Multimedia Tools and Applications (2022) 81:19967–19998
https://doi.org/10.1007/s11042-022-12802-6

You can try without visiting: a comprehensive survey on virtually try-on outfits

Hajer Ghodhbani [1] · Mohamed Neji [1,2] · Imran Razzak [3] · Adel M. Alimi [1,4]

Received: 8 August 2021 / Revised: 25 December 2021 / Accepted: 1 March 2022 /
Published online: 10 March 2022

Abstract

Since the last years and until now, technology has made fast progress for many industries, in particularly, garment industry which aims to follow consumer desires and demands. One of these demands is to fit clothes before purchasing them on-line. Therefore, many research works have been focused on how to develop an intelligent apparel industry to ensure the online shopping experience. Image-based virtual try-on is among the most potential approach of virtual fitting that tries on target clothes into customer's image, therefore, it has received considerable research efforts in the recent years. However, there are several challenges involved in development of virtual try-on that make it difficult to achieve naturally looking virtual outfit such as shape, pose, occlusion, illumination cloth texture, logo and text etc. The aim of this study is to provide a comprehensive and structured overview of extensive research on the advancement of virtual try-on. This review first introduces virtual try-on and its challenges followed by its demand in fashion industry. We summarize state-of-the-art image based virtual try-on for both fashion detection and fashion synthesis as well as their respective advantages, drawbacks, and guidelines for selection of specific try-on model followed by its recent development and successful application. Finally, we conclude the paper with promising directions for future research.

Keywords Virtual try-on · Fashion industry · Fashion detection · Fashion synthesis

✉ Hajer Ghodhbani
hajer.ghodhbani@regim.usf.tn; hajerghodhbani@ieee.org

[1] REsearch Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, 3038 Sfax, Tunisia

[2] National School of Electronics and Telecommunications of Sfax Technopark, BP 1163, CP 3018 Sfax, Tunisia

[3] Advanced Analytics Institute, University of Technology, Sydney, Australia

[4] Department of Electrical and Electronic Engineering Science, Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg, South Africa

1 Introduction

In the last few years and especially during COVID-19 pandemic, online shopping for clothes has become a common practice among millions of people around the world. It shows a great progress and become a habitual activity for many consumers. This progress is conducted by the implementation of virtual try-on technology that enables the customer to visualize the produce on themselves and see how certain the products look on them before purchasing. In 2012, Converse was the first brand that used virtual iPhone try-on by allowing their clients to use phone cameras to see how shoes looked on them, and post photos on social media as well as make online purchases [92]. This technology applies very well to shoes, apparel, accessories, jewelry as well as make-up, where consumers long for a sense of "touch and feel" and they have total freedom regarding decision making, trying, and choosing products at their own pace, without feeling the pressure to make a purchase.

Approximately, 40% customers are willing to spend more if they can try the product through virtual reality [92], due to the fact that try-on experience makes it much easy to explore the many other options as well as customize or personalize the products according to their body shape. For this reason, online shopping for clothes has earned its place deservedly. Popular fashion brands including L'Oréal, Baume, Sephora, Adidas, Nike and Snap are opting try-on technology in order to improve the connectivity with customer and gain a competitive advantage in the market. With statistical proof, the global fashion apparel has exceeded 3 trillion US dollars, in currently year, and presents 2 % of the world's Gross Domestic Product (GDP). In 2020, a revenue of 718 billion US dollars area attained in the fashion sector and an expectation to reach a growth of more than 8.4% for coming years [73].

During *COVID19* pandemic lockdown, most of the business went into kind of a crisis mode and not only big brands, but also small retailers are thinking how they can survive [81]. Taking our time in shops will be difficult in a post-Covid-19 world as a result, online shopping is ingrained significantly in our daily as trade become more and more like shopping in person thanks to the efforts of businesses to add new features and services with the intent of providing their customers the same support and comfort that they would have during an in-person shopping experience. This goal has been achieved by using the computer technology to develop virtual try on applications that assist the fit of garment product to make consumers know how cloths look on themselves, how both the top and bottom matches together, and how the size of clothes fits to them.

Therefore, Online shopping would give more information and availability of all kinds of products to encourage fashion trailers to make the best investment by exploring new sales methods and optimizing the technological process of purchasing clothes like virtual fitting system. These solutions draw a new picture of online shopping experience and bring it to a high level of reality and comfort. One of these improvements is to allow consumers buying clothes after trying them like in real shops because the existing systems cannot provide the possibility for users to try-on various fashion items according to their desires. Thus, fashion brands need to better satisfy customer preferences and engage them with the personalized shopping experience to make more informed and confident purchase decisions. In addition, allowing consumers to virtually try on clothes will not only enhance their shopping experience, but also increase the fashion industries sales because these solutions can play an important role to reduce return rates and improve customer satisfaction.

Instead of using current graphics tools that fail to meet the increasing demands for personalized visual content manipulation, there are many proposed algorithms to address

swapping clothes by using recent advances in computer vision tasks like fashion detection, fashion analysis or fashion synthesis. These solutions require considerable effort from researchers to perform the task of changing clothes with preserving details and identities. However, using current image editing technology e.g., Adobe Photoshop or Adobe illustrator cannot give a realistic result due to many challenges of changing clothing in 2D images, such as the deformation of the clothes, different poses, and different textures. Recent studies adopted deep-learning-based methods to encounter these problems and achieve more accurate results.

In the literature, a little number of fashion surveys are proposed [6, 42, 53, 71]. Recently, a summary on intelligent clothing analysis was made by Liu et al. [42]. In addition, Song and Mei [71] presented on overview of fashion development with the emergence with multimedia. Then, a general survey designs the whole picture of intelligent fashion without taken a specific issue [6]. Another survey [53] is proposed to present AI applications in the fashion apparel industry, but it is based only on the structured task-based multi-label classification works. Next and due to the rapid development of computer vision, many tasks are appeared within intelligent fashion, hence, many related works must be updated. In this direction, this survey aims to conduct a comprehensive literature review of deep learning methods applied in the fashion industry by citing research works published in the last years and mentioning their relationship to the early studies. Our contribution consists in responding to the following research questions:

- RQ1. What is the impact of adoption of *Artificial Intelligence* (AI) in the garment industry?
- RQ2. How virtual try on system are developed?
- RQ3. What are the common problems that need solving to ensure an intelligent fashion shopping?

In this paper, different sections are structured as follow: Section 2 outlines the research framework adopted to realize this research review. Section 3 is dedicated to virtual try-on applications, and divided into two parts, the first one presents the fashion detection tasks including fashion parsing, fashion synthesis, and landmark detection. The second one illustrates the works for fashion synthesis containing style transfer, pose transfer, and clothing simulation. Section 4 provides an overview of fashion benchmark datasets. Section 5 presents the performance of popular works on different tasks. Section 6 shows related applications and future directions. Finally, a conclusion is given in Section 7.

2 Research framework

In this study, a *Systematic Literature Review* (SLR) [29] is chosen to focus on research works related to virtual fitting system based on 2D images with deep learning methods and applied in the fashion industry. The SLR methodology adopted is shown in Fig. 1. The review process commenced with collecting and preparing data from scientific databases. Subsequently, articles were selected in different phases according to our research framework, and we have selected more than 100 articles from both journals and conference.

Articles in each tasks of the topic at hand such as fashion detection [10, 13, 14, 28, 30–32, 34, 35, 37–41, 43, 44, 52, 55–57, 64, 76–79, 83, 85, 93–95, 102, 103] and fashion synthesis

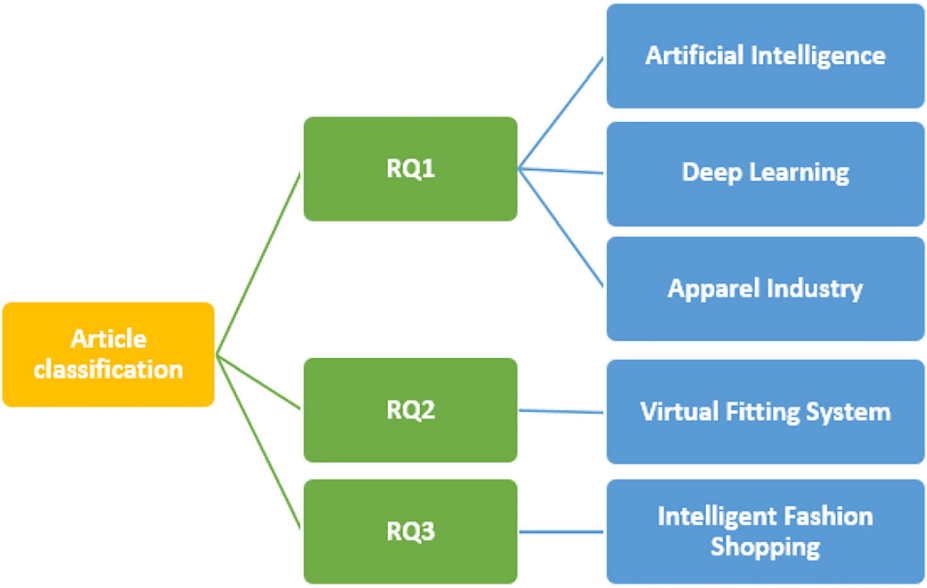

Fig. 1 Article Classification based on Research Questions

[3, 7, 9, 12, 15–17, 21, 22, 26, 27, 33, 48–51, 58, 59, 61, 62, 65–67, 69, 70, 72, 74, 84, 86, 89, 90, 96–98, 100, 101, 104, 106, 107], were retrieved from popular databases and engines such as Google scholar[1] and Research Gate[2] . Then, a screening process is used to select specific articles to address the research questions mentioned in previous section. Then, a categorization of research articles must be done according to the main steps used to develop image-based virtual fitting system with deep learning methods. After categorization, there is the process of information extraction and classification of the selected articles based on the key terms of research topic to address our research questions.

As shown in Fig. 1 that presented the article classification according to the research questions, RQ1 is focused on understanding the overall trend of AI in the Fashion industry. Hence, the focus of the screening process was limited to those articles discussing the implementation and execution of AI techniques to improve online shopping. RQ2 aimed at identifying the various stages on virtual fitting framework where the AI method was employed. RQ3 aims to understand the extent of online shopping problems which being a focus of research studies. These keys modules were considered during information extraction from research articles.

