Signal Processing: Image Communication 129 (2024) 117189

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Image-based virtual try-on: Fidelity and simplification

Tasin Islam *, Alina Miron, Xiaohui Liu, Yongmin Li

Brunel University London, Kingston Lane, Uxbridge, London, UB8 3PH, United Kingdom

ARTICLE INFO

Keywords:
Virtual try-on (VTON)
Generative Adversarial Network (GAN)
Fashion synthesis
Occlusion-handling
E-commerce

ABSTRACT

We introduce a novel image-based virtual try-on model designed to replace a candidate's garment with a desired target item. The proposed model comprises three modules: segmentation, garment warping, and candidate-clothing fusion. Previous methods have shown limitations in cases involving significant differences between the original and target clothing, as well as substantial overlapping of body parts. Our model addresses these limitations by employing two key strategies. Firstly, it utilises a candidate representation based on an RGB skeleton image to enhance spatial relationships among body parts, resulting in robust segmentation and improved occlusion handling. Secondly, truncated U-Net is employed in both the segmentation and warping modules, enhancing segmentation performance and accelerating the try-on process. The warping module leverages an efficient affine transform for ease of training. Comparative evaluations against state-of-the-art models demonstrate the competitive performance of our proposed model across various scenarios, particularly excelling in handling occlusion cases and significant differences in clothing cases. This research presents a promising solution for image-based virtual try-on, advancing the field by overcoming key limitations and achieving superior performance.

1. Introduction

The increasing popularity of e-commerce in the fashion industry has created significant prospects for virtual try-on systems to enhance consumers' shopping experiences. Virtual try-on technology generates a portrayal of an individual, showcasing the desired clothing item by employing a deep learning model to fuse images of the candidate and the selected apparel product. In order to provide a genuinely immersive experience, the virtual try-on system must uphold the integrity of the candidate's posture, physique, and distinctive features while simultaneously ensuring a seamless and natural adaptation of the garment to conform to the candidate's body shape.

The initial iterations of virtual try-on models follow a two-stage approach in generating the try-on image [1,2]. Subsequent researchers have enhanced the fidelity of these models by incorporating a segmentation module in their virtual try-on, enabling the preservation of non-targeted body parts, such as the head and arms [3,4] and improving the clothing warping process [3–6]. In recent developments, researchers have introduced innovative normalisation layers to augment the quality of image synthesis and enhance the efficacy of capturing input data [6,7].

A limited number of researchers have examined how virtual try-on models perform in scenarios where the sleeve of the target garment differs from the candidate's original clothing. Our investigations have revealed that certain previously acclaimed state-of-the-art models exhibit unsatisfactory performance when confronted with such scenarios. The implications of this observation highlight the need for further investigation and discussion within the research community to address this particular challenge and advance the capabilities of virtual try-on models in accommodating variations in sleeve design. Moreover, previous methodologies face considerable difficulty when dealing with occlusion in poses [1–3]. Instances where the candidate's arms are crossed in front of their torso, pose a significant challenge for virtual try-on models, as the accurate differentiation of these occluded body parts becomes exceedingly complex. Consequently, such occlusion scenarios can lead to the generation of unrealistic and visually distorted try-on images.

The correctness of the segmentation module's output is paramount in ensuring the fidelity of the virtual try-on model's generated try-on images. Through our analysis, we have identified a critical limitation in previous works' segmentation modules, specifically in cases where the length of the target sleeve differs from that of the original garment. Notably, the examples in Fig. 9 show that the segmentation module encounters challenges in accurately rendering the arm label for long-sleeved garments, consequently resulting in erroneous virtual try-on outcomes. In regard to the occlusion problem, these models [1–3] rely on an 18-keypoint pose map that lacks spatial information about the

* Corresponding author.
E-mail address: tasin.islam2@brunel.ac.uk (T. Islam).

https://doi.org/10.1016/j.image.2024.117189
Received 14 August 2023; Received in revised form 9 July 2024; Accepted 27 July 2024
Available online 16 August 2024

interconnections among key joints of the human body. Consequently, in scenarios involving occlusion, the models face considerable difficulty in accurately discerning and separating these interconnected joints. The absence of spatial relationships within the pose map leads to poor image synthesis.

This paper introduces a novel approach to virtual try-on, known as Simplified Virtual Try-On (SVTON), aiming to overcome the challenges discussed earlier. SVTON incorporates an innovative segmentation module capable of generating highly accurate clothing segments compared to previous methods. Additionally, SVTON effectively addresses the issue of occlusion by leveraging the RGB skeleton pose image as the input, diverging from the conventional practice of utilising an 18-keypoint pose map as employed by previous approaches. Notably, SVTON demonstrates improved efficiency due to containing fewer convolutional layers in the modules, resulting in the accelerated synthesis of try-on images.

SVTON encompasses three fundamental modules to enable its functionality. The first module, the Predictive Human Parsing Module (PHPM), facilitates the generation of segments for the torso and arms. The segments outline these body regions' boundaries, laying the foundation for subsequent operations. The second component of SVTON is the Geometric Matching Module (GMM), which undertakes the crucial task of warping the selected garment. By employing geometric transformation techniques, GMM ensures proper alignment and fitment of the clothing item onto the underlying body structure. The final module, the Try-On Module (TOM), takes the warped garment generated by GMM and effectively applies it to the candidate, resulting in the synthesised appearance of the clothing item on the candidate.

The contributions presented in this paper are the following:

- The segmentation module demonstrates a remarkable ability to generate accurate segments based on the target garment, independent of the original garment's influence.
- The integration of an RGB skeleton image represents a substantial breakthrough in addressing the occlusion problem inherent in virtual try-on systems. By leveraging the RGB skeleton image as an input, our model can recognise the spatial relationship among the key joints.
- The architectural optimisation employed in our model, featuring a reduced number of convolutional layers, grants an advantage in terms of synthesising try-on images faster compared to previous approaches. This efficiency enhancement not only accelerates the overall image synthesis process but also maintains a comparable or superior level of output quality.

This manuscript represents an extension of a previous work we published [8]. We conduct a broader scope of experimentation, including an ablation study and evaluations conducted on additional datasets. a broader scope of experimentation, including an ablation study and evaluations conducted on additional datasets. The organisation of this paper is as follows: We start by delving into the background of previous studies in Section 2. Subsequently, in Section 3, we describe our novel model, SVTON. Furthermore, we present our empirical findings and their corresponding analysis in Section 4. Lastly, drawing upon the culmination of our experiments and results, we draw conclusive insights in Section 5. Interested readers may access the source code for our research, which is openly available at https://github.com/1702609/SVTON.

2. Background

In this section, we undertake a comprehensive review of the generative model employed in virtual try-on systems as well as other fashion-related applications. Our analysis will distinguish between the various types of virtual try-on solutions currently available, explaining their respective mechanisms and operational processes. Moreover, we will critically evaluate the inherent limitations associated with these virtual try-on methodologies, shedding light on their areas of weakness.

2.1. Generative Adversarial Network (GAN)

Generative Adversarial Networks (GANs) [9] represent an innovative approach that leverages two neural networks to achieve high-quality image synthesis [10–13] and manipulation [14–16]. The fundamental principle of GANs involves a generator network that attempts to deceive a discriminator network, which, in turn, learns to distinguish between real and fake samples.

In order to control the generated output images in the realm of GANs, the adoption of Conditional GAN (cGAN) [17] emerges as a promising solution. Various methodologies exist for guiding the image generation process within cGANs. Notable examples encompass the utilisation of class labels [18,19], textual descriptions [20–23], attributes [24], and sketches [10,25]. These techniques enable cGANs to produce images aligned with specific criteria or desired characteristics. Consequently, the applications of cGANs, particularly in the domains of virtual try-on and fashion-related contexts [26,27], have gained significant relevance.

A majority of virtual try-on models have incorporated the utilisation of the GAN mechanism either as a whole or within specific modules [3,4,28,29]. By integrating GANs into the virtual try-on framework, these models have demonstrated the ability to generate try-on images with exceptional fidelity.

However, it is crucial to acknowledge the challenges associated with cGAN-based methods when confronted with substantial spatial deformations between the target clothing and the pose of the individual. Notably, CP-VTON [2] has demonstrated instances where cGAN-based approaches can exhibit unstable image generation under such conditions. Consequently, it becomes imperative to develop prerequisite methods that effectively guide cGANs during the image synthesis process, mitigating potential issues arising from large spatial deformations.

2.2. Diffusion model

In recent studies, the performance of diffusion models has surpassed that of GANs in the domain of image synthesis [30]. These innovative generative models operate by denoising a Gaussian distribution sample iteratively until a coherent image is generated [31].

Like GANs, output images of diffusion models can be controlled through textual descriptions [32] or even input images [33]. This flexibility opens up various possibilities for their application in the field of fashion. Several notable works have explored the utilisation of diffusion models in fashion-related tasks. For instance, DreamPose [34] leverages diffusion models to synthesise realistic fashion videos from input fashion images. Additionally, DiffFashion [35] utilises diffusion models to generate new textures for clothing items based on reference appearance images.

Although integrating diffusion models into fashion synthesis is still a relatively new phenomenon, it is evident that only a limited number of studies have been conducted thus far. However, given the promising results achieved and the potential of diffusion models in virtual try-on applications, it is plausible to expect that their utilisation in this area will become more prevalent in the near future.

