You can Try without Visiting: A Comprehensive Survey on Virtually Try-on Outfits

Hajer Ghodhbani [1,1], Adel Alimi [2], Mohamed Neji [2], and Imran Razzak [2]

[1]National Engineering School of Sfax (ENIS)
[2]Affiliation not available

November 8, 2023

Abstract

Our work aims to conduct a comprehensive literature review of deep learning methods applied in the fashion industry and, especially, the image-based virtual fitting task by citing research works published in the last years. We have summarized their challenges, their main frameworks, the popular benchmark datasets, and the different evaluation metrics. Also, some promising future research directions are discussed to propose improvements in this research field.

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** [4]

Abstract Virtual try-on enables the customers to try-on the products using their camera equipped devices. With the help of augmented reality, customers can contextually visualize the product they are interested, by virtually confirming the size, style, and the fit before making a purchase. 21% of retailers are already using virtual try-on in campaigns and 46% of retailers planned to deploy virtual solutions. 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 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 open research challenges and recommendations.

Keywords virtual try-on · fashion industry · fashion detection · fashion synthesis

Hajer Ghodhbani
E-mail: hajer.ghodhbani@regim.usf.tn

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

1 Introduction

In 2012, Converse first used virtual iPhone try-on and the users were able to use phone cameras to see how shoes looked on them, and post photos on social media as well as make online purchases. In the last few years 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. Virtual try-on technology enables the customer to visualize the produce on themselves and see how certain the products look on them before purchasing. It 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, 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. Big names 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 two percent 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 8.4% for 2021[1].

During COVID19 pandemic lockdown, most of the business went into kind of a crisis mode and not only big names but also small retailers are thinking how they can survive. Taking our time in shops will be difficult in a post-Covid-19 world as a result, online shopping is ingrained significantly in our daily 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 invest in the way to explore new sales methods and optimization of technological process of purchasing clothes like virtual fitting system. These solutions draw a new picture of online shopping experience and bring it to a high level of reality and comfort.

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 to achieve more accurate results.

In the literature, a little number of fashion surveys are proposed [2, 3]. Recently, a summary on intelligent clothing analysis was made by Liu et al. [2]. In addition, Song and Mei [3] 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 [4]. 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 planned improvements to extent research on this area?

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) [5] 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 140 articles from both journals and conference.

Articles were retrieved from popular databases and engines such as Google scholar [1] andResearch Gate [2] , then, a screening process is used to select the articles relevant 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.

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

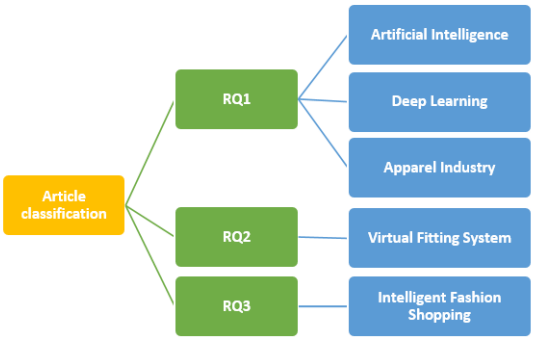

Fig. 1 Article Classification based on Research Questions

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

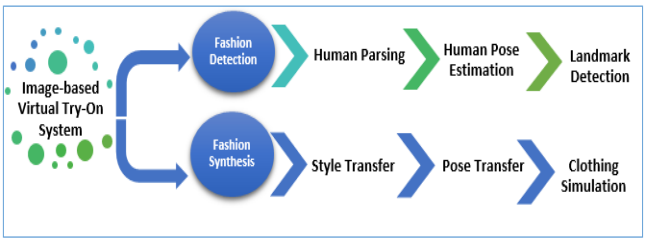

Fig. 2 Classification of based approaches for image-based virtual try-on System

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. [6].

Fig. 3 Examples of fashion parsing based on semantic segmentation [6].

In fashion domain, largest number of potential applications have been devoted to various tasks and particularly to human parsing [7-12]. At the begin, Yamaguchi et al. [7] proposed a model by merging the fashion parsing and the human pose estimation. Then, they proposed clothes parsing with a retrieval-based approach [9] to resolve the constrained parsing problem. After that, a weak supervision approach for fashion parsing is presented by Liu et al. [10] who resort to label images with color-category labels instead of pixel-level. These works conduct results far from being perfect because between pose estimation and clothing parsing there is no consistent targets. Thus, many studies realized in this context to fix this limitation, such as the work of Dong et al. [11] which proposed a traditional hand-crafted pipeline based on many hand-designed processing steps which lead this work to be imperfect solution. After that and with a contextualized approach, Liang et al. [12] handle the human parsing task by providing the clothing tags at the image level. 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. [15] 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, Liang et al. [12] designed a Contextualized CNN (Co-CNN) to take the context of cross-layer, global image-level, and local super-pixel. In 2018, Liao et al. [14] built a Matching CNN (M-CNN) network to solve the issues of parametric and non-parametric CNN-based methods. In the same year, Liang 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 [15], the same authors proposed a JPPNet network [16] to treat both the human parsing and human pose estimation task.

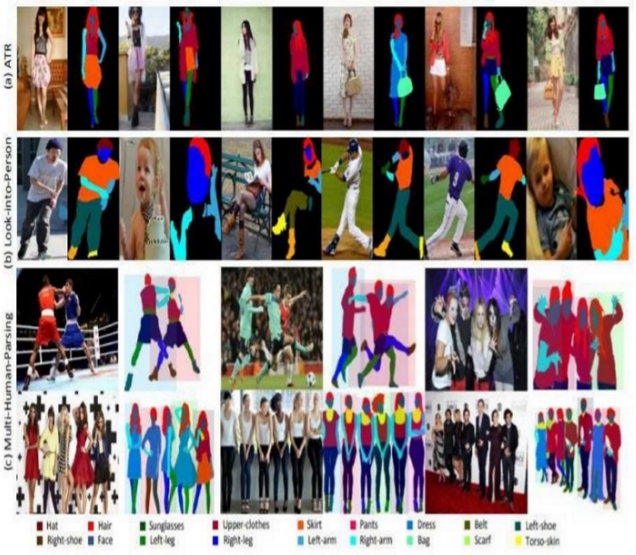

Fig. 4 Annotation examples for constructed: (a): ATR [12]; (b): LIP [16] and (c): Multi-Human Parsing: MHP [17].