3 Fashion virtual try-on

In recent years, advanced machine learning approaches have been successfully applied to various fashion-based problems. The topics of fashion research in the literature of image-based garment transfer are summarized in Fig. 2. One of the branches in fashion research is fashion detection, which aims to label each pixel in the scene (i.e., fashion parsing, landmark detection,

[1] https://scholar.google.com/
[2] https://www.researchgate.net

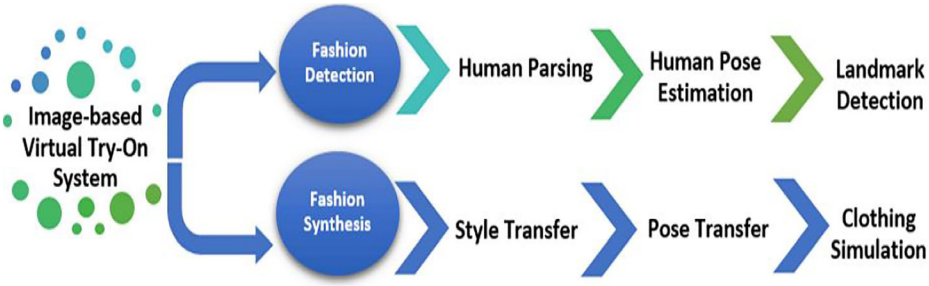

Fig. 2 Classification of based approaches for image-based virtual try-on System

and pose estimation), supported by fashion synthesis, which lead us a step closer to a fashion intelligent assistant.

3.1 Fashion detection

Fashion detection is an essential task for virtual try-on task, it consists of detecting the human body part to predict the region of clothing synthesis. To apply this task in virtual try-on systems, three aspects must be presented: Fashion parsing, Human Pose Estimation and Fashion landmark detection.

3.1.1 Fashion parsing

Fashion parsing or in other words human parsing with clothes classes, is a specific form of semantic segmentation. This task refers to generate pixel-level labels on the image which are based on the clothing items like hair, head, upper clothes, pants, etc. It is a very challenging problem since the number of garment types, the variation in configuration and appearance are enormous. In Fig. 3, we present an example of fashion parsing results generated by the work of Ji et al. [28].

In fashion domain, largest number of potential applications have been devoted to various tasks and particularly to human parsing [10, 39, 41, 93, 94]. At the beginning, Yamaguchi et al. [93] proposed a model by merging the fashion parsing and the human pose estimation. Then, they proposed clothes parsing with a retrieval-based approach [94] to resolve the constrained parsing problem. After that, a weak supervision approach for fashion parsing is presented by Liu et al. [41] who resort to label images with color-category labels instead of

Fig. 3 Examples of fashion parsing based on semantic segmentation [28]

pixel-level. These works conduct results far from being perfect because between pose estimation and clothing parsing there is no consistent targets. Many restrictions are presented with these hand-crafted methods because they need to be developed carefully.

To deal with these issues, many methods based on *Convolutional Neural Network* (CNN) are proposed such as the deep human parsing-based work of Liang et al. [10] which resorts to an active template regression for semantic labeling. Then and with the aim to improve the generated results of their human parsing work, a *Contextualized CNN* (Co-CNN) [39] is designed to take the context of cross-layer, global image-level, and local super-pixel. In parallel, they proposed a deep human parsing with *Active Template Regression* (ATR) [39] to ensure the human parsing task by decomposing an image of person into semantic fashion and body regions. In 2018, Liao et al. [40] built a *Matching CNN* (M-CNN) network to solve the issues of parametric and non-parametric CNN-based methods. In the same year, Gong et al. [13] implemented an important self-supervised method under the name of *Look Into Person* (LIP) to eschew the necessity of labeling the human joints in model training (Fig. 4). With the intent to ameliorate their previous work [13], the same authors proposed a *JPPNet* network [102] to treat both the human parsing and human pose estimation task.

Different from the previous mentioned works that only concentrated on single person parsing task, there are many others works [14, 64, 85, 103] which focus on treating the scenario with multiple views of persons. Zhao et al. [103] designed a deep *Nested Adversarial Network* (NAN) to understand humans in crowed scenes. Gong et al. [14] proposed the first attempt to explore a detection-free *Part Grouping Network* (PGN) used for the semantic part segmentation for assigning each pixel as a human part and the instance-aware edge detection to group semantic parts into distinct person instances. With the aim to manage, simultaneously, single and multiple human parsing, Ruan et al. [64] developed a *Context Embedding with Edge Perceiving* (CE2P) framework. Recently, hierarchical graph is used for human parsing tasks to improve parsing performance such as the work of Wang et al. [85] that considered the human body as a hierarchy of multi-level semantic parts to capture the human parsing information.

Fig. 4 Annotation examples for LIP [13] with appearance variability and different views

3.1.2 Human pose estimation

Advanced in computer vision are realized by many tasks especially with deep learning-based approaches such as *Human Pose Estimation* (HPE) that is applied in many fields like fashion fitting to get specific postures from human body by joints' localization. To overcome the challenges appeared with the task of HPE, many research efforts have been applied to the related fields. We present, in this section, recent research in HPE methods based on 2D images which are classified into two groups: single person pose estimation and multi-person pose estimation.

Single-person human pose estimation *Single-person Human Pose Estimation* (HPE) is related to the task of localizing human skeletal keypoints from an image or video data. In the following Figure (Fig. 5), we present results of Single-person HPE obtained from the DeepPose [79] trained on *Leeds Sports Pose* (LSP) dataset [30]. According to the different structures of HPE task, methods based on CNN can take different aspects such as regression methods and detection methods.

Regression-based methods produced joint coordinates by learning mapping directly from image [79]. The early deep learning-based network adopted by many researchers was *AlexNet* [31] due to its simple architecture. Toshev et al. [79] applied this network to learn joint coordinates from full images, and Li et al. [35] employed it as a multi-task framework to predict the joint coordinate from full image. However, Detection-based methods treat the body parts as detection targets based on two main representations: image patches and heatmaps of joint locations. The methods related to this category are intended to predict approximate locations of body parts [32] or joints [52].

Fig. 5 Example of human pose estimation from DeepPose [79] on the LSP Dataset [30]

Previous works attempt to adjust detected body parts into body models, but there are other recent works [57, 76–78] which aim to encode human body structure information into networks. Tang et al. [77] proposed a hierarchical representation of body parts, then, they extended their work [76] to learn specific features of part group. Then, they committed to improve the network structure by proposing a densely connected U-nets and efficient usage of memory [78]. For Peng et al. [57], they exploited data augmentation to avoid the need of more data during training.

Multi-person human pose estimation The second category of HPE methods is the *multi-person HPE* which aims to handle detection and localization tasks. It can be divided, according to its different level, into top-down methods and bottom-up methods. Top-down methods used bounding box and estimators of single-person pose to detect person from image and predict human poses. The bottom-up methods put into skeletons the prediction of 2D joints of persons in the image. Figure 6 shows examples of results from the work of Li et al. [38] that belongs to the bottom-up methods.

A combination of existing detection networks and single HPE networks used to implement the *Top-down HPE* methods [55, 56] that achieved state-of-the-art performance in almost benchmark datasets while the processing speed is dependent to the number of detected people. For *bottom-up HPE* methods, the main components include body joint detection and joint candidate grouping. The two components are handled separately for most algorithms. The bottom-up methods-based works realized perfect performance expect some conditions like human occlusions or complex background.

3.1.3 Fashion landmarks detection

Fashion landmark detection is an important task in fashion analysis, it aims to predict clothes keypoints which are very essential for fashion images understanding by getting discriminative representation. The local regions of fashion landmarks give more significant variances since the clothes are more complicated than human body joints. Figure 7 shows results generated by the fashion landmark detection approach.

Fig. 6 Example of multi-person HPE [38]

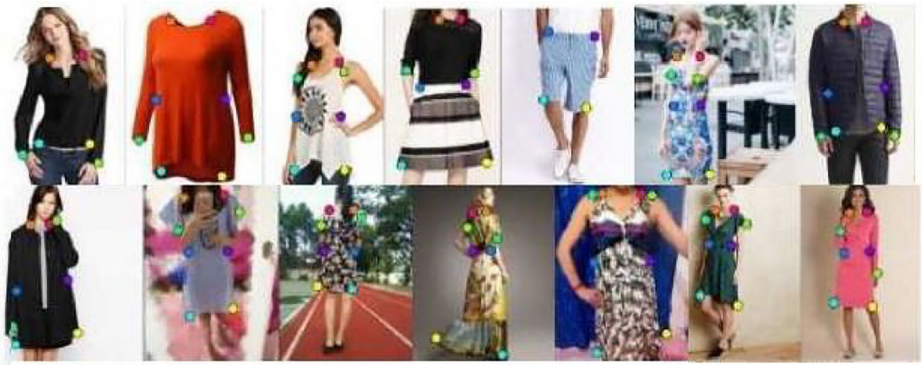

Fig. 7 Example of results from Fashion Landmark Detection approach [37]. First row illustrates the results on *DeepFashion-C* [43], second row presents results on *Fashion Landmark Dataset* (FLD) dataset [44]

For the first time, Liu et al. [43] presented fashion landmark concept and, in parallel, they proposed a deep model called *FashionNet* [43] applied on predicted clothing landmarks. Then, they proposed a deep fashion alignment framework [44] based on CNN. This Framework is trained on different datasets and evaluated on two fashion applications, clothing attribute prediction and clothes retrieval. Another regression model proposed by Yan et al. [95] used to relax constraint of clothing bounding box due to its difficult application. A more recent work [83] mentioned that optimization on regression model is hard, so, they proposed to directly predict a confidence map of positional distributions for each landmark. Lee et al. [34] resorted to contextual knowledge to achieve perfect performance on landmark prediction.

3.2 Fashion synthesis

Fashion synthesis is the task for generating new style across images and being able to imagine what that person would look in a different clothing style by synthesizing a realistic-looking image. In the following, we review existing methods for addressing the problem of generating images of people in clothing by focusing on style transfer, pose transformation, and physical simulation.