2.3. 3D virtual try-on

3D virtual try-on methodologies leverage simulated 3D clothing data to integrate and accurately apply garments onto a 3D avatar seamlessly. This advanced method enables the precise representation of intricate geometrical details, including the realistic rendering of clothing wrinkles, regardless of the avatar's pose or movement.

The Dressing Any Person (DRAPE) method [36] utilises a 3D approach to virtual fitting systems. The model training process involves using a dataset comprising 3D avatars characterised by diverse shapes in a consistent pose, as well as a dataset featuring a singular body shape

navigating through a range of poses. Notably, the method integrates a clothing deformation model performing geometrical transformations such as rigid rotation and calculating clothing shape variations. This model transfers the warped garments onto the 3D avatars by mapping the body shape parameters to the corresponding clothing shape parameters, thereby ensuring a realistic and accurate virtual try-on.

Sekine et al. propose a method [37] that effectively addresses the challenge of adjusting 2D clothing images to users by leveraging the analysis of 3D body shape models derived from single-shot depth images. Their approach entails an examination of the candidate's body shape, enabling the system to suggest appropriate clothing images characterised by shape similarity. As a result, the proposed method empowers candidates to virtually try on products that align with their unique physique, thereby enhancing the practicality of the virtual fitting experience.

ClothCap [38] comprises an automatic segmentation technique that employs a 3D avatar as a reference to extract precise 3D scan sequences. Additionally, it incorporates a sophisticated multi-mesh template tracking approach and a method specifically designed to adapt dynamic clothing to diverse body shapes. Collectively, ClothCap has exhibited exceptional capabilities in accurately dressing a 3D avatar with various types of clothing items.

TailorNet [39] employs a simple multi-layer perceptron (MLP) to predict the low-frequency geometry of clothing. For high-frequency geometry, each model consists of an MLP that predicts deformation based on the pose, and the weights of the mixture are determined using a kernel that evaluates similarity in style and shape. Through various experiments, the researchers demonstrate that TailorNet exhibits strong generalisation to new poses, accurately predicts garment fit based on body shape and retains detailed wrinkle information.

CloTH-VTON [40] introduces a hybrid methodology that combines 2D and 3D techniques. The proposed approach involves transforming the target clothing's 2D image into a 3D object, enabling more realistic deformations through physics calculations. By adopting this strategy, CloTH-VTON effectively utilises the strengths of the 2D method for generating or preserving body parts while also capitalising on the enhanced realism and flexibility offered by the 3D approach. Another noteworthy advancement in this field is M3D-VTON [41], which follows a similar approach of integrating both 2D and 3D techniques. It employs a 2D image-based virtual try-on process and subsequently infers a 3D representation of a person wearing the desired garments.

A critical limitation of 3D-based virtual try-on systems lies in their heavy reliance on 3D measurement data, rendering them impractical for integration within the context of online environments.

2.4. 2D virtual try-on

CAGAN [42] emerged as the pioneering model in the domain of 2D virtual try-on. It features a single network integrating multiple images to generate a try-on image. However, a significant drawback of this approach lies in its reliance on both the target and original clothing images during the inference process. Consequently, this dependency poses a significant limitation to practical utilisation, as it requires consumers to provide images of the original garment, which is unfeasible in real-world scenarios.

VITON [1] and CP-VTON [2] stand out as practical models due to their utilisation of the target clothing only. These models consist of two stages: warping the garment and conducting the try-on. In the initial stage, VITON employs Thin-Plate Spline (TPS) to align the candidate's body with the target garment. TPS parameters are calculated by establishing correspondences between keypoints on the target clothing and the candidate's body. However, TPS's drawback lies in its potential failure to preserve clothing logos and textures during the warping process.

To address this limitation posed by TPS, CP-VTON introduces the Spatial Transformation Network (STN) [43] in their model, featuring convolutional layers that extract high-level features from both the candidate and target clothing. These features are subsequently combined and fed into a regressor network responsible for predicting the optimal parameters for TPS. By leveraging this enhanced warping process, CP-VTON significantly improves the overall quality of garment alignment.

ACGPN [3] and LA-VITON [44] assert that the utilisation of STN alone is inadequate for effectively controlling TPS transformations, as evidenced by instances where STN failed to prevent distortions in clothing texture and pattern caused by TPS. In order to address this concern, both ACGPN and LA-VITON have devised respective solutions.

ACGPN tackles the issue by imposing a constraint on TPS that restricts shape deformation. Specifically, ACGPN employs second-order difference constraints on the TPS to prohibit unreasonable distortions in the shape of the garment and prevent undesirable texture disruptions. This approach effectively preserves the integrity of the garment's shape and ensures the preservation of its texture.

On the other hand, LA-VITON introduces a novel loss function known as the Grid Interval Consistency (GIC) loss. By focusing on the absolute difference between the x and y coordinates of the grid to be mapped, the GIC loss enforces consistency in the grid intervals, thereby mitigating distortions in the clothing texture and pattern.

Various models have proposed alternative approaches to enhance the performance and control of TPS. For instance, KP-VTON [45] leverages keypoint prediction in the target clothing as control points for TPS. This strategy enables finer control over the warping process, leading to further refinement of the try-on results.

WAS-VTON [46] incorporates Neural Architecture Search (NAS) techniques to discover clothing category-specific warping networks. By customising the warping module based on the specific clothing category, WAS-VTON achieves improved accuracy and effectiveness in the virtual try-on process.

Furthermore, CP-VTON+ [47], KP-VTON, and VITON-HD [6] have introduced modifications in their candidate representation to ensure that the warping module receives relevant and meaningful information. CP-VTON+ addresses the issue of misclassification by correcting the segmentation that wrongly categorised the neck and chest as background. In the case of KP-VTON, it replaces the conventional human parser with DensePose [48], a more precise body parts estimation method that is unaffected by the clothing worn by the candidate. This substitution enhances the accuracy of body part alignment during the warping process. Lastly, VITON-HD modifies the segment label by eliminating the clothing item's shape, enabling the warping module to focus solely on warping the garment in a manner that naturally matches the candidate's body.

Several virtual try-on methods have leveraged the segmentation module as an integral component of their models to enhance their performance [3–6]. The segmentation module plays a crucial role by providing regional boundaries that guide the dimension of the warped garment, thus improving the overall quality of the virtual try-on results.

In particular, ACGPN [3] and VTNFP [4] have demonstrated the effectiveness of incorporating segmentation labels into their methods. By preserving non-targeted body parts, such as hands, through the segmentation labels, these approaches enhance the realism of the generated try-on images and provide additional guidance for image synthesis.

Similarly, VTNFP follows a similar approach. Their method excels in producing more accurate and consistent labels by incorporating non-local operations that capture long-range dependencies and effectively eliminate patchy inconsistencies that may arise within the segments.

VITON-HD [6] and C-VTON [7] models have enhanced the quality and realism of synthesised try-on images by introducing novel conditional normalisation layers, which play a crucial role in improving the fidelity of the try-on results.

VITON-HD introduces the Alignment-Aware Segment (ALIAS) normalisation layer, specifically designed to address misaligned regions and enhance realism. The ALIAS normalisation effectively removes

irrelevant data from the clothing texture in these misaligned regions, replacing it with a generated clothing texture that aligns with the surrounding context. This process significantly contributes to the overall realism of the try-on image. Additionally, the normalising layer in VITON-HD leverages semantic information in an efficient manner, further improving the quality and coherence of the synthesised results.

Similarly, C-VTON utilises Context-Aware Normalisation (CAN) to achieve enhanced try-on image synthesis. The CAN normalisation layer efficiently utilises the information provided by the input images and delivers vital contextual cues to the generator. This enables the generator to better understand the contextual relationship between the clothing and the candidate, resulting in improved quality and realism of the synthesised try-on images.

The generation of realistic try-on images on significantly different clothing cases, along with the challenge of handling body part occlusion, has proven to be a persistent issue in several previous virtual try-on models. Notably, the segmentation module employed in these models often encounters difficulties in reliably producing accurate segments. Instances have been observed where the segmentation module erroneously generates segments for only a portion of the arm, even when the target garment is long-sleeved. Furthermore, these models struggle to render the arm when it occludes the body, further exacerbating the problem.

In light of these challenges, we have taken the initiative to conduct an evaluation of virtual try-on models, specifically focusing on their performance in significantly different clothing cases and their ability to handle body part occlusion. By thoroughly assessing and comparing these models, we aim to shed light on their strengths and limitations. In the subsequent sections, we will delve deeper into these aspects and present our findings and analyses.

3. Method

The proposed model, Simplified Virtual Try-On (SVTON), is composed of three distinct modules, each serving a specific role in the try-on process:

- Predictive Human Parsing Module (PHPM): This module plays a pivotal role in generating accurate body part segments based on the target garment. By leveraging useful input images, PHPM efficiently extracts essential information about the candidate's body, facilitating subsequent steps in the try-on process.
- Geometric Matching Module (GMM): The GMM module takes on the responsibility of aligning the target garment with the candidate's body. Through sophisticated geometric transformations and warping techniques, GMM ensures optimal alignment, thereby enabling a realistic and visually appealing try-on experience.
- Try-On Module (TOM): The TOM module serves as the final step in the try-on process, responsible for seamlessly merging the warped garment onto the candidate's body. Additionally, TOM possesses the capability to generate or preserve the non-targeted body parts, ensuring the overall coherence and naturalness of the final result.

Fig. 2 provides a visual representation of all the modules involved in the proposed model, along with their corresponding input images. The figure illustrates the system's comprehensive nature and highlights the interaction between PHPM, GMM, and TOM.