Different from the previous mentioned works that only concentrated on single person parsing task, there are many others works [17-19] which focus on treating the scenario with multiple views of persons. Zhao et al. [17] designed a deep Nested Adversarial Network (NAN) to understand humans in crowed scenes. This network is composed, respectively, of three Generative Adversarial Network (GAN) for semantic saliency prediction, instance-agnostic parsing, and instance-aware clustering. Gong et al. [19] 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. [18] developed a Context Embedding with Edge Perceiving (CE2P) framework. In recent years, hierarchical graph is used for human parsing tasks [20, 21] to improve parsing performance. Wang et al. [20] considered the human body as a hierarchy of multi-level semantic parts to capture the human parsing information. Gong et al. [21] designed a human parsing model untitled Graphonomy by including hierarchical graph into conventional parsing network with the exploitation of transfer learning technique.

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.

A- Single-person Human Pose Estimation

Single-person human pose estimation (HPE) is related to the task of localizing human skeletal key- points of from an image or video data. In the following Figure (Fig. 5), we present results of Single-person HPE obtained from MPII Human Pose dataset [22].

Fig. 5 Example of human pose estimation from DeepPose[27]

Most early, Single-person HPE methods began with a traditional way by adopting handcraft feature extraction and sophisticated body models to obtain local representations and global pose structures [23, 24]. Then, deep learning-based methods have resorted to neural

networks [25, 26] to extend the traditional works. 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 directly by learning mapping from image [27]. The early deep learning-based network adopted by many researchers was AlexNet [28] due to its simple architecture. Toshev et al. [27] applied AlexNet to learn joint coordinates from full images. Also, Pfister et al. [29] exploited this network to ensure the prediction of the human pose from videos. Then, Luvizon et al. [30] proposed a regression approach with Soft-argmax function to ensure the directly conversion of feature maps to joint coordinates. This framework enabled the learning of heat maps representations, without requiring more steps of artificial ground truth generation.

Due to the difficulty of prediction directly the joint coordinates from input images, many works interested to this challenge and proposed effective networks based on body model structure. Sun et al. [31] proposed a structure-aware regression method using bones instead of joints. Li et al. [32] employed an AlexNet as a multi-task framework to predict the joint coordinate from full images. For video sequences, Luvizon et al. [33] used a multi-task deep learning method to deal with both pose estimation and action recognition.

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 [28] or joints [34]. For a more supervision information and easy training, recent works [35, 36] used heatmaps based methods to indicate joint's ground truth location. Papandreou et al. [37] proposed a fully convolutional ResNet to ameliorate the representation of joint location with the prediction of dense heatmaps and offsets. GoogleNet proposed a network with multi-scale inputs [38] and ResNet-based network with deconvolutional layers [39] to ameliorate classic network. Many works [40-42] tackled the problem of design networks in a multi-stage style to refine results from coarse prediction.

Previous works attempt to adjust detected body parts into body models, but there are other recent works [43-45] which aim to encode human body structure information into networks. Yang et al. [43] proposed a CNN to predict joint locations in heatmap representation. An RNN was proposed in the work of Chu et al. [45] to output joint location one by one and transform kernels by a bi-directional tree to pass information between corresponding joints in a tree body model. Tang et al. [46] proposed a hierarchical representation of body parts, then, they extended their work [47] to learn specific features of part group. Additionally, Chou et al. [48] introduced adversarial training including two hourglass networks with same architecture. Chen et al. [49] proposed a CNN to effectively localize the human body parts by taking priors into account during training. Peng et al. [50] exploited data augmentation to avoid the need of more data during training. Luo et al. [51] exploited temporal information with RNN redesigned from CPM by changing multi-stage architecture with LSTM structure. Tang et al. [52] committed to improve the network structure by proposing a densely connected U-nets and efficient usage of memory. Feng et al. [53] adopted a model learning strategy called Fast Pose Distillation (FPD) to design Hourglass network.

B- 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. Fig. 6 shows examples of results from the work of Li et al. [54].

Fig. 6 Example of multi-person HPE [54]

A combination of existing detection networks and single HPE networks used to implement the Top- down HPE methods [22, 26, 37] 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- 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. Fig. 7 shows results generated by the fashion landmark detection approach.

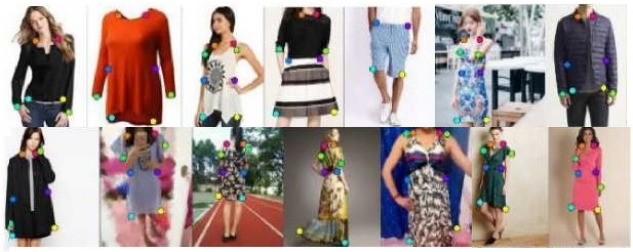

Fig. 7 Example of results from Fashion Landmark Detection approach [56]. First row illustrates the results on DeepFashion-C [57], second row presentes results on FLD dataset [55].

For the first time, Liu et al. [57] presented fashion landmark concept and, in parallel, they proposed a deep model called FashionNet [57] applied on predicted clothing landmarks. Then, they proposed a deep fashion alignment framework [55] 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. [58] used to relax constraint of clothing bounding box due to its difficult application. A more recent work [59] 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. [60] 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. [61] proposed the well-known 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. [62] transferred the input image or sketch to the corresponding texture (Fig. 8).