3.2.1 Style transfer

In fashion synthesis task, style transfer is an important step that aims to transfer the style between images. It can be applied in various kinds of image especially facial image and garment image. CNN- based methods applied on this task exploit the feature extraction to obtain style information from image. Isola et al. [26] proposed the style transfer work, pix2pix, which is a general solution for style transfer. For specific goal, based on a texture patch, the work of Xian et al. [90] transferred the input image or sketch to the corresponding texture (Fig. 8).

Driven by increasing power of deep generative models, popular virtual try-on applications have appeared [12, 16, 27, 50, 62, 84, 98]. Han et al. [16] proposed a two-stage pipeline called *VIrtual Try-On Network* (VITON) to transfer desired in-shop clothing onto a consumer's body by allowing the first stage to warp the input item to the desired deformation style and enabling the second stage to align the warped clothes to the consumer's image. Many approaches

Fig. 8 Examples of image style transfer by TextureGAN [90]

following this pipeline have been proposed with more competitive performance such as *CP-VTON* [84] and *CP-VTON+* [50], which adopt a *thinplate spline (TPS) transformation* learnable [9] based on Convolutional neural network architecture for geometric matching to align explicitly input clothing with body shape. All these works are powered by the use of TPS, thus, in the following Figure (Fig. 9) we present its application on VITON architecture [16].

However, results of these methods are limited in different cases (Fig. 10). One of the main causes resulting in such failed cases comes from warping stage which can be based on inaccurate clothing mask and warped target clothes image used to calculate TPS transformations, thus, its dependence on the shape context cannot be able to perform perfectly on the warping task, and this is the case on VITON [16]. Geometric matching module adopted in CP-VTON [84] utilizes grid points as control points for calculating TPS transformation to reduce image distortions in warped images, which can be seen Fig. 10.

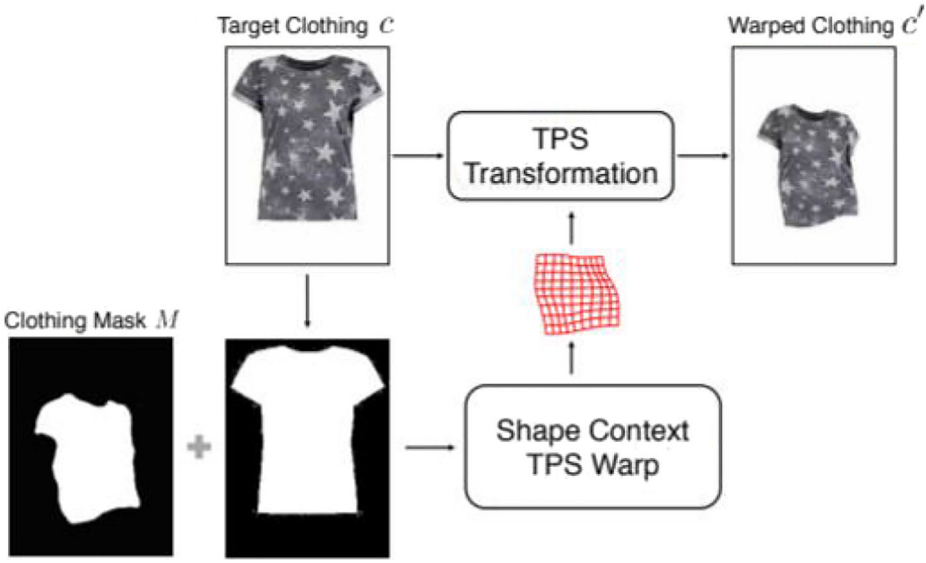

Fig. 9 Example of Warping a clothing image proposed by VITON [16]: Given the target clothing image and a clothing mask, the shape context matching is used to estimate the *TPS transformation* and generate a warped clothing image

Fig. 10 Results from the CP-VTON [84], CP-VTON+ [50] ACGPN [98] and CIT [62]

Then, a second-order difference constraint on Thin-Plate Spline (TPS) is proposed to produce geometric matching yet character retentive clothing images with the ACGPN network (Adaptive *Content Generating and Preserving Network*) [98]. This method characterized by the existence of an additional semantic generation module used to generate a semantic alignment of spatial layout. It presents important results but with no consideration of the latent global long-range interactive correlation between the person representation and the in-shop clothing. Despite the perfect results generated with these methods [16, 50, 84, 98], there are still a need to obtain more realistic image with no artifacts especially, when there are occlusions or large variations. For these reason a two-stage transformer pipeline is proposed under the name of *Cloth Interactive Transformer* (CIT) [62] to model the latent global relation in both stages (Fig. 10).

More recently, other works based on in-shop clothes items [12, 27] are proposed to deal with this same problem with the difference that most of the above methods [16, 50, 62, 84, 98] were relied on human segmentation of different body parts to enable the learning procedure of virtual try-on. However, ensure the human parsing task with high performance manner required important training of the corresponding models, for the reason that the poor quality of segmentation guide to highly-unrealistic generated images. To reduce this issue due to the

dependence to the masks as an inputs for the models, a *Warping U-Net for a Virtual Try-On* (WUTON) [27] is appeared as the first parser-free network without using of human segmentation for virtual try-on, as shown Fig. 11. Then, another work called *Parser Free Appearance Flow Network* (PF-AFN) [12] is proposed in the same context, to produce highly photo-realistic try-on images without human parsing (Fig. 11).

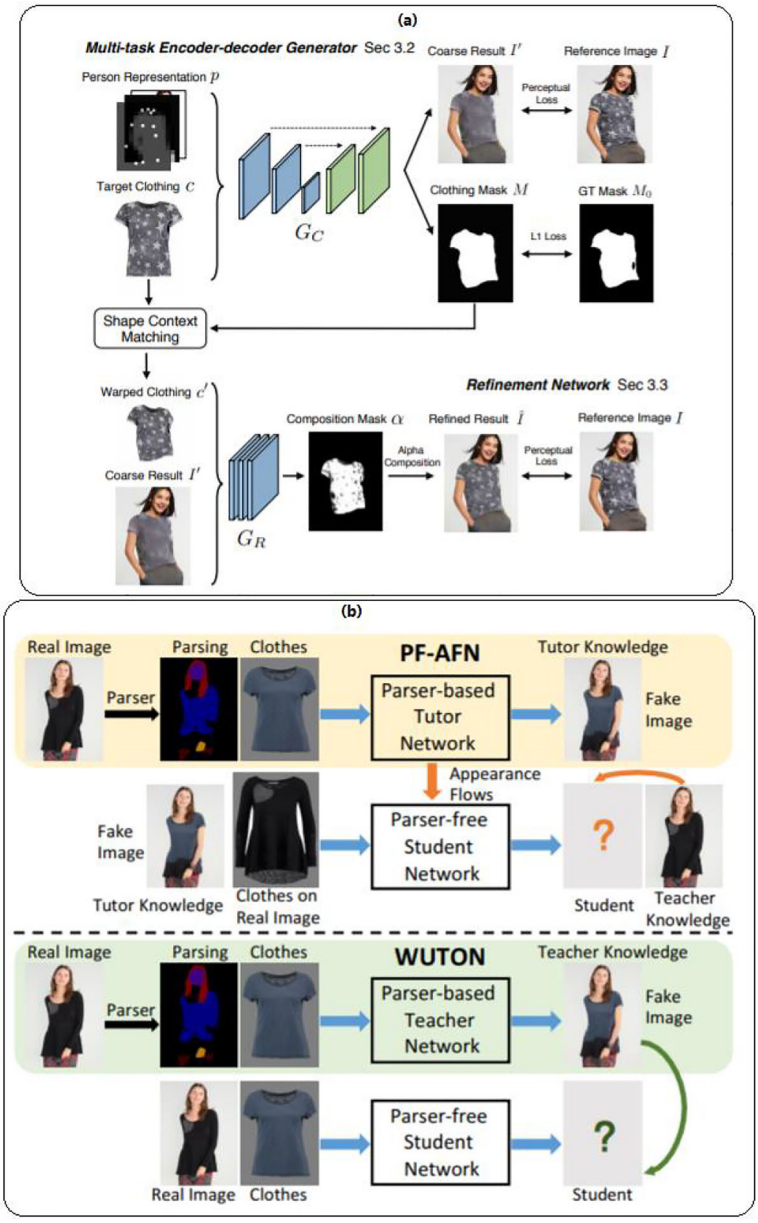

Fig. 11 Different architectures for warped Module: **a** based on segmentation mask from VITON [16], **b** without human segmentation from WUTON [27] and PF-AFN [12]

The previous works required in-shop clothing image for virtual try-on, but other existing models like *FashionGAN* [7] and *M2E-TON* [89] resolved this task basing on text description and model image by giving an input image and a sentence describing a different outfit. First, a GAN generates the segmentation map according to the description and then, another GAN ensures rendering of the output image by the segmentation map. Other works attempts to resolve the problem with arbitrary poses such as Fit-Me [21] which was the first work building virtual try-on dealing with this challenge. Then, *FashionOn* [22] applied the semantic segmentation to present more realistic results. Then, SwapNet [61] was one of the first works that expose the challenge of transferring all the clothing from one person's image onto the pose of another target person by operating in image-space. This is done by generating a mutually exclusive segmentation mask of the desired clothing into the desired pose.

Another virtual try-on network called *Vtnfp* [100] proposed a similar strategy to synthesize photo-realistic images given the images of clothed person and target clothing item. Zheng et al. [106] presented an architecture to try-on clothing with arbitrary poses by using the body shape mask prediction for pose transformation. Based in the same design strategy, Han et al. [17] proposed *ClothFlow* which is an appearance-flow-based generative model allowing the transfer of different appearances and synthesize clothed persons for posed-guided person image generation and virtual try-on.

Recently, various works [48, 51, 66, 67, 74, 96] address challenging problems of garment interchange between person's pictures with preserving the identity in the source and target images by developing an image-based virtual try-on network. Feng et al. [74] resolve the problems of visual details and the missing of body parts by maintain the structural between the generated image and the reference image. *Outfit-VITON* [74] allows the visualization of a cohesive outfit from multiple images of clothed human models, while fitting the outfit to the body shape and pose of the query person. Sarkar et al. [66, 67] achieve high-quality try-on results by aligning the given human images with a 3D mesh model via *DensePose* [79], estimating a UV texture map corresponding to the desired garments, and rendering this texture onto the desired pose (Fig. 12).