3.1. Candidate representation

The candidate representation plays a crucial role in providing essential information to enable the seamless functioning of the individual modules. In existing works such as VITON [1] and ACGPN [3], the 18-keypoint pose map has been commonly employed. However, we argue that relying solely on this representation can lead to the generation of unrealistic images in scenarios involving self-occlusions. This limitation arises from the lack of spatial relationship information among the joints within the pose map.

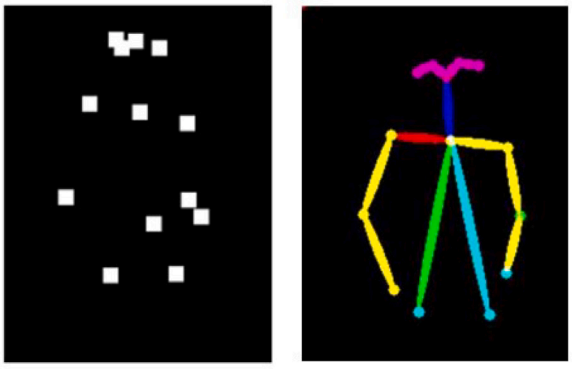


Fig. 1. Disparities between the two types of pose maps representing the human body. On the left, we observe the conventional 18-keypoint pose map, which lacks spatial connections with the remaining joints. In contrast, the pose map on the right showcases the enhanced representation achieved through the utilisation of RGB skeleton pose.

To address this issue, we propose the utilisation of the RGB skeleton pose as a superior alternative. The RGB skeleton pose shows the spatial relationships among the joints, offering a more comprehensive representation. Our proposed model can generate more realistic and visually appealing virtual try-on results by leveraging this enhanced pose information.

Fig. 1 visually demonstrates the advantage of employing the RGB skeleton pose, showcasing how the joints are accurately connected in the corresponding image. This visualisation serves to emphasise the importance of incorporating spatial relationship information within the candidate representation for improved virtual try-on outcomes.

We employ widely recognised and popular 2D pose estimators proposed by Cao et al. [49] and Simon et al. [50] to obtain the RGB skeleton pose image. These state-of-the-art pose estimation techniques have demonstrated remarkable accuracy and efficiency in capturing human body poses from 2D input images.

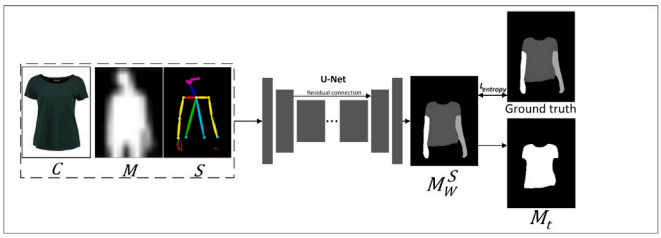
3.2. Predictive Human Parsing Module (PHPM)

PHPM, as depicted in Fig. 2a, serves as a segmentation module responsible for predicting appropriate labels for the torso and arms based on the target clothing. This module accomplishes this by taking input in the form of the RGB skeleton pose S, the candidate's body mask M, and the target clothing C, which are fed into our generative model.

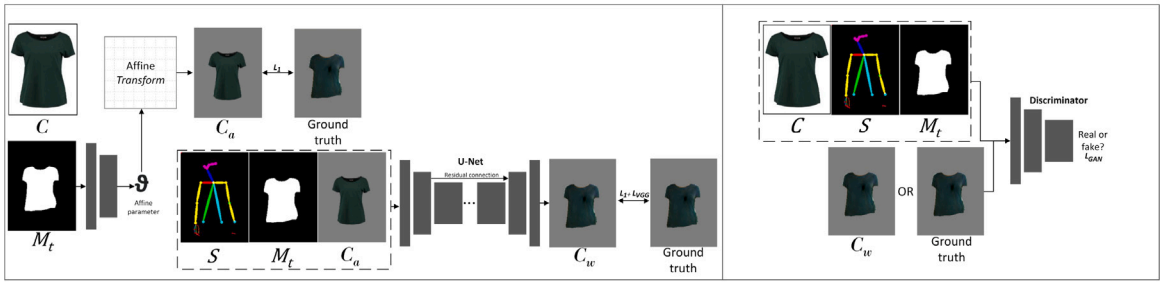
The generative model processes the input data and generates a four-channel image denoted as M_W^S. Each channel within M_W^S corresponds to a specific region, including the background, torso, and left and right arms. The purpose of the network is to acquire the necessary knowledge to generate reasonable body labels by effectively analysing the characteristics of the garment and its spatial relationship with the candidate's body.

We employed a two-step process to obtain the blurred variant of the candidate's body mask M. Firstly, we reduced the resolution of M by a factor of 16, effectively downsampling the image. Secondly, we resized the downsampled image back to its original dimension. This resizing operation introduces a blurring effect, resulting in the desired blurred variant of M. This procedure creates a smoothed representation of the candidate's body mask, which can enhance the overall quality and fidelity of subsequent steps in the virtual try-on process.

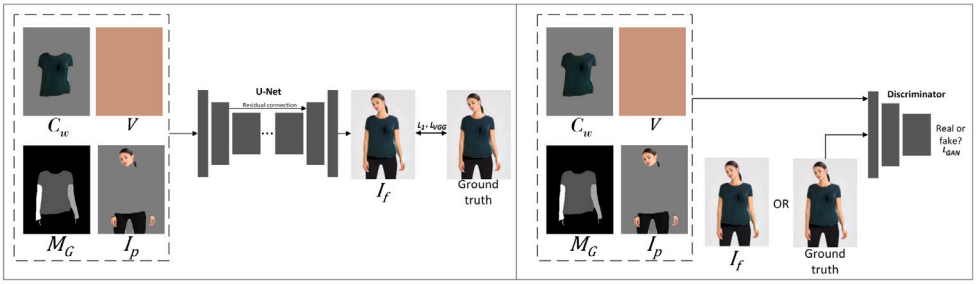
Many of the latest segmentation models, such as the segment anything model (SAM) [51], grounded DINO [52], and other vision transformer-based models [53], are not designed to predict how a



(a) Predictive Human Parsing Module (PHPM)



(b) Geometric Matching Module (GMM)



(c) Try-On Module (TOM)

Fig. 2. The structure of the proposed model. It involves three modules: PHPM, GMM, and TOM. (a) PHPM takes input variables such as the target garment image (C), RGB skeleton pose (S), and blurred binary mask (M) to generate a 4-channel image (M_W^S) that maps to the background, torso, and left and right arm regions. (b) GMM utilises a spatial transformation network (STN) with an affine transform, taking M_t and C to perform a pre-conditional warp on the target garment (C_a). Further refinement is achieved by passing M_t, C_a, and S through a generative model, resulting in the refined garment image (C_w). (c) Finally, TOM combines prerequisite elements, including the refined garment image (C_w), prerequisite image (I_p), and segment mask (M_G), along with average skin colour (V), using a generative model to synthesise the final try-on image (I_f). We utilise the discriminator for GMM and TOM during training.

segment may look on an object when a change is applied; they only segment images in their current form. Our model, on the other hand, is capable of predicting how a segment may look as if the person is already wearing the current clothing.

The application of cross-entropy loss is highly advantageous in deep learning models that involve predicting probabilities for multiple classes [54]. In the case of PHPM, which outputs four channels for segment prediction, cross-entropy loss plays a crucial role in evaluating the alignment between the generated segments and the ground truth labels. By formulating the loss function, denoted as L_{PHPM}:

$$L_{\text{PHPM}} = \lambda L_{\text{entropy}} \tag{1}$$

where L_{entropy} is the cross-entropy loss [54], and λ is the parameter to magnify the loss.

3.3. Geometric Matching Module (GMM)

GMM, depicted in Fig. 2b, plays a pivotal role in aligning the garment with the candidate's pose during the virtual try-on process. GMM incorporates a Spatial Transformation Network (STN) [43] in its initial stage to position the clothing accurately around the torso region. By leveraging the generated torso segment M_t and the target garment image C as inputs to the STN, a geometrically transformed and rotated image of the target garment, denoted as C_a, is obtained. This transformation facilitates the subsequent generative model in capturing intricate details such as complex textures and logos present in the clothing.

Our generative model is then employed to extract clothing features, including texture and logo, and additionally synthesise natural wrinkles on the garment. It utilises S, M_t, and C_a as inputs. The generative model's output is a warped garment image, referred to as C_w, which precisely conforms to the shape outlined by M_t, ensuring a seamless and natural fit of the clothing onto the candidate's body.

The utilisation of the Thin-Plate Spline (TPS) algorithm for the Spatial Transformation Network (STN) has been commonly employed in previous works. However, research studies such as ACGPN [3] and LA-VITON [44] have demonstrated instances where TPS may result in undesired deformation of the garment during the transformation process. To address this issue, they introduced a constraint in TPS to limit its ability to excessively deform the shape. In our approach, we opted for an affine transform instead, as it achieves comparable performance to TPS while maintaining garment integrity. Moreover, the use of affine transform offers the advantage of requiring fewer trainable parameters due to its reduced degrees of freedom in shape deformation. This characteristic enhances the efficiency of training the model.