Fig. 8 Examples of image style transfer by TextureGAN [62].

Driven by increasing power of deep generative models, popular virtual try-on applications have appeared [63-67]. Han et al. [63] proposed VIrtual Try-On Network (VITON) to try clothing on image of person by generating a coarse tried-on result and predicted the mask for the clothing item, then, a refinement network for the clothing region was employed to synthesize a more detailed result. This framework fails to handle large deformation, especially with more texture details, due to the imperfect shape-context matching for aligning clothes and body shape. CP-VTON model [64] was proposed to deal with this issue by handling the spatial deformation with a Geometric Matching Module, which explicitly aligned input clothing with body shape. In the same year, another work called La-viton is proposed to allow the generation of try-on images with preserving of appearance and characteristics of clothing items. Fig. 9 presents some results from these works [63-65].

The previous works required in-shop clothing image for virtual try-on, but other existing models like FashionGAN [75] and M2E-TON [69] 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 ensure rendering of the output image by the segmentation map.

Other works attempts to resolve the problem with arbitrary poses such as Fit-Me [69] which was the first work building virtual try-on dealing with this challenge.

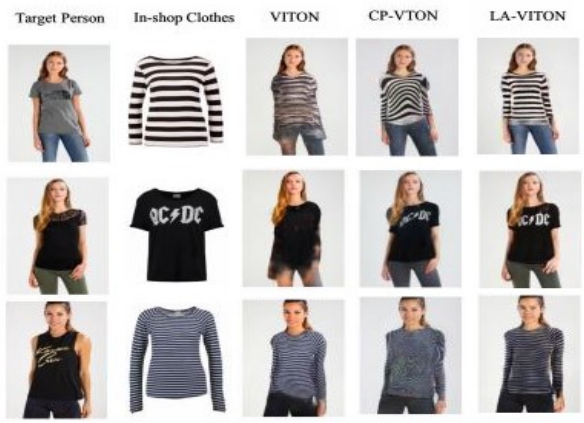

Fig. 9 Results from the VITON [63], CP-VTON [64] and LA-VITON [65].

Fig. 10 Garment transfer results generated by the work of Sarkar et al. [79].

Then, FashionOn [70] applied the semantic segmentation to present more realistic results. SwapNet [71] allow the transfer of all the clothing from one person's image onto the pose of another target person by operating in image-space. This is done by first generating a mutually exclusive segmentation mask of the desired clothing on the desired pose. Another virtual try-on network called Vtnfp [72] proposed a strategy to synthesize photo-realistic images given the images of a clothed person and a target clothing item. Firstly, the warped clothing is generated, followed by the body segmentation map of the person in target clothing, and finally a synthesis module is used to obtain the final image synthesis. Zheng et al. [73] 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. [74] 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 [71, 72, 76, 77] 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. [76] resolve the problems of visual details and the missing of body parts by maintain the structural between the generated image and the reference image. Then, Outfit-VITON [77] allowing 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. [78, 79] achieve high-quality try-on results by aligning the given human images with a 3D mesh model via DensePose [137], estimating a UV texture map corresponding to the desired garments, and rendering this texture onto the desired pose (Fig. 10).

The generative model, Attribute-decomposed GAN (ADGAN) [80], 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. In 2021, a similar conditioning model is adopted by Dressing in Order (DiOr) framework [35] to support 2D pose transfer, virtual try-on, and several fashion editing tasks. Despite this diversity of these systems, the ability to preserve details or to present, correctly, the shape and the texture presents 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. 11). To deal with this task, many works are proposed. Firstly, a pose guided person image generation PG2 [81] is presented with a two-stage adversarial network to achieve an early attempt on the challenging task of transferring a person to different poses. This framework generated both poses and appearance simultaneously by dividing the problem into two stages. Pose information are used in the first stage to generate human body structure in the desired image. Then and during the second stage, a deep convolutional GAN is used to treat the output of the first stage. This framework shows results for texture details which were highly blurred. To tackle this problem, the affine transform was employed to keep textures in the generated images better.

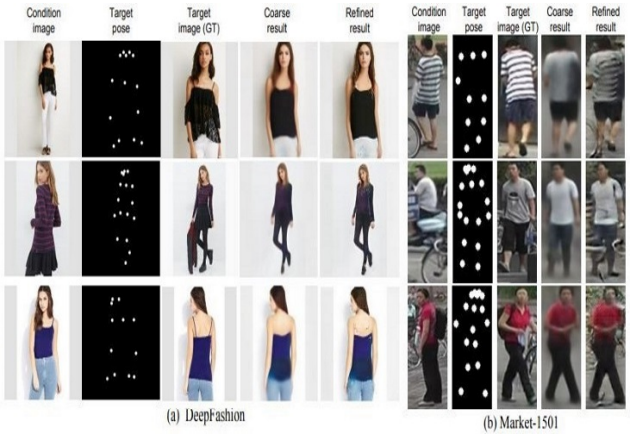

Fig. 11 Examples of pose transformation results generated by the work of Liqian Ma, et al. [81] from DeepFashion dataset [57] (a) and Market-1501 dataset [82] (b)..

The work of Siarohin et al. [83] used a deformable GAN to generate images of person according to a target pose which allowed the extraction of the articulated object pose by resorting to a keypoint detector. Other recent work [84] 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. [85] 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. [86] treat the limitation of data presented by the above research studies by borrowing the idea from [87] and leveraging cycle consistency. In 2019, the work of Song et al. [88] 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 subtasks to improve the appearance performance.

3.3 Clothing Simulation

For more amelioration 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 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 [89-91]. Yang et al. [91] 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. [89] aimed to dressing people in a different variation and pose, and clothing types with an automatic process. Thus, they proposed a model called DRAPE (DRessing Any PErson) to simulate clothes deformation with varying shape and pose (Fig. 12.). Pons-Moll et al. [90] proposed ClothCap approach 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.