In the current year, conditioning model is adopted by *Dressing in Order* (DiOr) [67] to support 2D pose transfer, virtual try-on, and several fashion editing tasks, and a *Complementary Transferring Network* (CT-Net) [96] is published to adaptively model different levels of geometric changes and transfer outfits between different people. Despite this diversity of these systems, the ability to preserve details or to present, correctly, the shape and the texture is still a challenging task.

3.2.2 Pose transformation

Pose transformation is a crucial task for fashion synthesis, it takes an input image of person and a target pose to generate images of this persons in different poses with the preserving of original identity (Fig. 13). To deal with this task, many works are proposed. Firstly, a pose guided person image generation PG2 [48] is presented with a two-stage adversarial network to achieve an early attempt on the challenging task of transferring a person to different poses by generating both poses and appearance simultaneously and using affine transform to keep textures in the generated results.

The work of Siarohin et al. [70] used a deformable GAN to generate images of person according to a target pose which allowed the extraction of the articulated object pose by

Fig. 12 Garment transfer results generated by the work of Sarkar et al. [67]

resorting to a keypoint detector. Guha et al. [3] address the problem of human pose synthesis with a modular generative neural network that synthesizes unseen poses by using four modules consisting of image segmentation, spatial transformation, foreground synthesis, and background synthesis. Si et al. [69] introduced a multi-stage pose-guided image synthesis framework which divided the network into three stages for pose transform in a novel 2D view, foreground synthesis, and background synthesis. Pumarola et al. [59] treat the limitation of data presented by the above research studies by borrowing the idea from [107] and leveraging cycle consistency.

Last year, the work of Song et al. [72] presented a solution for this limitation by proposing a novel approach which consisted of a decomposition of the hard mapping into semantic parsing transformation and appearance generation sub-tasks to improve the appearance performance. In addition, The generative model, *Attribute-decomposed GAN* (ADGAN) [49], produce realistic images with desired human attributes. The idea behind this work is to embed human attributes into the latent space as independent codes and then ensure the control of attributes via mixing and interpolation operations in explicit style representations.

Fig. 13 Examples of pose transformation results generated by PG2 work of Liqian Ma, et al. [48] from DeepFashion dataset [43] (**a**) and Market-1501 dataset [104] (**b**)

3.2.3 Clothing simulation

For more improvement of fashion synthesis performance, the use of clothing simulation is essential. The works mentioned in the previous section are about the 2D domain where clothing deformation is not considered to generate realistic appearance. This important task presented many challenges like the need of creating more realistic results in real-time running with the treatment of more complex garments.

Computer graphics tools was the traditional way for realistic clothes generation models [15, 58, 97]. Yang et al. [97] proposed an approach to recover a 3D mesh of garment with 2D physical deformations by capturing the global shape and geometry of the clothing and extracting important details of cloth from a single-view image. The recovered clothing can be addressed to other human bodies in variety of poses for virtual fitting task. Guan et al. [15] aimed to dress people in a different variation and pose, and clothing types with an automatic process. Thus, they proposed *DRAPE* (*DRessing Any PErson*) model to simulate clothes deformation with varying shape and pose (Fig. 14). Then, *ClothCap* [58] is proposed as a multi-part 3D model to simulate clothing deformation of people in motion from 4D scans. This model ensures the virtual try-on task by capturing a clothed person in motion, extracting their clothing, and retargeting the clothing to new body shapes.

The simulation of the physical deformation has important role to ensure more performance for fashion synthesis due to the generation of dynamic details, clothing-body interactions, and the 3D information. Wang et al. [86] interested on this task and proposed a semi-automatic method to learn the intrinsic physical properties with different postures to generate garment

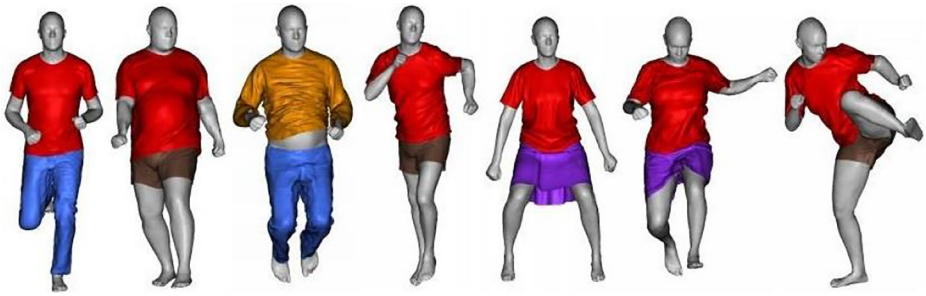

Fig. 14 Example of clothing simulation results obtained with DRAPE model [15]

animation which are shown in Fig. 15. The proposed model encoded the main information of the clothing shape and learned to reconstruct garment shape with physical properties by considering the intrinsic garment and the body motion.

To improve more realistic view to the garment on human body, Lahner et al. [33] proposed framework consisting of two modules. The first module aiming to recover shape deformations from 3D data of clothed persons in motion. The second module is a *conditional Generative Adversarial Network* (cGAN) that allowing to ensure realism and temporal consistency and lead the high-resolution details of clothing deformation sequences. Then, Santesteban et al. [65] proposed a two-level learning-based clothing animation method for virtual try-on simulation to ensure performance of the physical simulation with non-linear deformations of clothing. In addition, Yu et al. [101] proposed a physic-based simulation with performance capture called *SimulCap*. This model ensures tracking of people and clothing using a multi-layer surface. So, it combines the benefits of capture and physical simulation. The contribution of this work consisting of: (1) a multi-layer representation of garments and body including the undressed body surface and separate clothing meshes, (2) a physics-based performance capture procedure using body and cloth tracking for physical simulation and clothing-body interactions.

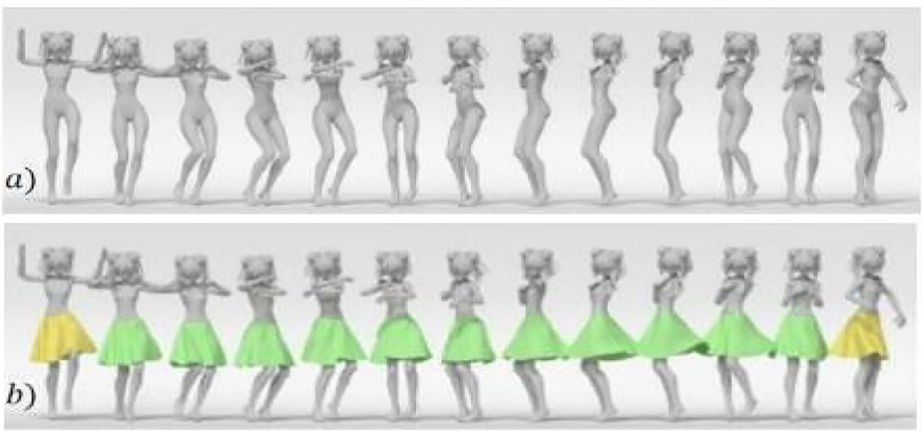

Fig. 15 Examples of physical simulation from the work of Wang et al. [86]

4 Benchmark datasets

Recent progress in virtual try-on systems have been driven by the building of fashion datasets, despite that, it is difficult to develop a universal dataset to evaluate the whole methods of virtual try-on because there are large variations in different tasks. Therefore, some researchers resort to create datasets to evaluate their proposed methods, this diversity makes the comparison on different algorithms very difficult. Datasets, also, bring more challenges and complexity through their expansion and improvement. This section discusses the popular publicly available datasets for virtual try-on tasks and their characteristics. Large number of benchmark datasets proposed to study fashion applications such as virtual try-on systems are summarized in Table 1.

As summarized in Table 1, for each task there are specific datasets with according setting. Market-1501 [104] and Deep-Fashion [43] are the most popular datasets for virtual try-on. FLD [44] is the most used dataset for fashion landmark detection. Several datasets were built to treat the fashion parsing task such as LIP dataset [13]. Datasets for physical simulation are different from other fashion tasks since the physical simulation is more related to computer graphics than computer vision. Dataset can be categorized into different types according to real data and created data especially when we are dealing with fashion physical simulation which interested on clothing-body interactions.

Despite the progress on 2D image-based fashion datasets like DeepFashion [43], DeepFashion2 [11] and FashionAI [109], the building of datasets basing on 3D clothing is almost rare or not sufficient for training like the digital wardrobe released by MG-Cloth [4]. Recently, Heming et al. [108] develop a comprehensive dataset named DeepFashion3D which is richly annotated and covers a much larger variations of garment styles.

5 Performance assessment

In image processing, measuring the perceptual assessments of generated results is an important step to validate research works. Therefore, there is an emerging demand for quantitative performance evaluation in image-based garment transfer, which is caused by the requirement to objectively judge the quality of virtual fitting systems to facilitate comparability of the various existing approaches and to measure their improvements.

5.1 Image quality assessment (IQA)

The measure of performance of computer vision tasks is ensured by image quality assessment methods which divided into objective or subjective methods. The last one is based on the perception of humans to evaluate the realistic appearance of generated images. With each year, the number of proposed IQA algorithms are progressively growing, by proposing new one or extending existing IQA algorithms. In this section, we present the most popular IQA algorithms used to evaluate tasks of image-based garment transfer.

5.2 IQA for fashion detection

For clothing fitting based on images, the fashion attributes must be first detected to predict the clothing style. Most works on clothing localization show validate results by using different metrics on different tasks such as landmark detection, pose estimation and human parsing.

