Module (SGM): The semantic generation module (SGM) is proposed to separate the target clothing region as well as to preserve the body parts of the person Clothes Warping Module (CWM): Clothes warping aims to fit the clothes into the shape of target clothing region with visually natural deformation according to human pose as well as to retain the character of the clothes Non-target Body Part Composition: It precisely preserves the non-target body part by combining the two masks from above two modules.

Content fusion module (CFM): This is composed of two main steps, in particular, Step 1 is designed to fully maintain the untargeted body parts as well as adaptively preserve the changeable body part (i.e. arms). Step 2 fills in the changeable body part by utilizing the masks and images generated from previous steps accordingly by an inpainting based fusion GAN.

Semantic generation module (SGM): The semantic generation module (SGM) is proposed to separate the target clothing region as well as to preserve the body parts (i.e., arms) of the person, without changing the pose and the rest human body details. Many previous works focus on the target clothes but overlook human body generation by only feeding the coarse body shape directly into the network, leading to the loss of the body part details. To address this issue, the mask generation mechanism is adopted in this module to generate semantic segmentation of body parts and target clothing region precisely. Specifically, given a reference image I, and its corresponding mask M, arms Ma and torso Mt are first fused into an indistinguishable area, resulting in the fused map MF shown in as one of the inputs to SGM. Following a two-stage strategy, the try-on mask generation module first synthesize the masks of the body parts MS ω (ω = {h, a, b} (h:head, a:arms, b:bottom clothes)), which helps to adaptively preserve body parts instead of coarse feature in the subsequent steps. we train a body parsing GAN G1 to generate MS ω by leveraging the information from the fused map MF, the pose map Mp, and the target clothing image Tc. Using the generated information of body parts, its corresponding pose map and target clothing image, it is tractable to get the estimated clothing region. In the second stage, MS ω, Mp and Tc are combined to generate the synthesized mask of the clothes MS c by G2. For training SGM, both stages adopt the conditional generative adversarial network (cGAN), in which a U-Net structure is used as the generator

DOI: 10.17148/IJARCCE.2023.12102

while a discriminator given in pix2pixHD is deployed to distinguish generated masks from their ground-truth masks. For each of the stages, the CGAN loss can be formulated as

$$L1 = Ex,y\ [\log (D (x, y))] + Ex,z\ [\log (1 - D (x, G (x, z)))] \qquad \text{Eq 1}$$

where x indicates the input and y is the ground-truth mask. z is the noise which is an additional channel of input sampled from standard normal distribution. The overall objective function for each stage of the proposed try-on mask generation module is formulated as Lm,

$$Lm = \lambda 1L1 + \lambda 2L2 \qquad \text{Eq 2}$$

where L2 is the pixel-wise cross entropy loss, which improves the quality of synthesized masks from generator with more accurate semantic segmentation results. λ1 and λ2 are the trade-off parameters for each loss term in, which are set to 1 and 10, respectively in our experiments. The two-stage SGM can serve as a core component for accurate understanding of body-parts and clothes layouts in visual try-on and guiding the adaptive preserving of image content by composition. We also believe SGM is effective for other tasks that need to partition semantic layout.

Clothes Warping Module (CWM):

Clothes warping aims to fit the clothes into the shape of target clothing region with visually natural deformation according to human pose as well as to retain the character of the clothes. However, simply training a Spatial Transformation Network (STN) and applying Thin-Plate Spline (TPS) [6] cannot ensure the precise transformation especially when dealing with hard cases (i.e. the clothes with complex texture and rich colors), leading to misalignment and blurry results.