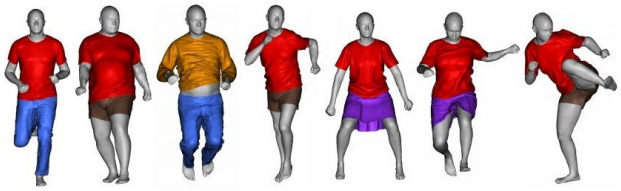

Fig. 12 Example of clothing simulation results obtained with DRAPE model [89]

The simulation of the physical deformation has an 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. [92] interested on this task and proposed a semi-automatic method to learn the intrinsic physical properties with different postures to generate garment animation which are shown in Fig. 13. 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.

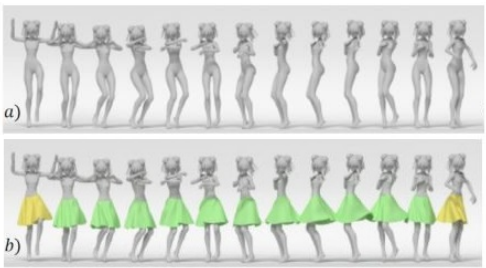

Fig. 13 Examples of physical simulation from the work of Wang et al. [92].

To improve more realistic view to the garment on human body, Lahner et al. [93] 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. [94] 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. [95] proposed a physics-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.

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 table1.

As summarized in table 1, for each task there are specific datasets with according setting. Market- 1501 [82] and Deep-Fashion [57] are the most popular datasets for virtual try-on. Fashion Landmark Dataset [55] is the most used dataset for fashion landmark detection. Several datasets were built to treat the fashion parsing task such as LIP dataset [15]. 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 Deep-Fashion [57], DeepFashion2 [96] and FashionAI [97], the building of datasets basing on 3D clothing is almost rare or not sufficient for training like the digital wardrobe released by MGN [44]. In 2020, Heming et al. [8] develop a comprehensive dataset named Deep Fashion3D which is richly annotated and coversa 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. 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.

5.2.1 Fashion parsing

In fashion Parsing, various metrics are used to evaluate proposed approaches on different datasets such as Fashionista [7] and LIP [15] and in terms of average Pixel Accuracy (aPA), mean Average Garment Recall (mAGR)

Table 1 Summary of the benchmark datasets for fashion tasks

Task	Dataset	Number of photos	Description	Publish time
Virtual Try-on	LookBook [107]	84,748	Composed by 9,732 top product images and 75,016 fashion model images	2016
	DeepFashion [57]	78,979	Selected from the In-shop Clothes Benchmark and associated with several sentences as captions and a segmentation map.	2016
	VITON [63]	32,506	Contained around 19,000 frontal-view woman and top clothing image pairs, yielding 16,253 pairs	2018
	FashionTryOn [73]	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
Fashion Parsing	Fashionista [7]	158,235	Outfit information in the form of tags, comments, and links	2012
	Paper Doll [9]	339,797	Annotated with metadata tags denoting characteristics, e.g., color, style, occasion, clothing type, brand	2013
	LIP [15]	50,462	- Focus on semantic understanding of person. 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 [113]	4,980	Instance-aware setting with fine-grained pixel-level annotations works with 7 body parts and 11 clothes categories	2017
	MHP v2.0 [17]	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)[19]	38,280	Multi-person images and Pixel-wise annotations in instance-level	2018
	ModaNet [111]	55,176	Annotated with pixel-level labels, bounding boxes, and polygons	2018
	DeepFashion2 [96]	491,000	Diverse images of 13 popular clothing categories from both commercial shopping stores and consumers. Labeled with scale, occlusion, zoom-in, viewpoint, category, style, bounding box, dense landmarks and per-pixel mask.	2019
Fashion landmark detection	DeepFashion-C [57]	289,222	Annotated with clothing bounding box, pose variation type, landmark visibility, clothing type, category, and attributes	2016
	Fashion Landmark Dataset (FLD) [70]	123,016	Annotated with clothing type, pose variation type, landmark visibility, clothing bounding box, and human body joint	2016
	Unconstrained Landmark Database (ULD) [58]	30,000	- Collected from fashion blogs, forums and the consumer-to shop retrieval benchmark of DeepFashion [57] - Contains substantial foreground scatters and background clutters	2017
	DeepFashion2 [96]	491,000	DeepFashion2 used in diverse tasks like fashion parsing, clothes detection, pose estimation, segmentation, and retrieval.	2019
Human Pose Estimation	MPII Human pose [22]	2.5104	Data are from YouTube videos. It covers 410 human activities, and each image is provided with activity label	2014
	MSCOCO [108]	328,000	Data are from Internet and it used for diverse activities.	2014
	AI Challenger [109]	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 [110]	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 [112]	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 [82]	32,668	Contains over 32,000 annotated boxes, plus a distractor set of over 500K images. Images produced using the Deformable Part Model (DPM) as pedestrian detector.	2015
	DeepFashion [57]	52,712	In-shop Clothes Retrieval Benchmark DeepFashion is used for pose transfer	2016

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 [7] in terms of accuracy, average precision, average recall and average F-1 score over pixels.

Table 2 Performance comparisons of fashion parsing methods (in %) [5]

Method	Dataaset	Evaluation Metrics			
		mIOU	aPA	mAGR	AVG.F-1
Yamaguchi et al.,[7]	Fashionista [7]	-	-	-	46.80
Liang et al.,[13]		-	-	-	69.30
Co-CNN [12]		-	-	-	83.78
CE2P [18]	LIP [15]	53.10	87.37	-	-
Wang et al.,[20]		57.74	88.03	-	-
Co-CNN [12]	ATR [12]	-	96.02	-	80.14
TGPNet [98]		-	96.45	-	81.76
Wang et al.,[20]		-	96.26	-	85.51

5.3.1 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.

5.3.2 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.