Table 1 Summary of the benchmark datasets for fashion tasks

Task	Dataset	Number of photos	Description	Publish time
Virtual Try-On	LookBook [8]	84,748	Composed by 9,732 top product images and 75,016 fashion model images	2016
	DeepFashion [43]	78,979	Selected from the In-shop Clothes Benchmark and associated with several sentences as captions and a segmentation map.	2016
	VITON [16]	32,506	Contained around 19,000 frontal-view woman and top clothing image pairs, yielding 16,253 pairs	2018
	FashionTryOn [106]	28,714	Comprising 28, 714 clothing person-person triplets with each consisting of a clothing item image and two model images in different poses.	2019
	FashionOn [22]	22,566	Pairs of person image wearing the same clothes in different poses.	2019
Fashion Parsing	Fashionista [93]	158,235	Outfit information in the form of tags, comments, and links	2012
	Paper Doll [94]	339,797	Annotated with metadata tags denoting characteristics, e.g., color, style, occasion, clothing type, brand	2013
	Chictopia10k [36]	10,000	Contains real-world annotated images in the wild with arbitrary postures, views and backgrounds	2015
	LIP [13]	50,462	■ Focus on semantic understanding of person and contains images with elaborated pixel-wise annotations with 19 semantic human part labels and 2D human poses with 16 key points. ■ Images collected from real-world scenarios contain human appearing with challenging poses and views, occlusions, and various appearances.	2017
	MHP v1.0 [105]	4,980	■ Instance-aware setting with fine-grained pixel--level annotations works with 7 body parts and 11 clothes categories.	2017
	v2.0 [85]	25,403	■ Annotated images with 58 fine-grained semantic categories: 11 body parts and 47 clothes categories ■ Captured images in real-world scenes from various viewpoints, poses, occlusion, interaction, and background	2018
	Crowd Instance-level Human Parsing (CIHP) [103]	38,280	■ Multi-person images ■ Pixel-wise annotations in instance-level	2018
	ModaNet [18]	55,176	Annotated with pixel-level labels, bounding boxes, and polygons	2018
	DeepFashion2 [109]	491,000	■ Diverse images of 13 popular clothing categories from both commercial shopping stores and consumers. ■ Labeled with scale, occlusion, zoom-in, viewpoint, and category, style, bounding box, dense landmarks and per-pixel mask.	2019
	Fashionpedia [24]	48,000	Containing 294 fine-grained attributes with high resolution (1710 × 2151)	2020
	RichWear [1]	322,198	Street fashion dataset containing various text labels for fashion analysis. The images are collected from an Asian social network site, focuses on street styles in Japan and other Asian areas.	2021
	DeepFashion-C [43]	289,222		2016

Table 1 (continued)

Task	Dataset	Number of photos	Description	Publish time
Fashion landmark detection			Annotated with clothing bounding box, pose variation type, landmark visibility, clothing type, category, and attributes	
	Fashion Landmark Dataset (FLD) [44]	123,016	Annotated with clothing type, pose variation type, landmark visibility, clothing bounding box, and human body joint	2016
	Unconstrained Landmark Database (ULD) [95]	30,000	■ Collected from fashion blogs, forums and the consumer-to shop retrieval benchmark of DeepFashion [43] ■ Contains substantial foreground scatters and background clutters	2017
	DeepFashion2 [109]	491,000	DeepFashion2 used in diverse tasks like fashion parsing, clothes detection, pose estimation, segmentation, and retrieval.	2019
Human Pose Estimation	MPII Human pose [60]	2.5104	■ Data are from YouTube videos. It covers 410 human activities, and each image is provided with activity label	2014
	MSCOCO [88]	328,000	■ Data are from Internet. It used for diverse activities.	2014
	AI Challenger [2]	300,000	■ Data are crawled from Internet. ■ Provide three sub-datasets for human keypoint detection, attribute based zero-shot recognition and image Chinese captioning.	2017
	PoseTrack [25]	550 video sequences	■ Focusses on 3 aspects: (1) single-frame multi--person pose estimation. (2) Multi-person pose estimation in videos. (3) Multi-person articulated tracking.	2017
Pose Transfer	Human3.6M [87]	3.6M	■ Containing 3.6 million different 3D articulated poses captured from a set of men and women actors. ■ provides synchronized 2D and 3D data (including time of flight, high quality image and motion capture data), accurate 3D human models of the actors, and mixed reality settings	2014
	Market-1501 [70]	32,668	■ Contains over 32,000 annotated boxes, plus a distractor set of over 500K images produced using the Deformable Part Model (DPM) as pedestrian detector.	2015
	DeepFashion [43]	52,712	In-shop Clothes Retrieval Benchmark DeepFashion is used for pose transfer	2016
	SMPL-NPT [5]	24,000	Contains 24,000 synthesized body meshes and used for 3D Pose Transfer	2020
	SMG-3D [54]	8,000	Contains 8,000 pairs of naturally plausible body meshes of 40 identities and 200 poses, 35 identities and 180 poses are used as the training set	2021
Clothing Simulation	MG-Cloth [108]	356 scans	Contains 3D scans of person with different body shapes, poses and clothes.	2019
	DeepFashion3D [99]	2,078 models	Contains 3D garment models with 10 different clothing categories and 563 garment instances	2020
	AFRIFASHION1600 [82]	1600	African fashion dataset curated to improve visibility, inclusion and familiarity of African fashion in computer vision tasks	2021

5.2.1 Fashion parsing

In fashion Parsing, various metrics are used to evaluate proposed approaches on different datasets such as Fashionista [93] and LIP [13] and in terms of average Pixel Accuracy (aPA), mean Average Garment Recall (mAGR), Intersection over Union (IoU), mean accuracy, average precision, average recall, average F-1 score over pixels and foreground accuracy. Table 2 report some quantitative results measured by these metrics. Most of the parsing methods are evaluated on Fashionista dataset [93] in terms of accuracy, average precision, average recall and average F-1 score over pixels. In addition, There are objective comparisons for virtual try-on, in terms of inception score (IS) [82] or structural similarity (SSIM) [19].

IS is used to evaluate the synthesis quality of images quantitatively. SSIM is utilized to measure the similarity between input and output images ranging from zero (dissimilarity) to one (similarity). Further, SSIM is used also for pose transfer to compare the luminance, contrast, and structure information in images to evaluate many state-of-the-art methods. Table 3 shows evaluation metrics including SSIM, IS, masked version SSIM (mask-SSIM), masked version of IS (mask-IS) and Detection Score (DS) [70] applied on Market-1501 dataset [104] and DeepFashion dataset [43].

5.2.2 Human pose estimation

Research in HPE has made significant progress during the last years which conducted to the appearance of different work that needed to be evaluated with different metrics to measure the performance of human pose estimation models. The most known metrics in this field are Percentage of Correct Parts (PCP), Percentage of Correct Keypoints (PCK) and Average Precision (AP) which can applied in different datasets.

Table 2 Performance comparisons of fashion parsing methods (in %) [28]

Method	Dataaset	Evaluation Metrics							
		mIOU	aPA	mAGR	Acc.	Fg.acc.	Avg.prec.	Avg.recall	AVG.F-1
Yamaguchi et al., [93]	**ATR** [39]	–	–	–	88.96	62.18	52.75	49.43	44.76
Liang et al., [10]		–	–	–	91.11	71.04	71.69	60.5	64.38
Co-CNN [39]		–	–	–	**96.02**	**83.57**	**84.95**	**77.66**	**80.14**
Yamaguchi et al., [93]	**Fashionista** [93]	–	–	–	89.98	65.66	54.87	51.16	46.80
Liang et al., [10]		–	–	–	92.33	76.54	73.93	66.49	69.30
Co-CNN [39]		–	–	–	**97.06**	**89.15**	**87.83**	**81.73**	**83.78**
CE2P [64]	**LIP** [13]	53.10	–	–	63.20	–	–	–	–
Wang et al., [85]		**57.74**	**–**	**–**	**68.80**	–	–	–	–
Co-CNN [39]	**ATR** [39]	–	96.02	–	–	83.57	84.95	77.66	80.14
TGPNet [47]		–	96.45	–	–	87.91	83.36	80.22	81.76
Wang et al., [85]		–	**96.26**	**–**	**–**	**87.91**	**84.62**	**86.41**	**85.51**

Table 3 Results of different state-of-the-art methods for fashion parsing [68]

Model	Market-1501 [104]						DeepFashion [43]			
	SSIM	IS	Mask-SSIM	Mask-IS	DS	pSSIM	SSIM	IS	DS	pSSIM
PG2 [48]	0.261	**3.495**	0.782	3.367	0.390	–	0.773	3.163	0.951	–
Def-GAN [72]	0.291	3.230	0.807	3.502	0.720	–	0.760	**3.362**	0.976	–
PATN [91]	0.81	3.162	0.799	3.737	**0.796**	0.6186	0.771	3.201	0.976	0.799
Loss function [68]	**0.312**	3.326	**0.810**	**3.807**	0.742	**0.6415**	**0.776**	3.262	**0.982**	**0.813**
Real Data	1.000	3.890	1.000	3.706	0.740	1	1.000	4.053	0.968	1

5.2.3 Fashion landmark detection

The most popular evaluation metrics in fashion detection are Normalized Error (NE) and Percentage of Detected Landmarks (PDL). NE is considered as the distance between predicted landmarks and ground-truth, while PDL is defined as the percentage of detected landmarks according to overlapping criterion. Typically, smaller values of NE or higher values of PDL indicate better results.

5.3 IQA for fashion synthesis

The image quality evaluation is essential for image generation methods to synthesize desired outputs. Recent image synthesis research commonly uses simple loss functions to measure the difference between the generated image and the ground truth, e.g., L1-norm loss, adversarial loss, and perceptual loss. Here, we will present related evaluation metrics to each tasks of fashion synthesis including style transfer, pose transfer and clothing simulation.

5.3.1 Style transfer and pose transfer

Image based garment transfer aims to transform a source person image to a target pose while retaining the appearance details. In this case two essential tasks are required to ensure this goal. That are, style transfer and pose transfer which are very challenging tasks especially in the case of human body occlusion, large pose transfer and complex textures and for measuring the quality of generated images common metrics are used. The evaluation for style transfer is generally based on subjective assessment by rating the results into certain degrees and the percentages of each degree are, then, calculated to evaluate quality of results.

5.3.2 Physical simulation

There are limited quantitative comparisons between physical simulation works. Most of them tend to calculate the qualitative results only within their work or show the vision comparison with related works. Figure 16 presents an example of these comparisons.

As shown in this section, the fashion assessment is based on inception score or human preference score. However, inception score focuses more on the image quality, regardless of the aesthetic factors. Human preference score obtained from a small group can be easily influenced by the users' personal preference or the environment. Thus, one of the challenging tasks in research domain is to build a novel fashion assessment metric that is objective and robust.