To train GMM, we employ a discriminator that follows a similar architecture to the discriminator utilised in Pix2PixHD [55]. We use the cGAN loss [17]; the loss is formulated as:

$$\begin{aligned} L_{\text{GAN}}(x, y) =& \mathbb{E}_{x,y}[\log D(x, y)] \\ &+ \mathbb{E}_x[\log(1 - D(x, G(x)))] \end{aligned} \tag{2}$$

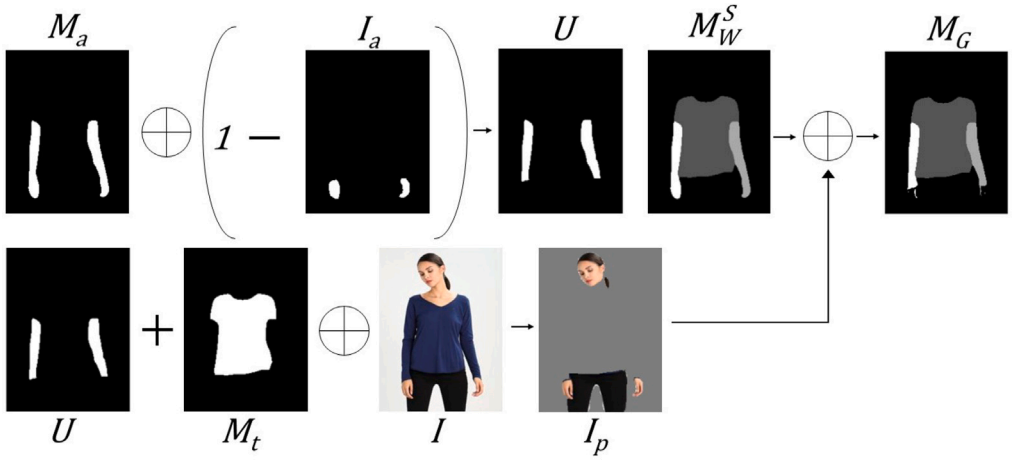


Fig. 3. Arithmetic process of generating M_G and I_p. The process involves an arithmetic procedure that encompasses element-wise multiplication and mask manipulation.

We calculate further losses for GMM by utilising the L1 and VGG loss functions. We formulate the loss function as L_{GMM}:

$$L_1(x, y) = |x - y| \tag{3}$$

$$L_{\text{VGG}}(x, y) = \lambda|\phi_5(x) - \phi_5(y)| \tag{4}$$

$$\begin{aligned} L_{\text{GMM}} =& L_1(C_w, C_{\text{gt}}) + L_1(C_a, C_{\text{gt}}) + L_{\text{VGG}}(C_w, C_{\text{gt}}) \\ &+ L_{\text{GAN}}(f, C_{\text{gt}}) \end{aligned} \tag{5}$$

where the symbols represent as follows: C_w denotes the warped garment; C_a is the prerequisite warp performed by affine transform; C_{gt} denotes the ground truth of the warped garment; f denotes the channel concatenation of M_t, C and S; G denotes the generator; D denotes the discriminator. The G for this module would be Fig. 4(a). L_{VGG} is the VGG perceptual loss [56] in which ϕ_5 represents the output of feature map of C_w and C_{gt} from the pre-trained VGG19 model. We use the fifth layer of the VGG network. λ is a parameter to control the loss value, which has the same value as Eq. (1).

3.4. Try On Module (TOM)

The generated segments play a vital role in providing guidance to the Try-On Module (TOM) (Fig. 2c), enabling it to determine the appropriate areas to preserve non-targeted body parts and to generate the arms. Nevertheless, before proceeding, pre-processing steps are necessary to get the desired input data. We show the method below:

$$U = M_a \otimes (1 - I_a) \tag{6}$$

$$I_p = (U + M_t) \otimes I \tag{7}$$

$$M_G = M_W^S \otimes I_p \tag{8}$$

where $\otimes$ denotes element-wise multiplication, we perform element-wise multiplication on I_a and M_a to produce U. The segment denoted as U serves as a crucial indicator for TOM, signifying the need to generate the appropriate arm length, particularly in cases where a transition occurs from long-sleeved to short-sleeved garments. I_p represents the preservable non-targeted body part, while M_G serves as a spatial guide, dictating the placement of the warped garment C_w and providing instructions for arm synthesis. To enhance the clarity and understanding of the intricate process involved in generating I_p and M_G, we have incorporated Fig. 3, which visually illustrates the steps in this arithmetic operation.

The determination of the average skin colour V involves a calculation process wherein the candidate's source image is analysed to derive the average pixel value specifically from the arm region.

Due to the inherent complexity involved, the generative model alone cannot accurately generate the intricate details of the candidate's hands. This limitation becomes evident when examining the try-on images produced by VITON [1] and CP-VTON+ [47] in the third row of Fig. 7, where the hands appear blurry, and the fingers are barely discernible. Our proposed approach addresses this challenge by focusing on preserving the original hands from the source image. This objective is accomplished through the creation of the maximum preservable region image, denoted as I_p. By excluding the hand regions from the corresponding mask M_G, we ensure that the generative model prioritises the preservation of the candidate's original hands. Both I_p and M_G play a pivotal role as vital inputs to the generative model, outlining the areas to be preserved (e.g., hands) and the regions requiring generation (e.g., arms). Consequently, with I_p, M_G, V and C_w as inputs, the generative model synthesises the final try-on image, denoted as I_f, which embodies a seamless integration of the clothing onto the candidate's body.

We train TOM with a discriminator [55] and utilise the VGG and L1 loss functions. The loss formula for TOM is presented as follows:

$$L_{\text{TOM}} = L_1(I_f, I) + L_{\text{VGG}}(I_f, I) + L_{\text{GAN}}(f, I) \tag{9}$$

where I_f denotes the final generated virtual try-on image; I denotes the ground truth; f denotes the channel concatenation of V, C_w, I_p, M_G. The G for TOM would be Fig. 4(b).

4. Experiments

The proposed model has undergone comprehensive evaluation in comparison to established methods such as VITON [1], CP-VTON+ [2], and ACGPN [3]. In order to present a thorough assessment, we provide an in-depth analysis of both qualitative and quantitative comparison results. Furthermore, we examine the individual contributions of each model component to the overall performance. The datasets employed in our experiments, as well as the implementation details of the networks, are also elaborated upon, ensuring transparency and reproducibility.

The matching of candidate and clothing images can be performed in two distinct settings: paired and unpaired. In the paired setting, the candidate is depicted wearing the original clothes, and this configuration is employed during training and for quantitative evaluation purposes. On the other hand, the unpaired setting involves pairing the candidate with a new garment, resembling the manner in which a consumer would utilise it.

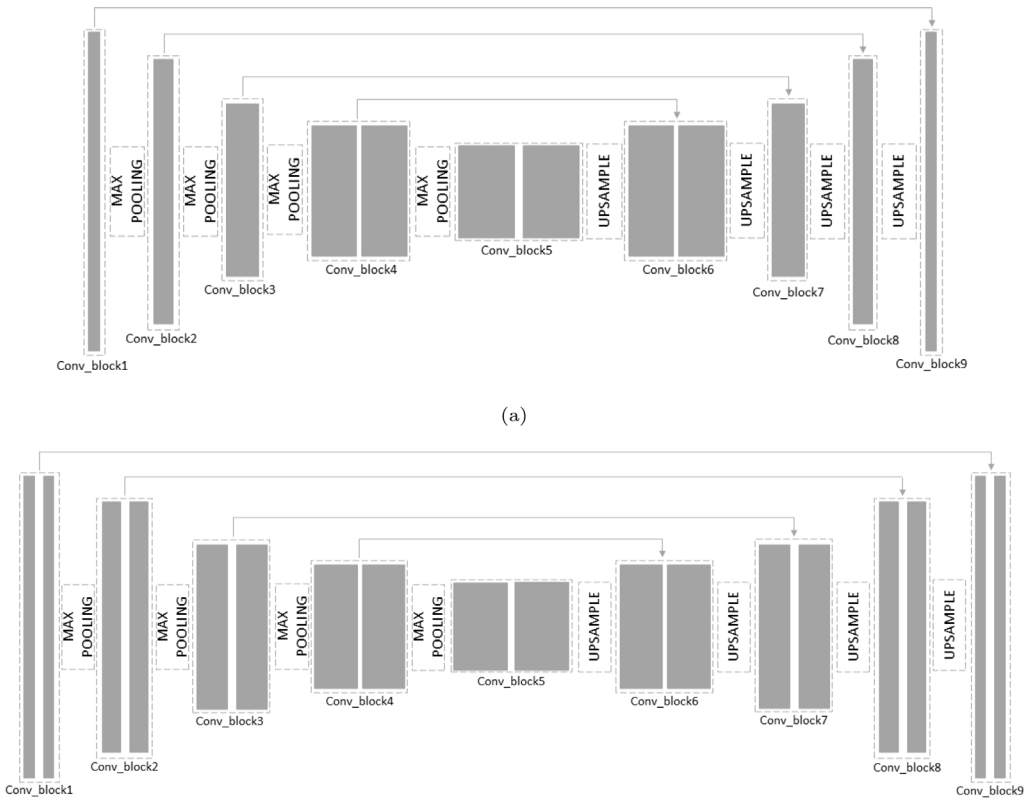


Fig. 4. The generative model architectures for the three model components PHPM, GMM and TOM. (a) The generative model used in both PHPM and GMM, which has fewer convolutional layers, making it more efficient for the PHPM to produce accurate labels and preserves clothing detail better for the GMM. (b) The generative model used in the TOM.