5.4 IQA for Fashion synthesis

The image quality evaluation is essential for image generation methods to synthesize desired outputs. Recent image synthesis research work commonly use 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.

5.4.1 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. Also, there are objective comparisons for virtual try-on, in terms of inception score (IS) or structural similarity (SSIM). IS [99] is used to evaluate the synthesis quality of images quantitatively. SSIM [100] 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 Structural Similarity (SSIM) [100], Inception Score (IS) [99], masked version of Structural Similarity (mask-SSIM) [100], masked version of Inception Score (mask-IS) [99] and Detection Score (DS) [96] applied on Market-1501 dataset [82] and DeepFashion dataset [57].

Table 3 Results of different state-of-the-art methods for fashion parsing [106]

Model	Market-1501 [82]			
	SSIM	IS	DS	pSSIM
PG2 [95]	0.261	3.395	0.390	-
Def-GAN [96]	0.291	3.230	0.720	-
PATN [102]	0.81	3.162	0.796	0.6186
Loss function [106]	0.312	3.326	0.742	0.6415
Real Data	1.000	3.890	0.740	1
	DeepFashion [72]			
PG2 [95]	0.773	3.163	0.951	-
Def-GAN [96]	0.760	3.362	0.976	-
PATN [102]	0.771	3.201	0.976	0.799
Loss function [106]	0.776	3.262	0.982	0.813
Real Data	1.000	4.053	0.968	1

5.4.2 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. Fig. 14 presents an example of these comparisons.

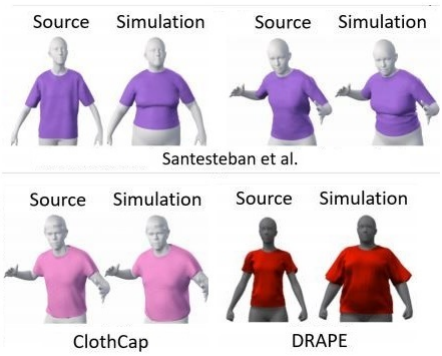

Fig. 14 Evaluation of the work of Santesteban et al. [94] compared with DRAPE [89] et ClothCap [105].

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 this research domain is to build a novel fashion assessment metric that is objective and robust.

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.

6.0.1 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 [14, 57]. Many other researchers have represented their ideas for feature extraction and accurate attribute, for fashion related images [55, 59]. 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 [104, 105], object detection and classification [22-29], Image Generation and Manipulation in [101-103]. Fig. 15 illustrates an overview of the AI application in the field of Fashion.

Fig. 15 Applications of artificial intelligence in fashion industry

Published literature presented in this survey show the potential of AI techniques for providing support in problem-solving processes involved in garment manufacturing. Despite these advantages, clothing companies do not widely use these advanced techniques due to challenges and limitations related to these field.

6.1 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. This goal can be achieved by using AI technology that has the power to better engage them with the personalized shopping experience that leads them to make more informed and confident purchase decisions.

Various fashion brands implemented online virtual fitting systems in attempt to reduce return rates and improve customer satisfaction. A virtual fitting would be a way to see the virtual effects, but it is still far from solved due to the challenge to virtually change the texture and pattern of clothes deformation and shading. Therefore, there are several challenges and issues can be observed, and they will direct the incoming studies in the field of adoption of AI techniques in clothing industry.

A/ 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. 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 preserving visual details of the clothing item, and maintaining the body parts of the person, during clothes interchange according to the person pose.

B/ 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. 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 thus also beneficial to consider how to optimize the model to achieve higher performance.

C/ 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 [57]. 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.

D/ 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, authors 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.

E/ Unsupervised Supervised Fashion Research

Most of current deep learning try-on systems depend on supervised learning 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 unsupervised or weakly supervised learning in fashion domain are necessary to. 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.

F/ Efficiency VS Accuracy

The best accuracy is achieved mostly by large models, but these large models may consume a lot of time and memory size which affects the efficiency of these models especially when applied to real time applications. Most existing models did not consider the processing time and memory size for the goal of achieving higher accuracy. The trade-off between ranking accuracy and the processing time is required and should be considered by authors that working in these fields.

G/ 2D/3D Virtual Try-On

As mentioned in this survey, current methods 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 mod-

eling for arbitrary poses is still challenging. Thus, new approaches should be proposed to capture detail of shape and clothing.

H/ 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.

6.2 Open Issues and Future Directions

Technology has always played an important role in fashion industry 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.

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, researchers' priority 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.

The implementation of virtual try-on application has the potential to provide a significant benefit to clothing e-tailers 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 in industry are appeared because real-world fashion is much more complex than in the experiments.

In our future work, we will aim to provide an efficient virtual try-on system for fashion retailers to ensure a better shopping experience for customers. This goal can be achieved by developing an intelligent system to understand fashion images. 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. 16.

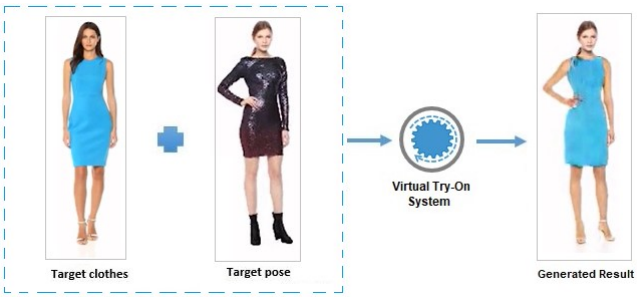

Fig. 16 Illustration of Virtual Try-On System 'function

Existing works mentioned in this survey 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. Recently, human re-rendering procedures are not able to recover the texture details properly. Fig. 17 show the result of the recent method of NHRR proposed by Sarkar et al. [78]

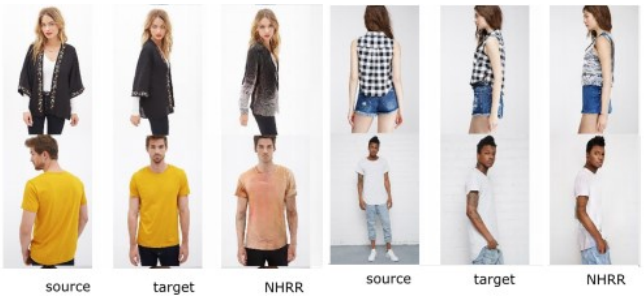

Fig. 17 Limitation of generated results of the virtual try-on task presented by the work of Sarkar et al. [78].