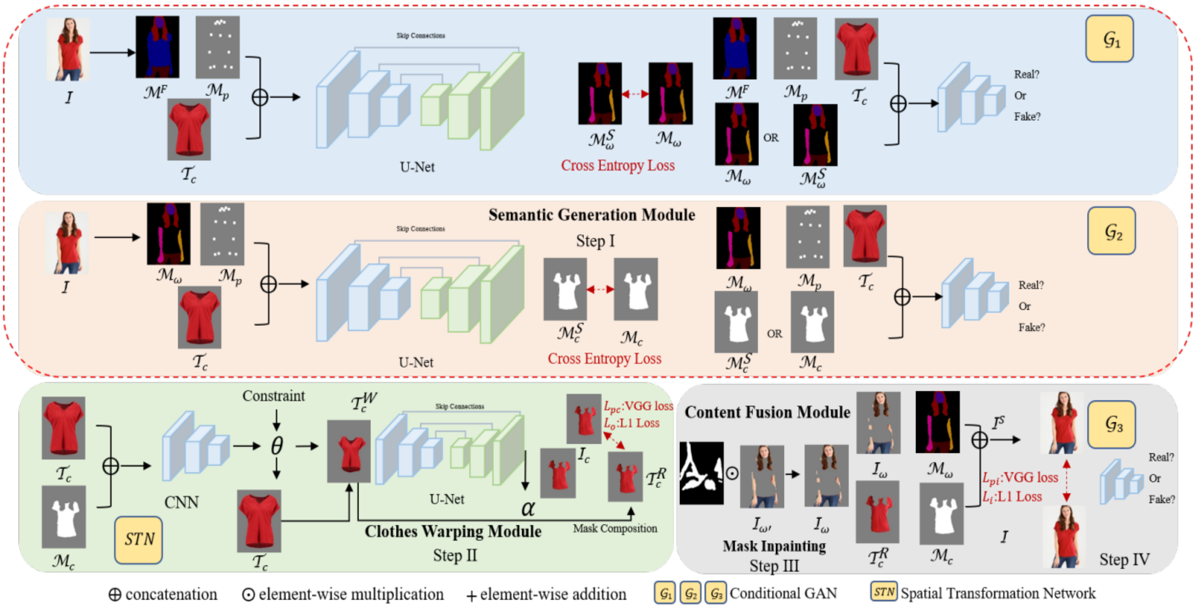

Fig 3.2 Architecture

To address these problems, we introduce a second-order difference constraint on the clothes warping network to realize geometric matching and character retention. As shown in , compared to the result with our proposed constraint, target clothes transformation without the constraint shows obvious distortion on shape and unreasonable mess on texture. Formally, given Tc and MS c as the input, we train the STN to learn the mapping between them. The warped clothing image T W c is transformed by the learned parameters from STN, where we introduce the following constraint L3

$$L3 = X \sum p\epsilon p\ \lambda r\ ||pp0||2\text{-}||pp1||2|+||pp2||2\text{-}||pp3||2| + \lambda s\ (|S (p, p0)\text{-}S (p, p1)|+|S (p, p2)\text{-}S (p, p3)|) \qquad \text{Eq 3}$$

where λr and λs are the trade-off hyper-parameters. Practically we can minimize max (L3 − Δ, 0) for restriction, and Δ is a hyper-parameter. As illustrated in Fig. 3, p(x, y) represents a certain sampled control point and p0(x0, y0), p1(x1, y1), p2(x2, y2), p3(x3, y3) are the top, bottom, left, right sampled control points of p(x, y), respectively in the whole control points set P; S (p, pi) = yi−y xi−x (i = 0, 1, 2, 3) is the slope between two points. L3 is proposed to serve as a constraint on TPS transformation by minimizing the metric distance of two neighbouring intervals for each axis and the distance of slopes, which maintains the collinearity

DOI: 10.17148/IJARCCE.2023.12102

Non-target Body Part Composition:

The composited body mask is composed by original body part mask, the generated body mask which is the region for generation, and synthesized clothing mask. It precisely preserves the non-target body part by combining the two masks. which are used to fully recover the nontargeted details in the following step to fully preserve and generate coherent body parts with the guidance. It is also worth noting that it can adaptively deal with different cases. For example, when transferring a T-shirt (short sleeve) to a person in long-sleeve only the within region will perform generation and preserve all the others

Content Fusion Module (CFM):

This is composed of two main steps, In particular, Step 1 is designed to fully maintain the untargeted body parts as well as adaptively preserve the changeable body part (i.e. arms). Step 2 fills in the changeable body part by utilizing the masks and images generated from previous steps accordingly by an inpainting based fusion GAN. making it possible to separate the regions of preservation and generation. To combine the semantic information, composited body mask MC ω and synthesized clothing mask MS care concatenated with the body part image Iω and refined clothing image T R c as the input. Thus, the texture information can be recovered by the proposed inpainting based fusion GAN, yielding the photo-realistic results. Therefore, in the inference stage, the network can adaptively generate the photo-realistic try-on image with rich details via the proposed CFM. Extensive experiments have proved that the proposed method can not only solve cases of easy and medium levels but hard cases with significant improvement.

IV. Conclusion

There exist many problems in the field of clothing industry such as mismatch of specific looks and fitting in clothing, lack of growth in E-commerce. The aim of this research is to propose a way to construct a software that will collect the images of users and targeted cloths and process it with the help of artificial intelligence and augmented reality by using semantic generation to separate the target clothing region as well as to preserve the body parts of the person without changing the pose and cloths wrapping to fit the clothes into the shape of target clothing region then fusing the above-mentioned methods. This application will prove to be portable and easy to use. It will be much helpful for the people who isn't tech-savvy.

.

REFERENCES

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

[2]. 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.

[3]. Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang and Ping Luo. "Deepfashion2: A versatile benchmark for re-detection, pose estimation, segmentation and re-identification of clothing images". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5337–5345, 2019.

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

[5]. Pandey, N., & Savakis, A. (2020). "Poly-GAN: Multi-Conditioned GAN for Fashion Synthesis". Neurocomputing Elsevier,volume 414, 13 November 2020, Pages 356-364.

[6]. Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. "Swapnet: Image based garment transfer". In ECCV (12), volume 11216 of Lecture Notes in Computer Science, pages 679–695. Springer, 2018

[7]. Kusam Lata, Mayank Dave, Nishanth K N. "Image-to-Image Translation Using Generative Adversarial Network", International Conference on Electronics Communication and Aerospace Technology [ICECA 2019], IEEE Conference Record # 45616; IEEE Xplore ISBN: 978-1-7281-0167-5

[8]. Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott and Larry S Davis. "Finet: Compatible and diverse fashion image inpainting". In Proceedings of the IEEE International Conference on Computer Vision, pages 4481–4491, 2019.

[9]. Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S.Davis. "VITON: an image-based virtual try-on network". In CVPR, pages 7543–7552. IEEE Computer Society, 2018.

[10]. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. "Generative image inpainting with contextual attention". In CVPR, pages 5505–5514. IEEE Computer Society, 2018.