4.1. Dataset

Our SVTON model has been trained on the VITON dataset [1], which comprises a substantial training set of 14,221 pairs of candidate images and their corresponding clothing. The researchers have allocated a separate testing set of 2032 image pairs. The dataset's images exhibit a standardised resolution of 256 × 192 pixels. We incorporated the RGB skeleton pose, leveraging the methodologies proposed in [49, 50].

For quantitative evaluation, we employed the testing sets derived from the VITON [1] and VITON-HD [6] datasets, each consisting of 2032 image pairs. For qualitative assessment, we only utilise the VITON test set. It is important to note that despite the similarity in dataset names, the images within them are distinct, featuring markedly different candidates and clothing items. By conducting evaluations on both datasets, we aimed to gain comprehensive insights into the performance of our model.

4.2. Implementation

Our generative model employs the U-Net architecture [57], which serves as the backbone across all three modules of our approach. However, we employ a strategic design decision by incorporating a reduced number of convolution layers in the PHPM and GMM modules.

The architectural design of PHPM and GMM's U-Net is depicted in Fig. 4(a), showcasing the underlying framework. The U-Net's encoder consists of five convolutional layers, each employing a kernel size of 3. The number of filters progressively increases through the layers, with values of 64, 128, 256, 512, and 512, respectively. Additionally, a max pooling operation is applied after each layer, effectively reducing the feature map dimensions by a factor of 2.

Within the latent space, two convolutional layers are employed, each utilising a kernel size of 3. The filter sizes for these layers are set at 1024 and 1024. As for the decoder, it comprises five convolutional layers, each with a kernel size of 3. The number of filters in the decoder layers follows a pattern of 512, 512, 256, 128, and 64. Notably, an upsampling operation is performed between each layer in the decoder, resulting in a doubling of the feature map resolution. Skip connections are employed to establish connectivity between the encoder and decoder.

The generative model of TOM retains the original U-Net architecture [57], as illustrated in Fig. 4(b). The kernel size for TOM's structure remains consistent at 3 throughout.

GMM utilises an STN consisting of five convolutional layers and a max pooling layer with a stride size of 2. This STN enhances the network's ability to manipulate spatial transformations.

The discriminator's structure is akin to that of Pix2PixHD [55]. It commences with four convolutional layers, employing a kernel size of 4. The respective numbers of filters for these layers are set at 64, 128, 256, and 1. Finally, a sigmoid function is appended to the output of the discriminator to facilitate appropriate classification.

We have trained the individual modules independently, each with a specific number of epochs. PHPM was trained for 20 epochs, GMM underwent 40 epochs, and TOM was trained for 100 epochs. The modules were trained in paired settings.

We employed the Adam optimiser, a popular choice in deep learning applications. The optimiser was configured with a learning rate hyperparameter of 0.0002, a value empirically determined to yield desirable results. Furthermore, we set the β_1 parameter to 0.5 and β_2 to 0.999.

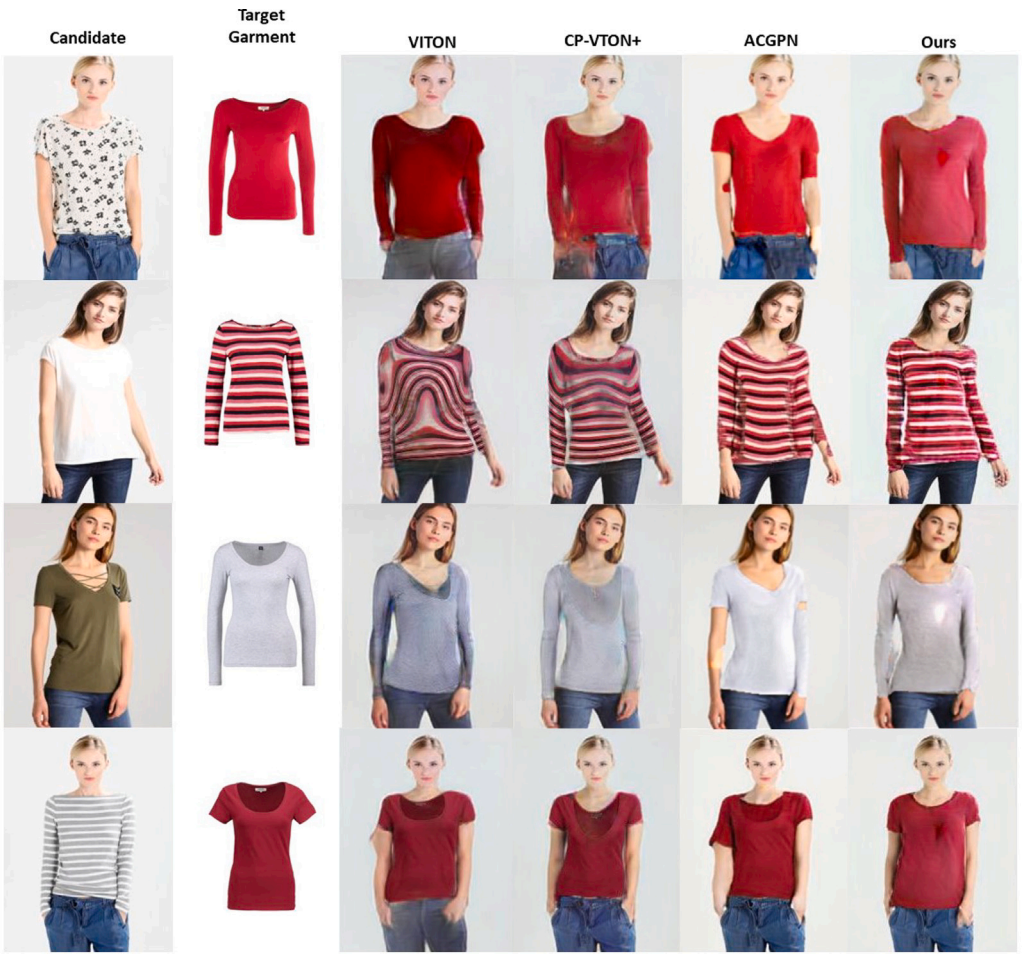


Fig. 5. Comparison of significant difference cases. Our proposed method showcases superior accuracy in try-on tasks involving substantial differences, such as transitioning from short sleeves to long sleeves or vice versa, outperforming previous approaches. The implementation of PHPM effectively utilises refined candidate representations (i.e. RGB skeleton pose and blurred mask) to generate precise labels, thus contributing to enhanced performance. ACGPN and our method rely on producing correct labels. Failure to do so may result in erroneous try-on images and inferior performance compared to non-segmentation methods such as VITON and CP-VTON+. The first and third rows affirm this observation.

4.3. Qualitative analysis

For a qualitative visual comparison of the proposed method against other models, we organise the unpaired setting results in three categories of significant different clothing, occlusion and easy cases.

4.3.1. Significant difference cases

One of the notable accomplishments of our work is in cases involving significant differences in candidate and garment pairing. This superiority is demonstrated in Fig. 5. In scenarios where a candidate wearing a short-sleeved garment is paired with a long-sleeved garment, or vice versa, our method excels by preserving intricate details and textures of the clothing, thereby generating accurate and visually pleasing try-on images.

By closely examining the 2nd and 4th rows of the figure, we observe that VITON and CP-VTON+ exhibit shortcomings such as colour alteration and failure to remove the back collar of the garment image. In contrast, our approach consistently maintains fidelity to the original clothing and successfully addresses these challenges. Furthermore, the 1st and 3rd rows indicate that ACGPN fails to synthesise short-sleeved to long-sleeved try-ons due to incorrect segmentation labelling, performing less favourably than VITON and CP-VTON+.

The effectiveness of our method can be attributed to the refined input data and a simpler network architecture employed by our segmentation module, which consistently produces accurate body labels regardless of the pairing. This is of utmost importance for methods utilising segmentation, such as our method and ACGPN, as an incorrect body label significantly degrades the performance, leading to inferior try-on results compared to methods that omit segmentation, such as VITON.

Furthermore, the clothing generated by our method exhibits a more visually pleasing appearance compared to VITON and CP-VTON+. This improvement is achieved by employing a generative model that refines geometrically warped garments, resulting in a more natural aesthetic. The generative model effectively adds realistic wrinkles and smooth textures to the garment. In contrast, VITON and CP-VTON+ primarily focus on geometric transformations and composite learning, capturing only a limited subset of the garment's textures and other properties. The poor performance of the composition mask is evident in its incorrect compositing of the back collar, leading to an unnatural appearance in the try-on images (e.g., 2nd and 4th rows).

4.3.2. Occlusion cases

An area where our model excels is in handling occluded cases. Previous works have faced challenges synthesising natural-looking try-ons when candidates' arms occlude their bodies. In contrast, our model consistently produces superior results in such scenarios, as demonstrated in Fig. 6.

In the 1st row of the figure, our model successfully creates a distinct boundary around the occluding arm by appropriately darkening it. On the other hand, ACGPN merges the arm with the torso, while VITON and CP-VTON+ render the arm in an unnatural manner. Remarkably,

Fig. 6. Comparison of occlusion cases. Our model distinguishes itself by successfully synthesising the arms of the candidate in occluded cases, providing accurate and realistic results. In contrast, other models either overlook this aspect or generate unnatural-looking arms, failing to preserve the integrity of body parts and often neglecting the hand altogether. The inclusion of RGB skeleton pose in our approach enables us to establish spatial relationships between body parts, facilitating effective arm synthesis even in challenging occluded scenarios.

our model is the only method capable of generating the correct try-on in the 2nd row, as it effectively preserves the occluded arm in front of the torso, resulting in a visually accurate output. Moreover, our model excels in preserving the hand in intricate regions (3rd row) and consistently maintains the integrity of all body parts (4th row). The RGB skeleton image has allowed the segmentation module to ensure that all body parts remain intact and helped the warping module to visualise how the arms are occluded.