Despite the important results given by these approaches 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, our future study will focus at providing realistic presentations of different target appearances of the consumers and allow them to virtually choose and try-on preferred clothing items, adjust size, style, and color of desired items by using the deep learning-based approaches. Towards this end, this system must realize at first fashion detection to localize where in the image a fashion item appears or where the different body parts are localized. Then, it would swap and interchange clothes between different images of persons and deal with the large variations on body poses and shapes.

7 Conclusion

Since the last years and until now, technology has made fast progress for many industries, in particu¬larly, 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. The 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. The technology advancement in fashion Apparel, pave the way for virtual try-ons since the digital garments can be projected into virtual environment. Virtual try-ons for style purposes were already recommended to effectively address suit and fit issues of online shopping, so, there is a growing number of software-developing companies creating diverse try-on solutions to ensure a virtual fitting system.

Despite advancements made by with AI technologies in fashion industry, modeling the real-world problems is still very limited and remain challenging. 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 solutions 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.

References

1. statista (2019) Apparel market worldwide. https://www.statista.com/study/54163/apparel-retail-worldwide/
2. Liu S, Liu L, Yan S (2014) Fashion analysis: Current techniques and future directions. IEEE MultiMedia, 21(2), 72-79.
3. Song S, Mei T (2018) When multimedia meets fashion. IEEE MultiMedia, 25(3), 102-108.
4. Cheng W H, Song S, Chen C Y, Hidayati S C, Liu J (2020) Fashion meets computer vision: A survey. arXiv preprint arXiv:2003.13988.
5. Johnsen T E, Miemczyk J, Howard M (2017) A systematic literature review of sustainable purchasing and supply research: Theoretical perspectives and opportunities for IMP-based research. Industrial Marketing Management, 61, 130-143.
6. Ji W, Li X, Zhuang Y, Bourahla OE, Ji Y, Li S, Cui J (2018) Semantic Locality-Aware Deformable Network for Clothing Segmentation. InIJCAI (pp. 764-770).
7. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In2012 IEEE Conference on Computer vision and pattern recognition (pp. 3570-3577). IEEE.
8. 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. InEuropean Conference on Computer Vision (pp. 512-530). Springer, Cham.
9. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2014) Retrieving similar styles to parse clothing. IEEE transactions on pattern analysis and machine intelligence. 29;37(5):1028-40.
10. Liu S, Feng J, Domokos C, Xu H, Huang J, Hu Z, Yan S (2013) Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia. 11;16(1):253-65.
11. Dong J, Chen Q, Xia W, Huang Z, Yan S (2013) A deformable mixture parsing model with parselets. InProceedings of the IEEE International Conference on Computer Vision (pp. 3408-3415).
12. Liang X, Xu C, Shen X, Yang J, Liu S, Tang J, Lin L, Yan S (2015) Human parsing with contextualized convolutional neural network. InProceedings of the IEEE international conference on computer vision (pp. 1386-1394).
13. Liang X, Liu S, Shen X, Yang J, Liu L, Dong J, Lin L, Yan S (2015) Deep human parsing with active template regression. IEEE transactions on pattern analysis and machine intelligence. 3;37(12):2402-14.
14. Liao L, He X, Zhao B, Ngo CW, Chua TS (2018) Interpretable multimodal retrieval for fashion products. InProceedings of the 26th ACM international conference on Multimedia (pp. 1571-1579).
15. 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. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 932-940).
16. Liang X, Gong K, Shen X, Lin L (2018) Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE transactions on pattern analysis and machine intelligence. 29;41(4):871-85.
17. Zhao J, Li J, Cheng Y, Sim T, Yan S, Feng J (2018) Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. InProceedings of the 26th ACM international conference on Multimedia (pp. 792-800).