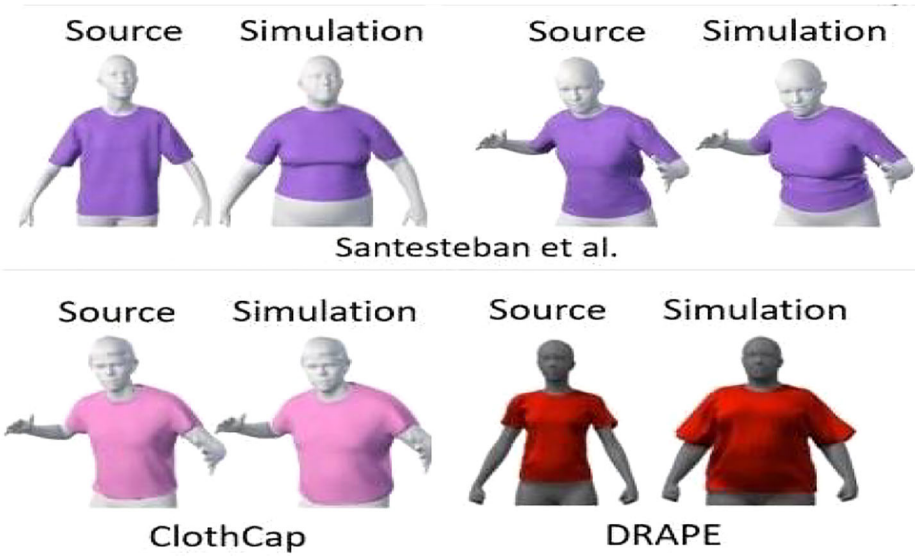

Fig. 16 Evaluation of the work of Santesteban et al. [4] compared with DRAPE [65] and ClothCap [101]

6 Application and future work

Automate the manual processes is a great achievement insured by technology advancements especially in the computer vision field. One of the largest industries that is influenced by technology advancement is Fashion Apparel. Due to computer vision powered tools, a great experience can be born for both retailers and consumers. In the following, we present the application of fashion technology uses in various areas and present the future works needed to realize the target benefits.

6.1 Application

Fashion is an ever-changing industry, where trends succeed one another, and companies must constantly rethink and adapt their products and strategies to maintain their position and assure customers' preference. AI based research appears to be a promising avenue for the fashion industry and can be applied for various activities to enhance the working on this area and maximize the financial gains. Creating AI systems that can understand fashion in images, can create a next-level customer experience like online fashion shopping because apparel industry is basically about visual, thus, it can be dealing with computer vision to recognize images just as we do by making computers understand images.

Here is where the future research work will bring value and become useful for fashion business by making smart shopping. The application of computer vision is mainly done for fashion image analysis, object detection and image retrieval [40, 43]. Many other researchers have represented their ideas for feature extraction and accurate attribute, for fashion related images [16, 62]. Recently, many researchers tried to explore and provide solutions for different fashion tasks using the concepts of artificial intelligence. Several works contributed for fashion recommendation in [20, 80], object detection and classification [37, 43, 44, 83, 95], Image

Generation and Manipulation in [17, 67, 70]. Figure 17 illustrates an overview of the AI application in the field of Fashion.

6.2 Challenges

Going completely online brings a vast number of challenges for fashion retailers and gives an inspiration for new innovative digital products like virtual fitting systems to make the wholesale process completely digital. Published literature presented in this survey show the potential of AI techniques for providing important solutions to implement intelligent systems. Despite that, clothing companies do not widely use these advanced techniques due to various limitations related to these field. A virtual fitting would be a way to see the virtual effects, but it is still far from solved due to several challenges.

Image-based virtual try-on is among the most potential approach of virtual fitting that tries on a target clothes into a customer's image and thus it has received considerable research efforts in the recent years, however, there are several challenges involved in development of virtual try-on that makes it difficult to achieve realistic outfit such as pose, occlusion, cloth texture, logo and text etc. In this section, we present the most important challenges which can be treated in the incoming studies in the field of adoption of AI techniques in clothing industry.

Try-on image generation Creating realistic images and videos of persons by considering the pose, shape and appearance is a crucial challenge related to the application of computer vision in many fields like movie production, content creation, visual effects, and virtual reality, etc., In virtual try-on, the body shape and the desired pose of the person highly influence the final appearance of the target clothing item [21, 22, 61, 89, 100, 106]. Thus, diverse questions must be asked to overcome many challenges: (1) How to deform the new clothing item and align it with the target person in a proper manner, and (2) How to generate the try-on image with

Fig. 17 Applications of AI techniques in fashion industry

preserving visual details of the clothing item, and maintaining the body parts of the person, during clothes interchange according to the person pose. Recently, diverse research works [51, 66, 67, 74, 96] take this challenge to respond to these questions and try to solve different issues related to the image generation but it seems that the necessity of obtaining photo-realistic images still persist, thus, the need to improve existing virtual try-on system.

Network efficiency It is a very important factor to apply algorithms in real-life applications. Diversity data can improve the robustness of networks to handle complex scenes with irregular poses, occluded body limbs and crowded people. The main issue is related to system performance which is still far from human performance in real-world settings [33, 65, 86, 101]. The demand for a more robust system consequently grows with it. Thus, it is crucial to pay attention to handling data bias and variations for performance improvements. Moreover, there is a definite need to perform the task in a light but timely fashion. It is also beneficial to consider how to optimize the model to achieve higher performance. Some existing methods used Transfer Learning and Data Augmentation [57, 61], but we need to focus for more performant methods to achieve high quality results within efficient network.

Virtual try-on DATASETS Datasets are very important for validating the new models. In particular, deep learning model needs large-scale data for training task. One of the early realistic and large-scale datasets in the fashion area is DeepFashion [43]. So, building new datasets would help quick progress in virtual try-ons and in some cases, there are a necessity to extend existing datasets by using different methods. 1) The GAN worked as a technique of data augmentation which helps in overcome the weakness of existing fashion datasets. 2) Synthetic technology can theoretically generate unlimited data while there is a domain gap between synthetic data and real data. 3) Cross-dataset supplementation to supplement 3D datasets with 2D datasets, can mitigate the problem of insufficient diversity of training data. 4) Transfer learning proves to be useful in this application. Therefore, how to create or extend a large-scale dataset constitutes a promising direction for both image-based dataset and video-based dataset.

Multi-modal virtual try-on Depending only on the appearance features such as clothing that extracted from RGB images are not robust enough against environment variations as shown in the above methods [7, 9, 12, 16, 21, 22, 27, 50, 61, 62, 84, 89, 98, 100, 106]. Thus, authors should try to combine multiple modalities with complementary information for the final task to improve the accuracy. So, using deep learning on multimodal data is one of new directions in virtual try-on. Also, one of the challenges in the multimodal, needs to be considered in new studies, is developing a framework that handles missing features or modalities that occur by occlusions or pose variations. In the last year, some research works present their interest to this challenge [45, 46].

Unsupervised /supervised fashion research Most of current deep learning try-on systems depend on supervised learning [16, 50, 62, 84, 98] which train labeled data in the same environment. So, training annotation data in new and real-world environments will conduct to high annotation cost while the deep learning models need enormous data for training and labelling presents a tedious and time-consuming process. To overcome this problem and relieve the labelling burden, it is very useful to work with unsupervised models to extract discriminative features from unlabeled dataset instead of supervised or weakly supervised

learning. In fact, current AI approaches require a lot of labeled data to achieve decent accuracy in their predictions. However, since labeling often requires expensive human labor and much time, AI techniques need to evolve toward Unsupervised Learning models that do not require labeled data to train the AI models. The use of this kind of learning begin with some works and become most in-demand in last year [59, 72, 75, 100].

2D/3D virtual try-on As mentioned in this survey, current methods such as [7, 9, 12, 16, 17, 21, 22, 27, 48, 50, 51, 61, 62, 66, 67, 70, 74, 84, 89, 96, 98, 100, 104, 106] are still far from the built of an ideal virtual try-on system for many reasons related to the input data. Firstly, clothes deformation and occlusion make the garment rendering process very hard. Also, 3D human body modeling for arbitrary poses is still challenging [2, 4, 5, 87]. Thus, new approaches should be proposed to capture detail of shape and clothing.

Fashion generation conditioned on text Although the advancement on the development of intelligent fashion systems, the automatic synthesis of photo-realistic images from text is needed to obtain perfect results in the design process and to generate realistic images. This need is due to the diverse attributes of fashion images in color, pattern, style, etc. So, research works must focus on how handling complex conditions as well as data sources should be inspired. This challenge is treated with some studies for fashion intelligent system such as Semantic-Spatial Aware GAN [23] and Inspirational adversarial image generation [63].

6.3 Open issues and future directions

Technology has always played an important role in fashion industry and started a more profound and faster transformation that is changing the way in which customers shop and interact with products and brands. At the same time, companies are adopting these technologies to ensure a best shopping experience. Virtual try-on applications present the irreplaceable technology in fashion industry, it provides important benefits to the apparel industries and allows to try-on garment before purchasing, improves accuracy, and suggests well-fitted garment for body type. Throughout the pandemic, virtual try-on has offered a great service to e-consumers and brands unable to demo their products offline.

Virtual try-on solutions represent fit to body, as well as garment pattern design, style, colors to get the perfect results of clothing fitting because the main purpose of retailers is to prove virtual try-on matching with the real garments. Thus, the priority of researchers is to identify the key challenges and the critical success factors that determine the effectiveness of the implementations of digital technologies in the online garment industry to bridge the gap between physical and digital shopping and to attain the challenge of reaching the people wherever they are which has been needed during the pandemic, and will continue to be with the rise of e-commerce.

The implementation of virtual try-on application has the potential to provide a significant benefit to clothing e-retailers but their adoption in the clothing sector is still limited, and even the technological advances, the existing try-on applications are not completely developed yet and still not matured to obtain target results. Most of them are not realistic enough to feel comfortable when try-on a garment item because the structure of clothes is not coherent and done in an artificial manner. Therefore, there are still many unresolved challenges and gap between research and practical applications such as those mentioned in the previous section.

This crucial challenges in adopting fashion technologies for fashion industry are appeared because real-world fashion is much more complex than in the experiments.