The key factor enabling our model's exceptional performance in occluded cases is the utilisation of RGB skeleton pose. Our approach surpasses other methods that rely on an 18-keypoint pose map. The 18-keypoint pose map assigns each channel to a single joint of the human body. However, it does not provide any information about how each channel is connected to each other. For instance, the model would not be able to determine if the channel indicating the left hand of the person is connected to the adjacent channel pointing to the left elbow. In other words, spatial relationships do not exist in this pose map.

On the other hand, using the RGB skeleton pose clearly shows the spatial relationships among the keypoints. It provides essential information on how the joints of a human body are connected. For example, the RGB skeleton pose clearly demonstrates that the joint in the left hand is connected to the left elbow but not to the right elbow. This is what allows our model to exhibit a superior ability to handle occluded scenarios and maintain the integrity of the entire body structure.

4.3.3. Easy cases

In the case where the candidate is paired with a target garment of similar sleeve length, our model demonstrates robust performance as illustrated in Fig. 7. All models were consistent and accurately preserved sleeve lengths. This finding highlights our model's versatility, as it is capable of handling both easy and significant difference pairings, surpassing previous works that are more suitable for the former.

Even in the aforementioned easy cases, SVTON possesses notable advantages over previous methods. Notably, the last two rows showcasing occluded cases exemplify SVTON's ability to synthesise try-ons with a highly naturalistic appearance. These results reinforce the effectiveness of our model, as it consistently generates visually pleasing and realistic outcomes even in scenarios where occlusion is present, setting it apart from other approaches.

4.3.4. Comparison on the VITON-HD dataset

As opposed to Figs. 5–7 where they were evaluated against the test set of VITON dataset, in this section, we have conducted experiments on the VITON-HD dataset [6]. Our aim is to showcase the performance of our model in a more diverse scenario and make it easier to compare it with more advanced virtual try-on models like the VITON-HD. As shown in Fig. 8, VITON-HD generates higher resolution and better quality virtual try-on images. However, our outcomes were comparable, and our performance was similar. The figure clearly depicts that both models can accurately apply clothing to the candidate. There

Fig. 7. Comparison of easy cases. Our proposed approach and previous works demonstrate similar outcomes in terms of performance in easy cases. Across all models, the sleeve length is effectively preserved, ensuring appropriateness and consistency throughout the try-on process.

were instances where our model performed better than VITON-HD. For example, in the last row of the figure, our model preserved the hand, whereas VITON-HD synthesised the hand unrealistically.

Furthermore, our model has the advantage of requiring less computational resources and synthesising images at a faster rate. Therefore, our model may be more attractive to businesses with a low budget and cannot afford to invest in powerful GPUs and other computational resources. They may prefer to use a slightly weaker but more efficient model.

4.4. Quantitative analysis

The structural similarity (SSIM) [58] metric measures the similarity between the generated image and the original image by assessing their luminance, contrast, and structural characteristics. The SSIM index quantifies the degree of agreement between the two images, where higher values indicate a stronger correspondence.

The fréchet inception distance (FID) [59,60] metric uses the Inception network [61] to extract feature representations from original and generated images. This metric measures the discrepancy between the feature distributions of the two image sets by calculating the fréchet distance. Importantly, a decreased FID score indicates a stronger resemblance between the feature distributions of the generated images and those of the real images.

The inception score (IS) [62] is a metric designed to assess the performance of generative models. It evaluates the diversity and visual appeal of the generated images by passing them through a classifier that has been pre-trained. The score is determined by computing the output probabilities and is based on the KL divergence between the class distribution of the generated images and the class distribution of a large collection of real images. A higher IS indicates that the generated images exhibit greater diversity and visual appeal.

The learned perceptual image patch similarity (LPIPS) [63] metric utilises a deep neural network that has undergone fine-tuning to evaluate the perceptual similarity of images. This network is specifically trained to capture human perception regarding image quality. By calculating the dissimilarity between the feature maps of two images across

Table 1
Quantitative comparisons among different techniques performed on VITON [1] and VITON-HD [6] test set. The table showcases the performance of our approach in relation to VITON [1], CP-VTON+ [47], and ACGPN [3]. Higher values indicate better results for SSIM and IS, while lower values are desirable for FID and LPIPS.

Method	Paired settings				Unpaired settings		
	SSIM ↑	FID ↓	IS ↑	LPIPS ↓	FID ↓	IS ↑	LPIPS ↓
VITON dataset							
VITON	0.801	19.463	2.946	0.0818	30.403	2.567	0.155
CP-VTON+	0.828	16.800	**3.012**	0.0714	31.594	2.520	0.281
ACGPN	0.843	**13.318**	2.805	0.0737	20.728	2.541	**0.131**
SVTON	**0.854**	15.662	2.719	**0.0647**	**17.607**	**2.615**	0.135
VITON-HD dataset							
CP-VTON+	0.828	31.150	2.948	0.1236	30.026	**3.254**	0.158
ACGPN	**0.829**	**20.834**	2.943	0.1028	25.770	2.957	0.146
SVTON	0.819	21.708	**3.092**	**0.0909**	**23.211**	2.873	**0.140**

Fig. 8. Comparison conducted on the VITON-HD dataset [6]. Although VITON-HD produces higher resolution outputs for virtual try-on images, our model is equally effective and more efficient in terms of computational resources and inference time.

various spatial scales, LPIPS determines their perceptual distance. The average of these dissimilarity values produces an overall score, where a lower LPIPS score indicates that the generated images demonstrate greater perceptual similarity to the real images.

Table 1 presents a comprehensive overview of the quantitative performance of our proposed method along with other state-of-the-art models, namely VITON [1], CP-VTON+ [47], and ACGPN [3], on both the VITON [1] and VITON-HD [6] test sets. In the paired setting, we employed all the metrics discussed earlier to evaluate the performance. However, in the unpaired setting, we excluded the SSIM metric due to the absence of ground truth for candidates wearing different clothing, as SSIM would not provide accurate scores in this scenario.

In the paired setting of both the VITON and VITON-HD test sets, our method and ACGPN show varying performance in different metrics. While we outperform ACGPN in one metric, they surpass us in another. Notably, according to the SSIM metric, both our models demonstrate a higher resemblance to the ground truth in terms of synthesising try-on images compared to VITON and CP-VTON+. Moreover, our models exhibit a similar level of image quality, as indicated by the closely aligned scores of FID, IS, and LPIPS. This suggests that both our models produce try-on images at a comparable level of visual fidelity.

In the unpaired setting, our method outperforms ACGPN in terms of FID and IS scores on the VITON test set. Additionally, in the VITON-HD test set, we achieve better scores than ACGPN for both FID and LPIPS metrics. These results indicate that our model consistently generates higher-quality try-on images, irrespective of the target garment. The performance advantage positions our model favourably, as consumers will utilise the unpaired setting in real-world scenarios. This highlights the practical relevance and superiority of our approach in producing visually appealing and accurate try-on images.

Furthermore, VITON-HD has achieved a FID score of 11.988, IS score of 3.198 and LPIPS score of 0.121 on the VITON-HD test set in unpaired settingsCP, which is better than our model listed in Table 1. However, it is important to note that the comparison is not entirely suitable, as VITON-HD produces much higher-resolution virtual try-on images. It is not enough to evaluate quantitative measures alone, as computation resources and time taken to achieve the result also need to be considered. Our model is significantly more efficient and demands less video memory of a GPU, making it a more affordable option for small businesses.

4.5. Ablation study

To thoroughly assess the impact of individual components, an ablation study has been conducted on PHPM and GMM, utilising diverse input images. The primary objective of this study was to gain a comprehensive understanding of the specific influences exerted by these components.

The ablation study is presented in Table 2 highlights the compelling advantages achieved when incorporating the truncated generative model (Fig. 4(a)) in conjunction with the utilisation of RGB skeleton pose. This integration empowers our model to operate at peak performance, delivering unparallelled levels of fidelity in its output.

We have conducted an ablation analysis on PHPM to assess its performance under various training conditions. Specifically, we employed three different techniques: training on an 18-keypoint pose map,

Fig. 9. Ablation study of PHPM. Ablation study of PHPM. The 18-keypoint pose creates random artefacts and spots in the arm label, which will expose bare skin in the final synthesised image. Using the standard U-Net (Fig. 4(b)) with RGB skeleton pose can reduce the occurrence of random patches. Our truncated U-Net (Fig. 4(a)) further improves the performance of segmentation synthesis and ensures that body labels remain intact.

Table 2
Ablation study of PHPM and GMM. The findings derived from this table indicate that employing RGB Skeleton pose alongside a truncated generative model yields noteworthy enhancements in the performance of both modules. This combined approach leads to the generation of virtual try-on images characterised by superior fidelity and quality.

Method	SSIM ↑	FID ↓	IS ↑	LPIPS ↓
PHPM				
18-keypoint	0.715	–	–	–
Standard U-Net	0.866	–	–	–
Normal mask	0.764	–	–	–
Ours	**0.872**	–	–	–
GMM				
18-keypoint	0.810	32.991	3.639	0.0858
Standard U-Net	**0.899**	**23.771**	3.822	**0.0461**
Ours	0.898	24.846	**3.989**	0.0481

utilising the generative model depicted in Fig. 4(b), and employing an unblurred mask. Our proposed approach consistently achieved higher SSIM scores than these methods, indicating superior results.