18. Ruan T, Liu T, Huang Z, Wei Y, Wei S, Zhao Y (2019) Devil in the details: Towards accurate single and multiple human parsing. InProceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4814-4821).
19. Gong K, Liang X, Li Y, Chen Y, Yang M, Lin L (2018) Instance-level human parsing via part grouping network. InProceedings of the European Conference on Computer Vision (ECCV) (pp. 770-785).
20. Wang W, Zhang Z, Qi S, Shen J, Pang Y, Shao L (2020) Learning compositional neural information fusion for human parsing. InProceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5703-5713).
21. Gong K, Gao Y, Liang X, Shen X, Wang M, Lin L (2019) Graphonomy: Universal human parsing via graph transfer learning. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7450-7459).
22. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2d human pose estimation: New benchmark and state of the art analysis. InProceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).
23. Chen X, Yuille A (2014) Articulated pose estimation by a graphical model with image dependent pairwise relations. arXiv preprint arXiv:1407.3399.
24. Gkioxari G, Hariharan B, Girshick R, Malik J (2014) Using k-poselets for detecting people and localizing their keypoints. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3582-3589).
25. Jain A, Tompson J, Andriluka M, Taylor GW, Bregler C (2013) Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302.
26. Ouyang W, Chu X, Wang X (2014) Multi-source deep learning for human pose estimation. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2329-2336).
27. Toshev A, Szegedy C (2014) Deeppose: Human pose estimation via deep neural networks. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 1653-1660).
28. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Communications of the ACM. 60(6):84-90.
29. Pfister T, Simonyan K, Charles J, Zisserman A (2014) Deep convolutional neural networks for efficient pose estimation in gesture videos. InAsian Conference on Computer Vision (pp. 538-552). Springer, Cham.
30. Luvizon DC, Tabia H, Picard D (2019) Human pose regression by combining indirect part detection and contextual information. Computers & Graphics. 85:15-22.
31. Sun X, Shang J, Liang S, Wei Y (2017) Compositional human pose regression. InProceedings of the IEEE International Conference on Computer Vision (pp. 2602-2611).
32. Li S, Liu ZQ, Chan AB (2014) Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. InProceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 482-489).
33. Luvizon DC, Picard D, Tabia H (2018) 2d/3d pose estimation and action recognition using multitask deep learning. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5137-5146).
34. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. InEuropean conference on computer vision (pp. 483-499). Springer, Cham.
35. Cui A, McKee D, Lazebnik S (2021) Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3940-3945).
36. Jain A, Tompson J, LeCun Y, Bregler C (2014) Modeep: A deep learning framework using motion features for human pose estimation. InAsian conference on computer vision (pp. 302-315). Springer, Cham.
37. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy K (2017) Towards accurate multiperson pose estimation in the wild. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 4903-4911).
38. Rafi U, Leibe B, Gall J, Kostrikov I (2016) An Efficient Convolutional Network for Human Pose Estimation. InBMVC (Vol. 1, p. 2).
39. B. Xiao, H. Wu, Y. Wei (2018) Simple baselines for human pose estimation and tracking. In: The European Conference on Computer Vision (ECCV).
40. Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. InProceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 4724-4732).
41. Yang W, Li S, Ouyang W, Li H, Wang X (2017) Learning feature pyramids for human pose estimation. Inproceedings of the IEEE international conference on computer vision (pp. 1281-1290).
42. Belagiannis V, Zisserman A (2017) Recurrent human pose estimation. In2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 468-475). IEEE.
43. Sun K, Xiao B, Liu D, Wang J (2017) Deep high-resolution representation learning for human pose estimation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5693-5703).
44. Bhatnagar BL, Tiwari G, Theobalt C, Pons-Moll G (2019) Multi-garment net: Learning to dress 3d people from images. InProceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5420-5430).
45. Gkioxari G, Toshev A, Jaitly N. (2016) Chained predictions using convolutional neural networks. InEuropean Conference on Computer Vision (pp. 728-743). Springer, Cham.
46. Tang W, Yu P, Wu Y (2018) Deeply learned compositional models for human pose estimation. InProceedings of the European conference on computer vision (ECCV) (pp. 190-206).
47. Tang W, Wu Y (2019) Does learning specific features for related parts help human pose estimation?. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1107-1116).
48. Chou CJ, Chien JT, Chen HT (2018) Self adversarial training for human pose estimation. In2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 17-30). IEEE.
49. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).
50. 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. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2226-2234).
51. Güler RA, Neverova N, Kokkinos I (2018) Densepose: Dense human pose estimation in the wild. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 7297-7306).

52. Tang Z, Peng X, Geng S, Wu L, Zhang S, Metaxas D (2018) Quantized densely connected u-nets for efficient landmark localization. InProceedings of the European Conference on Computer Vision (ECCV) (pp. 339-354).
53. Zhang F, Zhu X, Ye M (2019) Fast human pose estimation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3517-3526).
54. Li J, Su W, Wang Z (2020) Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. InProceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11354-11361).
55. Liu Z, Yan S, Luo P, Wang X, Tang X (2016) Fashion landmark detection in the wild. InEuropean Conference on Computer Vision (pp. 229-245). Springer, Cham.
56. Li Y, Tang S, Ye Y, Ma J (2019) Spatial-aware nonlocal attention for fashion landmark detection. In2019 IEEE International Conference on Multimedia and Expo (ICME) (pp. 820-825). IEEE.
57. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 1096-1104).
58. Yan S, Liu Z, Luo P, Qiu S, Wang X, Tang X (2017) Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. InProceedings of the 25th ACM international conference on Multimedia (pp. 172-180).
59. Wang W, Xu Y, Shen J, Zhu SC (2018) Attentive fashion grammar network for fashion landmark detection and clothing category classification. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4271-4280).
60. Lee S, Oh S, Jung C, Kim C (2019) A global-local embedding module for fashion landmark detection. InProceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).
61. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
62. 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. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8456-8465).
63. Han X, Wu Z, Wu Z, Yu R, Davis LS (2018) Viton: An image-based virtual try-on network. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 7543-7552).
64. Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M (2018) Toward characteristic-preserving image-based virtual try-on network. InProceedings of the European Conference on Computer Vision (ECCV) (pp. 589-604).
65. Jae Lee H, Lee R, Kang M, Cho M, Park G (2019) LA-VITON: a network for looking-attractive virtual try-on. InProceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).
66. Lewis KM, Varadharajan S, Kemelmacher-Shlizerman I (2021) VOGUE: Try-On by StyleGAN Interpolation Optimization. arXiv preprint arXiv:2101.02285.
67. 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. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7850-7859).
68. Hsieh CW, Chen CY, Chou CL, Shuai HH, Cheng WH (2019) Fit-me: Image-based virtual try-on with arbitrary poses. In2019 IEEE International Conference on Image Processing (ICIP) (pp. 4694-4698). IEEE.
69. Wu Z, Lin G, Tao Q, Cai J (2019) M2e-try on net: Fashion from model to everyone. InProceedings of the 27th ACM International Conference on Multimedia (pp. 293-301).
70. 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. InProceedings of the 27th ACM International Conference on Multimedia (pp. 275-283).
71. Raj A, Sangkloy P, Chang H, Lu J, Ceylan D, Hays J (2018) Swapnet: Garment transfer in single view images. InProceedings of the European Conference on Computer Vision (ECCV) (pp. 666-682).
72. Yu R, Wang X, Xie X (2019) Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. InProceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10511-10520).
73. Zheng N, Song X, Chen Z, Hu L, Cao D, Nie L (2019) Virtually trying on new clothing with arbitrary poses. InProceedings of the 27th ACM International Conference on Multimedia (pp. 266-274).
74. Han X, Hu X, Huang W, Scott MR (2019) Clothflow: A flow-based model for clothed person generation. InProceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10471-10480).
75. Cui YR, Liu Q, Gao CY, Su Z (2018) Fashiongan: Display your fashion design using conditional generative adversarial nets. InComputer Graphics Forum (Vol. 37, No. 7, pp. 109-119).
76. Sun F, Guo J, Su Z, Gao C (2019) Image-based virtual try-on network with structural coherence. In2019 IEEE International Conference on Image Processing (ICIP) (pp. 519-523). IEEE.
77. Neuberger A, Borenstein E, Hilleli B, Oks E, Alpert S (2020) Image based virtual try-on network from unpaired data. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5184-5193).
78. Sarkar K, Mehta D, Xu W, Golyanik V, Theobalt C (2020) Neural re-rendering of humans from a single image. InEuropean Conference on Computer Vision (pp. 596-613). Springer, Cham.
79. 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.
80. Men Y, Mao Y, Jiang Y, Ma WY, Lian Z (2020) Controllable person image synthesis with attribute-decomposed gan. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5084-5093).
81. Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L (2017) Pose guided person image generation. arXiv preprint arXiv:1705.09368.
82. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. InProceedings of the IEEE international conference on computer vision (pp. 1116-1124).
83. Siarohin A, Sangineto E, Lathuiliere S, Sebe N (2018) Deformable gans for pose-based human image generation. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3408-3416).
84. Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J (2018) Synthesizing images of humans in unseen poses. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8340-8348).
85. SI Chenyang, WANG Wei, WANG Liang, et al. (2018) Multistage adversarial losses for pose-based human image