Following this objective, we present in this paper an interesting review of literature for the virtual try-on task, which can provide researchers with explicit research directions, facilitates their access to the related studies and improve the visibility of adopted methods. Thus, this literature review help to understand from existing works how we can implement an efficient virtual try-on system and how we can understand fashion image. However, people would show different views of themselves in the desired clothing product before making purchasing decision. Considering this objective, a virtual try-on system must be designed and developed, where given a person image, a desired pose, and a target clothing item, it can generate the try-on look of the person with the target appearances and desired poses. We illustrate this process in Fig. 18.

Towards this end, Most of the systems presented in this paper proceed as follow: they realized at first fashion detection to localize where in the image a fashion item appears or where the different body parts are localized. Then, they swap and interchange clothes between different images of persons and deal with the large variations on body poses and shapes via deep learning models. These studies show that there is significant progress has been made in this direction using learning-based image generation tools, such as GANs, and authorize various range of applications, such as human appearance interchange, virtual try-on, motion transfer, and novel appearances synthesis. However, because of the under constrained nature of these tasks, most existing methods have restriction in the visual quality on generated results and present observable artefacts such as blurring of small details, lose facial identity, unrealistic distortions of the body parts and garments as well as severe changes of the textures. The major procedures are not able to recover the texture details properly. Figure 19 show the result of the recent method of NHRR proposed by Sarkar et al. [67].

Despite the important results given by the approaches discussed in this survey, and the power of measuring technologies developed with deep learning methods, several limitations persist like the lack of perfection and the incorrect fit on the human body. Therefore, future studies should focus at providing realistic presentations of different target appearances of the consumers and allow them to virtually choose and try-on preferred clothes, adjust size, style, and color of desired items by using the deep learning-based approaches.

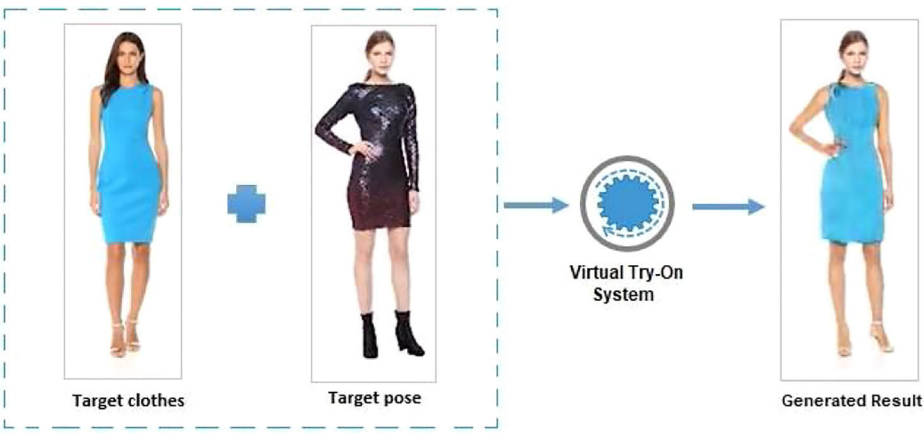

Fig. 18 Illustration of the idea of Virtual Try-On System

Fig. 19 Limitation of generated results of the virtual try-on task presented by the work of Sarkar et al. [67]

7 Conclusion

The advancements made with AI technologies in fashion industry have not yet reach the goal of modeling the real-world problems which is still very limited and remain challenging, and this is because important hurdles exist at various levels. Thus, the implementation of the AI techniques into this task requires a careful consideration of the various practical features existing in the clothing industry to ensure optimal solutions. The different studies on intelligent fashion analysis surveyed in this paper are just the beginning of this wide research domain because up to now, enormous research efforts have been spent on these tasks and will continue to grow and expand due to the enormous profit potential in the ever-growing fashion industry. This future directions must bridge the gap between research and real industry demand by adding new features and services with the intent of providing customers the same support and comfort that they would have during an in-person shopping experience.

Code availability Not applicable.

Data availability Not applicable.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2d human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3686–3693
2. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B (2018) Posetrack: a benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5167–5176

3. Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J (2018) Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8340–8348
4. Bhatnagar BL, Tiwari G, Theobalt C, Pons-Moll G (2019) Multi-garment net: learning to dress 3d people from images. In: proceedings of the IEEE/CVF international conference on computer vision, pp 5420–5430
5. Chen H, Liu X, Li X, Shi H, Zhao G (2019) Analyze spontaneous gestures for emotional stress state recognition: a micro-gesture dataset and analysis with deep learning. In: 2019 14th IEEE international conference on Automatic Face & Gesture Recognition (FG 2019), pp 1–8. IEEE
6. Cheng WH, Song S, Chen CY, Hidayati SC, Liu J (2021) Fashion meets computer vision: a survey. ACM Comput Surv 54(4):1–41
7. Cui YR, Liu Q, Gao CY, Su Z (2018) Fashiongan: display your fashion design using conditional generative adversarial nets. Comput Graph Forum 37(7):109–119
8. Dalmia A, Joshi S, Singh R, Raykar V (2018) Styling with attention to details. arXiv preprint arXiv: 1807.01182
9. Donato G, Belongie S (2002) Approximate thin plate spline mappings. In: European conference on computer vision. Springer, Berlin, Heidelberg, pp 21–31
10. Dong J, Chen Q, Xia W, Huang Z, Yan S (2013) A deformable mixture parsing model with parselets. In: Proceedings of the IEEE international conference on computer vision, pp 3408–3415
11. Ge Y, Zhang R, Wang X, Tang X, Luo P (2019) Deepfashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5337–5345
12. Ge Y, Song Y, Zhang R, Ge C, Liu W, Luo P (2021) Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8485–8493
13. Gong K, Liang X, Zhang D, Shen X, Lin L (2017) Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 932–940
14. Gong K, Liang X, Li Y, Chen Y, Yang M, Lin L (2018) Instance-level human parsing via part grouping network. In: Proceedings of the European conference on computer vision (ECCV), pp 770–785
15. Guan P, Reiss L, Hirshberg DA, Weiss A, Black MJ (2012) Drape: dressing any person. ACM Trans Graph 31(4):1–10
16. Han X, Wu Z, Wu Z, Yu R, Davis LS (2018) Viton: an image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7543–7552
17. Han X, Hu X, Huang W, Scott MR (2019) Clothflow: a flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10471–10480
18. Hariharan B, Cardie C, Adam H, Jia M, Shi M, Sirotenko M, Belongie S, Cui Y (2020) Fashionpedia: ontology, Segmentation, and an Attribute localization dataset
19. Hore A, Ziou D (2010) Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on pattern recognition, pp 2366–2369. IEEE
20. Hou M, Wu L, Chen E, Li Z, Zheng VW, Liu Q (2019) Explainable fashion recommendation: a semantic attribute region guided approach. arXiv preprint arXiv:1905.12862
21. Hsieh CW, Chen CY, Chou CL, Shuai HH (2019) Fit-me: image-based virtual try-on with arbitrary poses. In: 2019 IEEE international conference on image processing (ICIP), pp 4694–4698. IEEE
22. Hsieh CW, Chen CY, Chou CL, Shuai HH, Liu J, Cheng WH (2019) FashionOn: semantic-guided image-based virtual try-on with detailed human and clothing information. In: Proceedings of the 27th ACM international conference on multimedia, pp 275–283
23. Hu K, Liao W, Yang MY, Rosenhahn B (2021) Text to image generation with semantic-spatial aware GAN. arXiv preprint arXiv:2104.00567
24. Huang FH, Lu HM, Hsu YW (2021) From street photos to fashion trends: leveraging user-provided Noisy labels for fashion understanding. IEEE Access 9:49189–49205
25. Ionescu C, Papava D, Olaru V, Sminchisescu C (2013) Human3. 6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell 36(7):1325–1339
26. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
27. Issenhuth T, Mary J, Calauzenes C (2020) Do not mask what you do not need to mask: a parser-free virtual try-on. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, august 23–28, 2020, proceedings, part XX 16, pp 619–635. Springer International Publishing

28. Ji W, Li X, Zhuang Y, Bourahla OE, Ji Y, Li S, Cui J (2018) Semantic locality-aware deformable network for clothing segmentation. In: IJCAI, pp 764–770
29. Johnsen TE, Miemczyk J, Howard M. A (2017) A systematic literature review of sustainable purchasing and supply research: theoretical perspectives and opportunities for IMP-based research. Ind Mark Manag, 61, pp 130–143
30. Johnson S, Everingham M (2010) Clustered pose and nonlinear appearance models for human pose estimation. In: bmvc. Vol. 2, No. 4, p. 5
31. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst 25:1097–1105
32. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
33. Lahner Z, Cremers D, Tung T (2018) Deepwrinkles: accurate and realistic clothing modeling. In: Proceedings of the European conference on computer vision (ECCV), pp 667–684
34. Lee S, Oh S, Jung C, Kim C. A (2019) A global-local embedding module for fashion landmark detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops
35. Li S, Liu ZQ, Chan AB (2014) Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, p 482-489
36. Li J, Zhao J, Wei Y, Lang C, Li Y, Feng J (2017) Towards real world human parsing: multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206, pp 3193–3202
37. Li Y, Tang S, Ye Y, Ma J (2019) Spatial-aware non-local attention for fashion landmark detection. In: 2019 IEEE international conference on multimedia and expo (ICME), pp 820–825. IEEE
38. Li J, Su W, Wang Z (2020) Simple pose: rethinking and improving a bottom-up approach for multi-person pose estimation. Proceedings of the AAAI conference on artificial intelligence 34(07):11354–11361
39. Liang X, Xu C, Shen X, Yang J, Liu S, Tang J, Lin L, Yan S (2015) Human parsing with contextualized convolutional neural network. In: Proceedings of the IEEE international conference on computer vision, pp 1386–1394
40. Liao L, He X, Zhao B, Ngo CW, Chua TS (2018) Interpretable multimodal retrieval for fashion products. In: Proceedings of the 26th ACM international conference on multimedia, pp 1571–1579
41. Liu S, Feng J, Domokos C, Xu H, Huang J, Hu Z, Yan S (2013) Fashion parsing with weak color-category labels. IEEE Trans Multimed 16(1):253–265
42. Liu S, Liu L, Yan S (2014) Fashion analysis: current techniques and future directions. IEEE Multimed 21(2):72–79
43. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1096–1104
44. Liu Z, Yan S, Luo P, Wang X, Tang X (2016) Fashion landmark detection in the wild. In: European conference on computer vision. Springer, Cham, pp 229–245
45. Liu J, Song X, Chen Z, Ma J (2020) MGCM: multi-modal generative compatibility modeling for clothing matching. Neurocomputing 414:215–224
46. Liu L, Zhang H, Zhou D (2021) Clothing generation by multi-modal embedding: a compatibility matrix-regularized GAN model. Image Vis Comput 107:104097
47. Luo X, Su Z, Guo J, Zhang G, He X (2018) Trusted guidance pyramid network for human parsing. In: Proceedings of the 26th ACM international conference on multimedia, pp 654–662
48. Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L (2017) Pose guided person image generation. Advances in neural information processing systems 31 (NIPS 2017), pp 406–416
49. Men Y, Mao Y, Jiang Y, Ma WY, Lian Z (2020) Controllable person image synthesis with attribute-decomposed Gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5084–5093
50. Minar MR, Tuan TT, Ahn H, Rosin P, Lai YK (2020) Cp-vton+: clothing shape and texture preserving image-based virtual try-on. In: CVPR Workshops
51. Neuberger A, Borenstein E, Hilleli B, Oks E, Alpert S (2020) Image based virtual try-on network from unpaired data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5184–5193
52. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, Cham, pp 483–499
53. Omid Mohammadi S, Kalhor A (2021) Smart fashion: a review of AI applications in the Fashion & Apparel Industry. arXiv e-prints, pp arXiv–a2111