The rationale behind the observed performance discrepancy can be explained by examining Fig. 9. In the case of the 18-keypoint pose map, it is evident that the arm segment exhibits noticeable patches (as seen in the 2nd and 3rd rows) due to the lack of spatial information regarding the interconnections between joints. Consequently, the resulting segmentation is suboptimal.

Furthermore, the standard generative model (Fig. 4(b)) demonstrates limitations in preserving the integrity of segmented regions. For instance, in the 2nd row and 5th column of Fig. 9, the generated output fails to generate the candidate's hand. This weakness undermines the overall quality of the parsing results.

Lastly, Fig. 11 showcases the advantages of employing a blurred mask over a conventional mask. The segmentation module must generate smooth segments that facilitate the natural fit of the new target garment on the candidate. By employing a blurred mask, the information pertaining to the previous clothing becomes unseeable, resulting in visually smoother and more natural-looking segments, as shown in the 1st row.

The ablation study conducted on GMM reveals substantial advantages associated with the utilisation of the RGB skeleton pose. Both our proposed approach and the standard generative model (depicted in Fig. 4(b)) have outperformed the approach employing the 18-keypoint pose map. The SSIM metric affirms that the warped garments synthesised by our approach and the standard generative model approach closely resemble the ground truth, indicating the effective preservation and alignment of details and texture. Moreover, the quality and realism of the warped garments generated by our approaches surpass the method using the 18-keypoint pose map, as indicated by FID, IS, and LPIPS. Fig. 10 visually demonstrates that both our proposed approach and the standard generative model exhibit improved handling of occlusion cases. Although the standard generative model approach has quantitatively outperformed our approach by a small margin, their scores are highly similar. Therefore, we have selected the shorter generative model for its efficiency in accelerating the try-on process.

4.6. Inference time

To evaluate the efficiency of our model, we conducted a comparative analysis of its inference time with two other models: ACGPN [3] and VITON-HD [6], as shown in Table 3. Both of these models have a similar tripartite structure to ours. We experimented with these models on an RTX 2070 GPU. For our method and ACGPN, we used 200

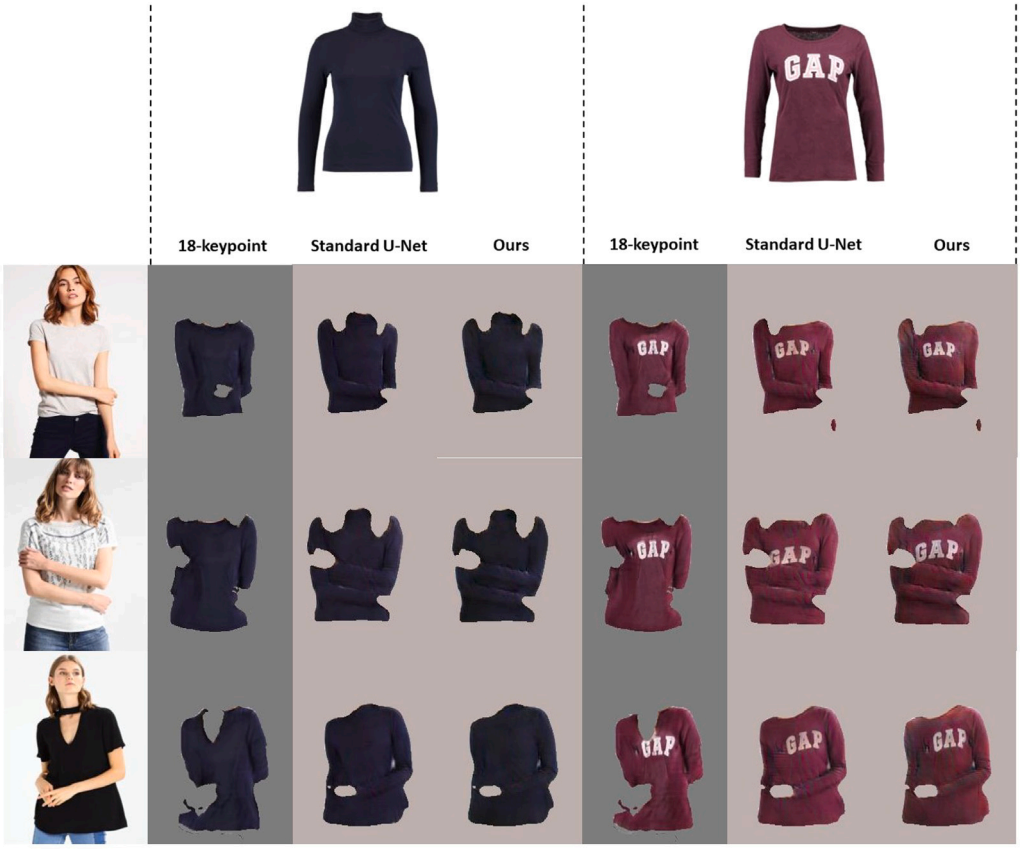


Fig. 10. Ablation study of GMM. Ablation study of GMM. The use of an 18-keypoint pose map is not sufficient for GMM to handle cases of occlusion, as the spatial relationship is not illustrated. However, the use of an RGB skeleton on both standard (see Fig. 4(b)) and truncated U-Net (see Fig. 4(a)) provides a much clearer solution for handling occlusion. This approach is able to effectively synthesise the arms, which are visible and clearly show the person crossing them.

Table 3
Performance comparison. The average time required to synthesise a single virtual try-on image using the same hardware. Our method demonstrates the lowest inference time due to the reduction of parameters.

	Inference time (ms)	# Parameters (millions)	Time saved (%)
ACGPN	164.3	136	50.2
VITON-HD	809.6	135	89.9
Ours	**81.9**	98	–

images from the VITON [1] test set, while VITON-HD utilised the high-resolution test set of the VITON-HD [6] dataset, for which we also provided 200 images. The results demonstrate that our approach outperforms ACGPN and VITON-HD by reducing the inference time by 50.2% and 89.9%, respectively.

The increase in speed can be attributed to our use of our truncated U-Net, in which we reduced the number of convolution layers, thereby reducing the number of parameters. From Table 3, we can see that VITON-HD is slower than ACGPN, even though it has slightly fewer parameters. This is because it operates at a much higher image resolution than our model and the ACGPN model. The speed of inference is not solely determined by the number of parameters but also by the resolution of the data it is processing.

Our approach has a significant advantage: it can serve a larger consumer base and deliver results more quickly, even when running on weaker hardware. This distinct advantage offers considerable benefits for businesses looking for streamlined and efficient processes or companies that do not have the budget to pay for expensive GPUs.

4.7. Limitations

Despite the advancements achieved by the proposed methods, it is essential to acknowledge their limitations. One such limitation is the candidate representation's inadequate information regarding the leg-torso boundary, leading to the PHPM synthesising body segment labels of inconsistent sizes. This limitation is illustrated in Fig. 12, where the failure of PHPM to generate accurate body segment sizes is demonstrated. The examples highlight how TOM creates a visible gap between body parts due to the insufficient generation of the torso label by PHPM, thus failing to connect it with the leg. One potential solution to address this issue is to incorporate distinct images of the candidate's head and legs as the input, allowing PHPM to generate the correct length for the torso label. For instance, methods like VITON-HD [6] initially include the head and leg labels in the candidate representation before passing it to the segmentation module, enabling a more accurate synthesis of the torso label.

5. Conclusion

In conclusion, we have introduced Simplified Virtual Try-On (SVTON), a novel image-based virtual try-on model designed to tackle challenging scenarios characterised by significant differences between the original and target clothing, as well as substantial overlapping of body parts. The proposed model stands out for its two key features: (1) a candidate representation based on an RGB skeleton image, which enhances the spatial relationships among joints and leads to more robust and accurate segmentation results, and (2) the utilisation of a truncated generative model in both the segmentation and warping modules, with the warping module incorporating an efficient affine transform.

We conducted a comprehensive evaluation of the proposed model against state-of-the-art approaches such as VITON, CP-VTON+, and ACGPN. The evaluation encompassed qualitative comparisons, quantitative analyses, ablation studies, and computational runtime assessments. The results from these evaluations consistently demonstrate

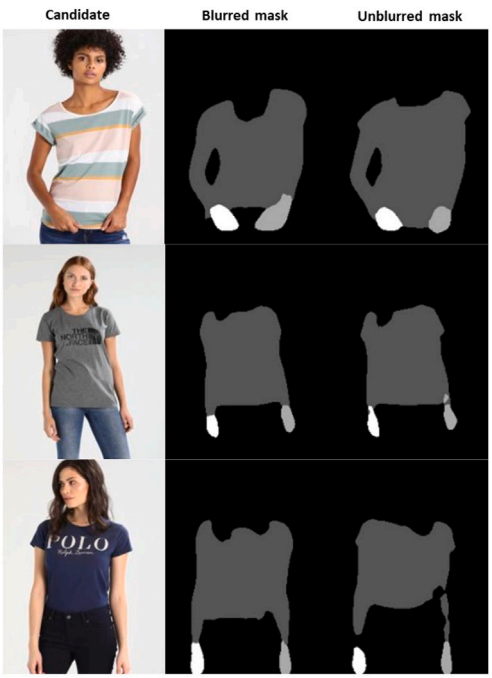


Fig. 11. Effects of utilising the blurred and unblurred mask for PHPM. In the 1st column, a blurred mask is employed, which effectively conceals the shape of the original clothing. This blurred mask facilitates the generation of smoother body segments by the PHPM, resulting in visually appealing output. Conversely, in the 2nd column, an unblurred mask is utilised, exhibiting a coarse shape that closely resembles the original clothing. This poses a challenge for the PHPM, as it struggles to remove the coarse shape information provided by the unblurred mask.

Fig. 12. Limitations of the proposed model. We do not provide the boundary of the candidate's head and legs to the Predictive Human Parsing Module, which may cause them to generate incorrect sizes of the body label. This will have a knock-on effect on subsequent modules and produce undesirable try-on.

the superior performance of the proposed model across various scenarios. Furthermore, the proposed model exhibits notable efficiency advantages, outperforming previous models in terms of computational runtime.

Overall, our research underscores the effectiveness of SVTON in addressing complex virtual try-on challenges and establishes its superiority over existing models. The proposed model's enhanced segmentation accuracy, spatial understanding, and computational efficiency make it a promising solution for practical applications in the field of virtual try-on technology.

CRediT authorship contribution statement

Tasin Islam: Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Alina Miron:** Validation, Supervision, Methodology. **Xiaohui Liu:** Supervision, Methodology. **Yongmin Li:** Supervision, Methodology.

Declaration of competing interest

The authors declare that we have no conflicts of interest related to the research or writing of the manuscript.

Data availability

I have shared a link to my code and dataset via GitHub.

Acknowledgments

This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/T518116/1.

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, L.S. Davis, Viton: An image-based virtual try-on network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7543–7552.
[2] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, Toward characteristic-preserving image-based virtual try-on network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 589–604.
[3] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, P. Luo, Towards photo-realistic virtual try-on by adaptively generating-preserving image content, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7850–7859.
[4] R. Yu, X. Wang, X. Xie, Vtnfp: An image-based virtual try-on network with body and clothing feature preservation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10511–10520.
[5] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, J. Yin, Towards multi-pose guided virtual try-on network, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 9025–9034, http://dx.doi.org/10.1109/ICCV.2019.00912.
[6] S. Choi, S. Park, M. Lee, J. Choo, VITON-HD: High-resolution virtual try-on via misalignment-aware normalization, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[7] B. Fele, A. Lampe, P. Peer, V. Struc, C-VTON: Context-driven image-based virtual try-on network, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022.
[8] T. Islam, A. Miron, X. Liu, Y. Li, SVTON: Simplified virtual try-on, in: 21st IEEE International Conference on Machine Learning and Applications, ICMLA, 2022.
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[10] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[11] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, 2019, arXiv:1903.07291. URL: http://arxiv.org/abs/1903.07291.
[12] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018, arXiv:1710.10196. URL: http://arxiv.org/abs/1710.10196.
[13] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[14] Y. Jo, J. Park, SC-FEGAN: Face editing generative adversarial network with user's sketch and color, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1745–1753, http://dx.doi.org/10.1109/ICCV.2019.00183.
[15] C.-H. Lee, Z. Liu, L. Wu, P. Luo, MaskGAN: Towards diverse and interactive facial image manipulation, 2019, arXiv:1907.11922. URL: http://arxiv.org/abs/1907.11922.
[16] H. Dong, X. Liang, Y. Zhang, X. Zhang, Z. Xie, B. Wu, Z. Zhang, X. Shen, J. Yin, Fashion editing with adversarial parsing learning, 2019, arXiv:1906.00884. URL: http://arxiv.org/abs/1906.00884.

[17] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[18] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: International Conference on Machine Learning, PMLR, 2017, pp. 2642–2651.

[19] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2018, arXiv:1809.11096. URL: http://arxiv.org/abs/1809.11096.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, 2016, arXiv:1605.05396. URL: http://arxiv.org/abs/1605.05396.

[21] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.

[22] Z. Qi, J. Sun, J. Qian, J. Xu, S. Zhan, PCCM-GAN: Photographic text-to-image generation with pyramid contrastive consistency model, Neurocomputing 449 (2021) 330–341, http://dx.doi.org/10.1016/j.neucom.2021.03.059.

[23] S. Pande, S. Chouhan, R. Sonavane, R. Walambe, G. Ghinea, K. Kotecha, Development and deployment of a generative model-based framework for text to photorealistic image generation, Neurocomputing 463 (2021) http://dx.doi.org/10.1016/j.neucom.2021.08.055.

[24] W. Shen, R. Liu, Learning residual images for face attribute manipulation, 2016, arXiv:1612.05363. URL: http://arxiv.org/abs/1612.05363.

[25] Y. Lei, W. Du, Q. Hu, Face sketch-to-photo transformation with multi-scale self-attention GAN, Neurocomputing 396 (2020) http://dx.doi.org/10.1016/j.neucom.2020.02.024.

[26] L. Liu, H. Zhang, Y. Ji, Q.M. Jonathan Wu, Toward AI fashion design: An attribute-GAN model for clothing match, Neurocomputing 341 (2019) http://dx.doi.org/10.1016/j.neucom.2019.03.011.

[27] J. Liu, X. Song, Z. Chen, J. Ma, MGCM: Multi-modal generative compatibility modeling for clothing matching, Neurocomputing 414 (2020) http://dx.doi.org/10.1016/j.neucom.2020.06.033.

[28] S. Honda, Viton-gan: Virtual try-on image generator trained with adversarial loss, 2019, arXiv preprint arXiv:1911.07926.

[29] N. Pandey, A. Savakis, Poly-GAN: Multi-conditioned GAN for fashion synthesis, Neurocomputing 414 (2020) http://dx.doi.org/10.1016/j.neucom.2020.07.092.

[30] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.

[31] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[33] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.

[34] J. Karras, A. Holynski, T.-C. Wang, I. Kemelmacher-Shlizerman, DreamPose: Fashion image-to-video synthesis via stable diffusion, 2023, arXiv preprint arXiv:2304.06025.

[35] S. Cao, W. Chai, S. Hao, Y. Zhang, H. Chen, G. Wang, DiffFashion: Reference-based fashion design with structure-aware transfer by diffusion models, 2023, arXiv preprint arXiv:2302.06826.

[36] P. Guan, L. Reiss, D.A. Hirshberg, A. Weiss, M.J. Black, Drape: Dressing any person, ACM Trans. Graph. 31 (4) (2012) 1–10.

[37] M. Sekine, K. Sugita, F. Perbet, B. Stenger, M. Nishiyama, Virtual fitting by single-shot body shape estimation, in: Int. Conf. on 3D Body Scanning Technologies, Citeseer, 2014, pp. 406–413.

[38] G. Pons-Moll, S. Pujades, S. Hu, M. Black, ClothCap: Seamless 4D clothing capture and retargeting, ACM Trans. Graph. (Proc. SIGGRAPH) 36 (4) (2017) URL: http://dx.doi.org/10.1145/3072959.3073711. Two first authors contributed equally.

[39] C. Patel, Z. Liao, G. Pons-Moll, TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2020.

[40] M.R. Minar, H. Ahn, Cloth-VTON: Clothing three-dimensional reconstruction for hybrid image-based virtual try-ON, in: Asian Conference on Computer Vision, ACCV, 2020.

[41] F. Zhao, Z. Xie, M. Kampffmeyer, H. Dong, S. Han, T. Zheng, T. Zhang, X. Liang, M3D-VTON: A monocular-to-3D virtual try-on network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 13239–13249.

[42] N. Jetchev, U. Bergmann, The conditional analogy gan: Swapping fashion articles on people images, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2287–2292.

[43] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[44] H.J. Lee, R. Lee, M. Kang, M. Cho, G. Park, LA-VITON: A network for looking-attractive virtual try-on, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE, 2019, pp. 3129–3132, http://dx.doi.org/10.1109/ICCVW.2019.00381, URL: https://ieeexplore.ieee.org/document/9021949/.

[45] P. Lai, N.T. Nguyen, S.-T. Chung, Keypoints-based 2D virtual try-on network system, J. Korea Multimedia Soc. 23 (2020) 186–203.

[46] Z. Xie, X. Zhang, F. Zhao, H. Dong, M.C. Kampffmeyer, H. Yan, X. Liang, WAS-VTON: Warping architecture search for virtual try-on network, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 3350–3359, http://dx.doi.org/10.1145/3474085.3475490.

[47] M.R. Minar, T.T. Tuan, H. Ahn, P. Rosin, Y.-K. Lai, Cp-vton+: Clothing shape and texture preserving image-based virtual try-on, in: CVPR Workshops, 2020.

[48] R.A. Güler, N. Neverova, I. Kokkinos, DensePose: Dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[49] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: CVPR, 2017.

[50] T. Simon, H. Joo, I. Matthews, Y. Sheikh, Hand keypoint detection in single images using multiview bootstrapping, in: CVPR, 2017.

[51] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[52] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023, arXiv preprint arXiv:2303.05499.

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[54] K.P. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, 2012.

[55] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, 2017, arXiv:1711.11585. URL: http://arxiv.org/abs/1711.11585.

[56] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.

[57] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[58] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).

[60] M. Seitzer, PyTorch-fid: FID score for PyTorch, 2020, Version 0.3.0. https://github.com/mseitzer/pytorch-fid.

[61] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.

[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Adv. Neural Inf. Process. Syst. 29 (2016).

[63] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.