synthesis. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 118-126.
86. Pumarola A, Agudo A, Sanfeliu A, Moreno-Noguer F (2018) Unsupervised person image synthesis in arbitrary poses. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8620-8628).
87. ZHU, Jun-Yan, PARK, Taesung, ISOLA, Phillip, et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In : Proceedings of the IEEE international conference on computer vision. p. 2223-2232.
88. Song S, Zhang W, Liu J, Mei T (2019) Unsupervised person image generation with semantic parsing transformation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2357-2366).
89. Guan P, Reiss L, Hirshberg DA, Weiss A, Black MJ (2012) Drape: Dressing any person. ACM Transactions on Graphics (TOG). 31(4):1-0.
90. Pons-Moll G, Pujades S, Hu S, Black MJ (2017) ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics (TOG). 36(4):1-5.
91. 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.
92. Wang TY, Shao T, Fu K, Mitra NJ (2019) Learning an intrinsic garment space for interactive authoring of garment animation. ACM Transactions on Graphics (TOG). 38(6):1-2.
93. Lahner Z, Cremers D, Tung T (2018) Deepwrinkles: Accurate and realistic clothing modeling. InProceedings of the European Conference on Computer Vision (ECCV) (pp. 667-684).
94. Santesteban I, Otaduy MA, Casas D, inventors; Seddi Inc, assignee (2021) Learning-based animation of clothing for virtual try-on. United States patent application US 16/639,923.
95. 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. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5504-5514).
96. 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. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5337-5345).
97. Zou X, Kong X, Wong W, Wang C, Liu Y, Cao Y (2019) Fashionai: A hierarchical dataset for fashion understanding. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).
98. Luo X, Su Z, Guo J, Zhang G, He X (2018) Trusted guidance pyramid network for human parsing. InProceedings of the 26th ACM international conference on Multimedia (pp. 654-662).
99. Hore A, Ziou D (2010) Image quality metrics: PSNR vs. SSIM. In2010 20th international conference on pattern recognition (pp. 2366-2369). IEEE.
100. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 13(4):600-12.
101. Kim BK, Kim G, Lee SY (2019) Style-controlled synthesis of clothing segments for fashion image manipulation. IEEE Transactions on Multimedia. 22(2):298-310.
102. Wu Q, Zhu B, Yong B, Wei Y, Jiang X, Zhou R, Zhou Q (2021) ClothGAN: generation of fashionable Dunhuang clothes using generative adversarial networks. Connection Science. 33(2):341-58.
103. Singh M, Bajpai U, Vijayarajan V, Prasath S (2019) Generation of fashionable clothes using generative adversarial networks: A preliminary feasibility study. International Journal of Clothing Science and Technology.
104. Liu J, Song X, Chen Z, Ma J (2020) MGCM: Multi-modal generative compatibility modeling for clothing matching. Neurocomputing. 414:215-24.
105. Turkut Ü, Tuncer A, Savran H, Yılmaz S (2020) An Online Recommendation System Using Deep Learning for Textile Products. In2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-4). IEEE.
106. Shi H, Le Wang 0003, Tang W, Zheng N, Hua G (2020) Loss Functions for Person Image Generation. InBMVC.
107. Yoo D, Kim N, Park S, Paek AS, Kweon IS (2016) Pixel-level domain transfer. InEuropean conference on computer vision (pp. 517-532). Springer, Cham.
108. Puri D (2019) COCO Dataset Stuff Segmentation Challenge. In2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-5). IEEE.
109. 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.
110. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B (2018) Posetrack: A benchmark for human pose estimation and tracking. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 5167-5176).
111. Zheng S, Yang F, Kiapour MH, Piramuthu R (2018) Modanet: A large-scale street fashion dataset with polygon annotations. InProceedings of the 26th ACM international conference on Multimedia (pp. 1670-1678).
112. 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 transactions on pattern analysis and machine intelligence. 36(7):1325-39.
113. Li Jianshu, Zhao Jian., Wei Yunchao, Lang, et al. (2017) Towards real world human parsing: Multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206, 3193-3202.