54. Oyewusi WF, Adekanmbi O, Ibejih S, Osakuade O, Okoh I, Salami M. (2021) AFRIFASHION1600: a contemporary African fashion dataset for computer vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3968–3972
55. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy K (2014) 2d human pose estimation: new benchmark and state of the art analysis. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 3686–3693
56. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. and Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4903–4911
57. Peng X, Tang Z, Yang F, Feris RS, Metaxas D (2018) Jointly optimize data augmentation and network training: adversarial data augmentation in human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2226–2234
58. Pons-Moll G, Pujades S, Hu S, Black MJ (2017) ClothCap: seamless 4D clothing capture and retargeting. ACM Trans Graph 36(4):1–15
59. Pumarola A, Agudo A, Sanfeliu A, Moreno-Noguer F (2018) Unsupervised person image synthesis in arbitrary poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8620–8628
60. Puri D (2019). COCO dataset stuff segmentation challenge. In: 2019 5th international conference on computing, communication, control and automation (ICCUBEA), pp 1–5. IEEE
61. Raj A, Sangkloy P, Chang H, Lu J, Ceylan D, Hays J (2018) Swapnet: garment transfer in single view images. In: Proceedings of the European conference on computer vision (ECCV), pp 666–682
62. Ren B, Tang H, Meng F, Ding R, Shao L, Torr PH, Sebe N (2021) Cloth interactive transformer for virtual try-on. arXiv preprint arXiv:2104.05519
63. Rozière B, Riviere M, Teytaud O, Rapin J, LeCun Y, Couprie C (2021) Inspirational adversarial image generation. IEEE Trans Image Process 30:4036–4045
64. Ruan T, Liu T, Huang Z, Wei Y, Wei S, Zhao Y (2019) Devil in the details: towards accurate single and multiple human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence Vol. 33, No. 01, pp 4814–4821
65. Santesteban I, Otaduy MA, Casas D, inventors; Seddi Inc, assignee (2021) Learning-based animation of clothing for virtual try-on. U.S. Patent Application 16/639,923
66. Sarkar K, Mehta D, Xu W, Golyanik V, Theobalt C (2020) Neural re-rendering of humans from a single image. In: European conference on computer vision. Springer, Cham, pp 596–613
67. Sarkar K, Golyanik V, Liu L, Theobalt C (2021) Style and pose control for image synthesis of humans from a single monocular view. arXiv preprint arXiv:2102.11263
68. Shi H, Le Wang, Tang W, Zheng N, Hua G (2020) Loss Functions for Person Image Generation. In: BMVC
69. Si C, Wang W, Wang L, Tan T (2018) Multistage adversarial losses for pose-based human image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 118–126
70. Siarohin A, Sangineto E, Lathuiliere S, Sebe N (2018) Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3408–3416
71. Song S, Mei T (2018) When multimedia meets fashion. IEEE Multimed 25(3):102–108
72. Song S, Zhang W, Liu J, Mei T (2019) Unsupervised person image generation with semantic parsing transformation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2357–2366
73. Statista (2020) Global Apparel Market Statistics & Facts. https://www.statista.com/topics/5091/apparel-marketworldwide. Accessed 10 July 2021
74. Sun F, Guo J, Su Z, Gao C (2019) Image-based virtual try-on network with structural coherence. In: 2019 IEEE international conference on image processing (ICIP), pp 519–523. IEEE
75. Sun S, Li X, Li J (2021) January. UCCTGAN: unsupervised clothing color transformation generative adversarial network. In: 2020 25th international conference on pattern recognition (ICPR), pp 1582–1589. IEEE
76. Tang W, Wu Y (2019) Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1107–1116
77. Tang W, Yu P, Wu Y (2018) Deeply learned compositional models for human pose estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 190–206
78. Tang Z, Peng X, Geng S, Wu L, Zhang S, Metaxas D (2018) Quantized densely connected u-nets for efficient landmark localization. In: Proceedings of the European conference on computer vision (ECCV), pp 339–354

79. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1653–1660
80. Turkut Ü, Tuncer A, Savran H, Yılmaz S. (2020) June. An online recommendation system using deep learning for textile products. In: 2020 international congress on human-computer interaction, optimization and robotic applications (HORA), pp 1–4. IEEE
81. Vandana M (2020) Here are 3 ways businesses can survive and thrive through COVID-19 and beyond. World Economic Forum. https://www.weforum.org/agenda/2020/07/here-are-3-ways-for-businesses-to-survive-and-thrive-through-covid-19. Accessed 12 July 2021
82. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
83. Wang W, Xu Y, Shen J, Zhu SC (2018) Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4271–4280
84. Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M (2018) Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV), pp 589–604
85. Wang W, Zhang Z, Qi S, Shen J, Pang Y, Shao L (2019) Learning compositional neural information fusion for human parsing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5703–5757
86. Wang TY, Shao T, Fu K, Mitra NJ (2019) Learning an intrinsic garment space for interactive authoring of garment animation. ACM Trans Graph 38(6):1–12
87. Wang J, Wen C, Fu Y, Lin H, Zou T, Xue X, Zhang Y (2020) Neural pose transfer by spatially adaptive instance normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5831–5839
88. Wu J, Zheng H, Zhao B, Li Y, Yan B, Liang R, Wang W, Zhou S, Lin G, Fu Y, Wang Y (2017) Ai challenger: a large-scale dataset for going deeper in image understanding. arXiv preprint arXiv: 1711.06475
89. Wu Z, Lin G, Tao Q, Cai J (2019) M2e-try on net: fashion from model to everyone. In: Proceedings of the 27th ACM international conference on multimedia, pp 293–301
90. Xian W, Sangkloy P, Agrawal V, Raj A, Lu J, Fang C, Yu F, Hays J (2018) Texturegan: controlling deep image synthesis with texture patches. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8456–8465
91. Xu M, Chen Y, Liu S, Li TH, Li G (2021) Structure-transformed texture-enhanced network for person image synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13859–13868
92. Yahong Z (2021) Virtual try-on: the next big thing in luxury business. Hapticmedia. https://hapticmedia.com/blog/virtual-try-on. Accessed 10 June 2021
93. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 3570–3577
94. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2014) Retrieving similar styles to parse clothing. IEEE Trans Pattern Anal Mach Intell 37(5):1028–1040
95. Yan S, Liu Z, Luo P, Qiu S, Wang X, Tang X (2017) Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In: Proceedings of the 25th ACM international conference on multimedia, pp 172–180
96. Yang F, Lin G (2021) CT-Net: Complementary Transfering Network for Garment Transfer with Arbitrary Geometric Changes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9899–9908
97. Yang S, Ambert T, Pan Z, Wang K, Yu L, Berg T, Lin MC (2016) Detailed garment recovery from a single-view image. arXiv preprint arXiv:1608.01250
98. Yang H, Zhang R, Guo X, Liu W, Zuo W, Luo P (2020) 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, pp 7850–7859
99. Yoo D, Kim N, Park S, Paek AS, Kweon IS (2016) Pixel-level domain transfer. In: European conference on computer vision. Springer, Cham, pp 517–532
100. Yu R, Wang X, Xie X (2019) 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, pp 10511–10520
101. Yu T, Zheng Z, Zhong Y, Zhao J, Dai Q, Pons-Moll G, Liu Y (2019) Simulcap: single-view human performance capture with cloth simulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5504–5514

102. Zhao J, Li J, Yu C, Sim T, Yan S, Feng J (2018) Look into person: joint body parsing & pose estimation network and a new benchmark. IEEE Trans Pattern Anal Mach Intell 41(4):871–885
103. Zhao J, Li J, Yu C, Sim T, Yan S, Feng J (2018) Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: Proceedings of the 26th ACM international conference on Multimedia, pp 792–800
104. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
105. Zheng S, Yang F, Kiapour MH, Piramuthu R (2018) Modanet: a large-scale street fashion dataset with polygon annotations. In: Proceedings of the 26th ACM international conference on multimedia, pp 1670–1678
106. Zheng N, Song X, Chen Z, Hu L, Cao D, Nie L (2019) Virtually trying on new clothing with arbitrary poses. In: Proceedings of the 27th ACM international conference on multimedia, pp 266–274
107. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
108. Zhu H, Cao Y, Jin H, Chen W, Du D, Wang Z, Cui S, Han X (2020) Deep Fashion3D: a dataset and benchmark for 3D garment reconstruction from single images–supplemental materials
109. Zou X, Kong X, Wong W, Wang C, Liu Y, Cao Y (2019) Fashionai: a hierarchical dataset for fashion understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops