Image-Based Virtual Try-on System: A Survey of Deep Learning-Based Methods

Hajer Ghodhbani[1,*], Mohamed Neji[1], Adel M. Alimi[1]

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

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. Most of these works focus on the virtual try-on task to develop Image-based virtual fitting systems which present various challenging issues since persons can appear in differents poses and views. In last years, many studies have developped by using deep learning methods to face the challenges of pose variation, occlusion and illumination changes. Thus, we reviewed, in this paper, a large range of research works focused on using deep learning methods in image-based virtual fitting solutions by summarizing their challenges, their main frameworks and the popular benchmark datasets used for training. Hence, an overview of different evaluation metrics is presented with some examples of performance comparison, and lastly, some promising future research directions are discussed.

Keywords: Garment industry, deep learning, image-based virtual fitting, intelligent apparel industry

1. Introduction

In the few last years, 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. For this reason, online shopping for clothes has earned its place deservedly. 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) [1]. 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 [2].

The main reasons of online shopping growth in the last years, is that this kind of 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

*Corresponding author

Email addresses: hajerghodhbani@ieee.org (Hajer Ghodhbani), mohamed.neji@ieee.org (Mohamed Neji), adel.alimi@ieee.org (Adel M. Alimi)

technology to develop virtual try on applications that assist the fit of garment product in order 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 across images 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. Hence, recent studies adopted deep-learning-based methods to encounter these problems and to achieve more accurate results.

In the literature, a few fashion surveys are proposed [3, 4]. Most Early, an overview on intelligent facial and clothing analysis was presented by Liu et al. [3]. In 2018, Song and Mei [4] introduced the development of fashion tasks due to its emergence with multimedia. Next, a general survey painted the global picture of intelligent fashion without taken a specific problem [5]. Then 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 research 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.
The contribution of this work consists in responding to the following research questions:

- RQ1. What is the impact of using Artificial Intellignce (AI) and deep learning (DL) on fashion apparel 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 literature's review which is divided into two main 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 and presents the performance of popular works on different tasks. Section 5 shows related applications and future directions. Finally, a conclusion is given in Section 6.

2. Research Framework

In this study, a Systematic Literature Review (SLR) [6] 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, we have selected more than 130 articles from both journals and conference. Articles were retrieved from popular databases and engines such as Google scholar[1] and Research 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.

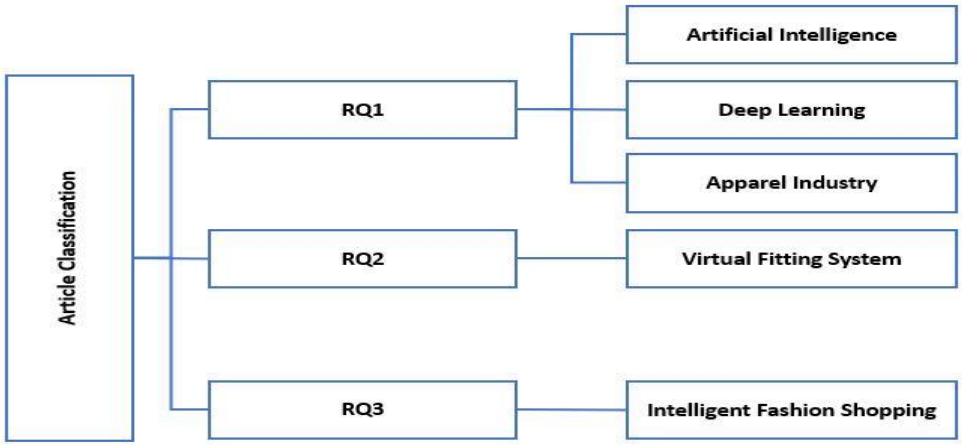

Fig. 1. Article Classification based on Research Questions

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. Review of literatures

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.

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

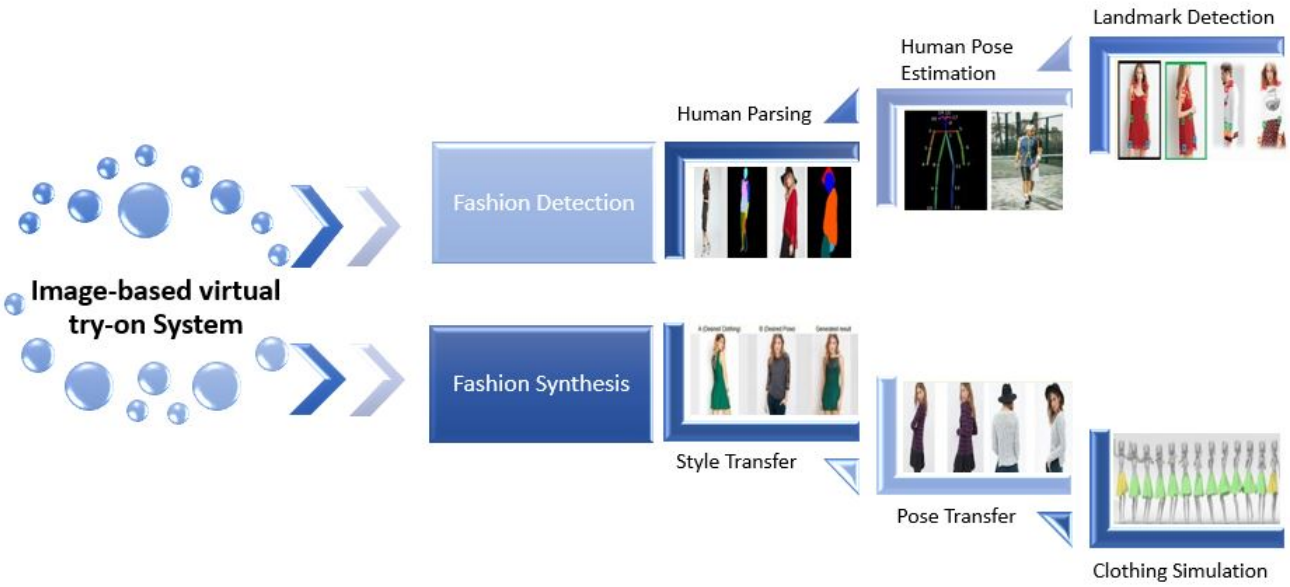

Fig. 2. Claasification of based approaches for image-based virtual try-on System

3.1. Fashion Detection

Fashion detection is an essential task for virtual try-on task, it consists to detect 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. [7]

Fig. 3. Examples of fashion parsing based on semantic segmentation [7].

In fashion domain, largest number of potential applications have been devoted to various tasks and particularly to human parsing [8–14]. The beginning is with Yamaguchi et al. [8] who proposed a combination between fashion parsing and human pose estimation. Then, they proposed clothes parsing with a retrieval-based approach [9, 10] to resolve the constrained parsing problem. After that, a weak supervision approach for fashion parsing are presented by Liu et al. [11] who resort to label images with color-category labels instead of pixel-level. The inconsistent targets between pose estimation and clothing parsing presented by these works leads to obtain results far from the perfect. Therefore, Other research studies attempted to relax this restriction, such as the work of Dong et al. [12] which proposed a traditional hand-crafted pipeline that wasn't considered as a perfect solution because many hand designed processing steps were needed. Then, Liang et al. [13, 14] treat the human parsing with the contextualized approach by providing the clothing tags at the image level. These hand-crafted approaches present many limitations because they need to be designed carefully.

To fix this problem, some approaches based on CNN are exploited like the framework of Liang et al. [15] based on deep human parsing with active template regression for semantic labeling. With the intent to ameliorate the parsing results of their previous work, Liang et al. [14] developed a Contextualized CNN (Co-CNN) architecture to capture, synchronously, the context of *cross-layer*, *global image-level*, and *local super-pixel*. In 2018, Liao et al. [16] 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. [17] implemented an important self-supervised method under the name "Look Into Person" (LIP) to eschew the necessity of labeling the human joints in model training (Fig. 4). Following their previous work [17], the same authors proposed a JPPNet network [18] to deal with both the human parsing and human pose estimation task.

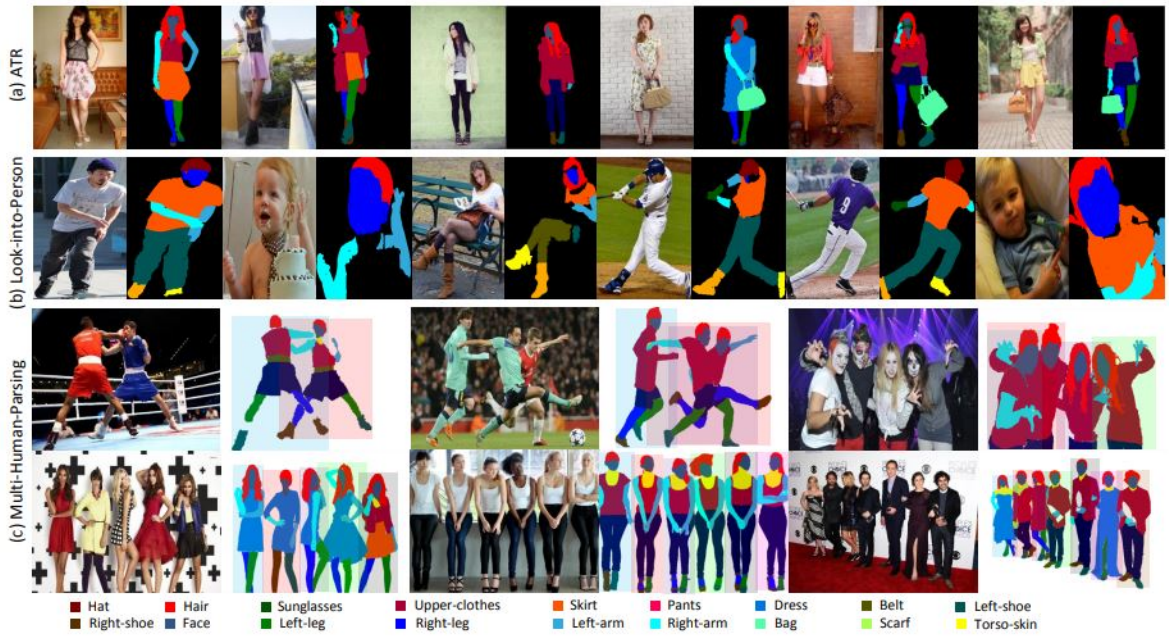

Fig. 4. Annotation examples for constructed: (a): ATR [14]; (b): LIP [18] and (c): Multi-Human Parsing:MHP [19].

Different from the abovementioned works that only focused on single person parsing task, several works [20–22] focus on handling the scenario with multiple persons. A deep Nested Adversarial Network (NAN) is presented in the work of Zhao et al. [20] 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. [22] proposed the first attempt to explore a detection-free Part Grouping Network (PGN) used for semantic part segmentation, instance-aware edge detection and instance-level human parsing. In 2019, Ruan et al. [21] presented a Context Embedding with Edge Perceiving (CE2P) framework to deal with both single and multiple human parsing. Most recently, hierarchical graph is considered for human parsing tasks [23, 24] to ensure perfect parsing performance. Wang et al. [23] considered the human body as a hierarchy of multi-level semantic parts to capture the human parsing information. Basing on transfer learning technique, Gong et al. [24] designed a human parsing model untitled Graphonomy by including hierarchical graph into conventional parsing network.

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 related to HPE, many research efforts have been devoted to the related fields. We present, in this section, recent researches 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) refers to the task of localizing human skeletal keypoints of a person from an image or video frames. In the following Figure (Fig. 5), we present some results of Single-person HPE obtained from *MPII Human Pose dataset* [25] and *Leeds Sports Poses (LSP) dataset* [26].

Fig. 5. Example of human pose estimation on (a)MPII [25] and (b)LSP [26]

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 [27–29]. Then, deep learning-based methods have resorted to neural networks [30, 31] 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 [32]. The early deep learning-based network adopted by many researches was AlexNet [33] due to

its simple architecture. Toshev et al. [32] applied AlexNet to learn joint coordinates from full images. Also, Pfister et al. [34] exploited this network to ensure the prediction of the human pose from videos. Then, Luvizon et al. [35] 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. Nibali et al. [36] proposed a numerical coordinate regression by using CNN to calculate joint coordinates from heatmaps.

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. [37] proposed a structure-aware regression method using bones instead of joints. Li et al. [38] employed an AlexNet as a multi-task framework to predict the joint coordinate from full images. a R-CNN architecture [39] is used to detect person, estimate pose, and classify action. Fan et al. [38] proposed a dual-source deep CNNs which take image patches to combine both local and contextual information to generate an output designed with a combination of joint detection and joint localization. For video sequences, Luvizon et al. [40] 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 [41]. For a more supervision information and easy training, recent works [42, 43] used heatmaps based methods to indicate joint's ground truth location. Papandreou et al. [44] 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 [45] and ResNet-based network with deconvolutional layers [46] to ameliorate classic network. Many works [47–51] 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 [52–59] which aim to encode human body structure information into networks. Yang et al. [52] proposed a CNN to predict joint locations in heatmap representation. An RNN was proposed [54] to output joint location one by one. Chu et al. [54] proposed to transform kernels by a bi-directional tree to pass information between corresponding joints in a tree body model. Tang et al. [59] proposed a hierarchical representation of body parts, then, they extended their work [60] to learn specific features of part group. Additionally, Chou et al. [61] introduced adversarial training including two hourglass networks with same architecture. Chen et al. [62] proposed a CNN to effectively localize the human body parts by taking priors into account during training. Peng et al. [63] exploited data augmentation to avoid the need of more data during training. Luo et al. [64] exploited temporal information with RNN redesigned from CPM by changing multi-stage architecture with LSTM structure. Tang et al. [65] committed to improve the network structure by proposing a densely connected U-nets and efficient usage of memory. Feng et al. [66] 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. [67]

Fig. 6. Example of multi-person HPE [67]

A combination of existing detection networks and single HPE networks used to implement the Top-down HPE methods [25, 26, 44, 53] that achieved state-of-the-art performance in almost benchmark datasets while the processing speed is depend 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 [68, 69] 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 an example of fashion landmark detection from the work of Liu et al. [70].

Fig. 7. Example of results from Fashion Landmark Detection approach [71]. First row presented the results on Deep-Fashion-C test set [72], and second row shows results on FLD dataset [70].

For the first time, Liu et al. [72] presented fashion landmark concept and, in parallel, they proposed a deep model called FashionNet [72] applied on predicted clothing landmarks. Then, they proposed

a deep fashion alignment framework [70] 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. [73] used to relax constraint of clothing bounding box due to its difficult application. A more recent work [74] mentioned that optimization on regression model is hard, so, they proposed to predict directly a confidence map of positional distributions for each landmark. Lee et al. [75] resorted to contextual knowledge to achieve perfect performance on landmark prediction. Ge et al. [76] built a Match R-CNN model to deal with their proposed versatile benchmark Deepfashion2 [77].

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. [78] proposed the well-known style transfer work, pix2pix, which is a general solution for style transfer. For specific goal, based on a texture patch, these works [79, 80] transferred the input image or sketch to the corresponding texture. An example of image style transfer from TextureGAN [80] is shown in Fig. 8.

Fig. 8. Examples of image style transfer by TextureGAN [80].

Han et al. [81] 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 [82] was proposed to deal with this issue by handling the spatial deformation with a Geometric Matching Module, which explicitly aligned the input clothing with the body shape. Fig. 9 presents some results of VITON [83] and CP-VTON [82].

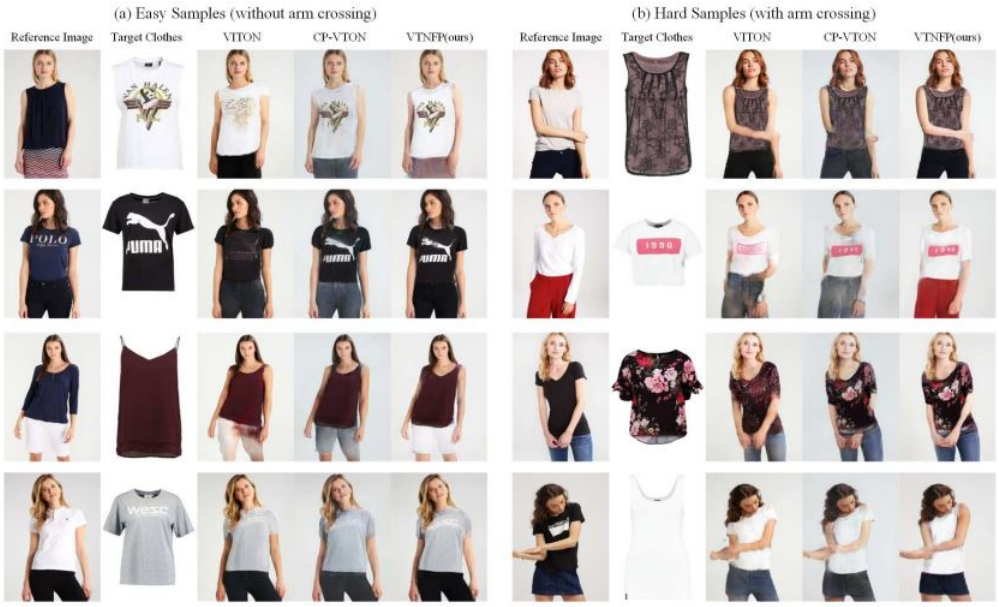

Fig. 9. Examples of results of VITON [83] and CP-VTON [82].

The previous works needed in-shop clothing image for virtual try-on, but there are other proposed models like FashionGAN [84] and M2E-TON [85] presented target try-on clothing image based 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. M2E-TON [85] was able to try on clothing from different images of people, and with different poses.

Other works attempts to resolve the problem with arbitrary poses such as Fit-Me [86] which was the first work building virtual try-on dealing with this challenge. Then, FashionOn [87] applied the semantic segmentation to present more realistic results. SwapNet [88] proposed a pipeline to transfer garment information across images with arbitrary clothing, body poses, and shapes by operating in image-space. Vtnfp [89] proposed a design strategy which generates, firstly, warped clothing, followed by body segmentation map of the person wearing the target clothing, and ending with a try-on synthesis module to fuse together all information for a final image synthesis. In 2019, Zheng et al. [90] proposed an architecture to Virtually trying on new clothing with arbitrary poses by using the body shape mask prediction for pose transformation. The work of Han et al. [91] focus on transferring the appearance naturally and synthesizing novel result by proposing ClothFlow model. In addition to their approaches related to image-based virtual try-on, Dong et al. [92] presented a Flow-Navigated Warping GAN for Video (FW-GAN) which aimed to synthesize a video of virtual try-on results.

Recent works [88, 89, 93, 94] address challenging task of transferring garment between person's pictures with preserving the identity in the source and target images. To solve the problems of body parts missing and visual details, Feng et al. [93] proposed a novel image-based virtual try-on network which maintain the structural consistency between the generated image and the original image by human parsing. Then, Outfit-VITON [94] synthesizes a cohesive outfit from multiple images of clothed human models, while fitting the outfit to the body shape and pose of the query person.

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. Some examples of pose transformation are presented in Fig. 10.

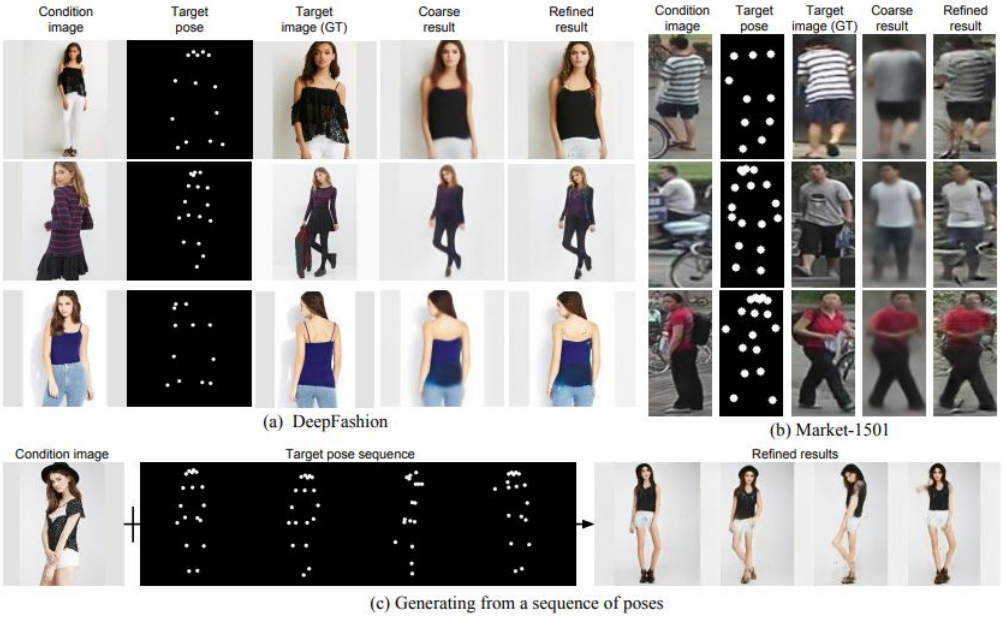

Fig. 10. Examples of pose transformation [95].

A two-stage adversarial network PG2 [95] achieved 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.

The work of Siarohin et al. [96] 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 [97] 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. [98] 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*.

Previous research works present data limitation that was taken by Pumarola et al. [99] which borrowed the idea from [100] by leveraging cycle consistency. Different approaches[101, 102] aimed to model body shape but they didn't show good results in the appearance of reference images. In 2019, the work of Song et al. [103] 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 abovementioned synthesis works are within the 2D domain where the 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.

The traditional way to simulate realistic clothes is the building of models by using computer graphics tools [104–107]. For learning both stretching and bending in real cloth samples, Wang et al. [106] proposed a piecewise linear elastic model. For learning the physical properties of clothing on different human body shapes and poses, Guan et al. [104] designed a pose-dependent model to simulate clothes deformation. Pons-Moll et al. [105] designed ClothCap to simulate clothing deformation of people in motion. As shown in Fig. 11, they separated garments from the human body to estimate the body shape and pose, then, they tracked the 3D deformations of the clothing over time from 4D scans to help simulate the physical clothing deformations in different human posture.

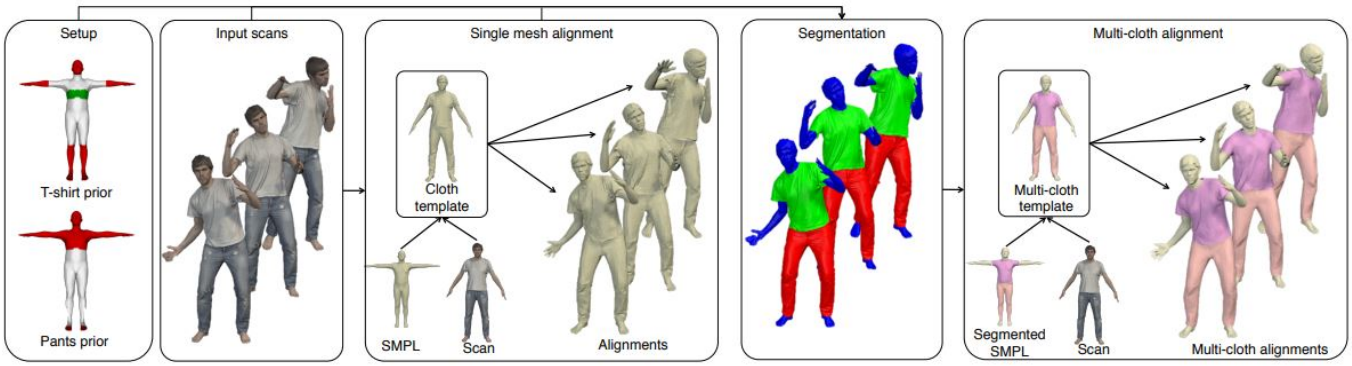

Fig. 11. Example of clothing simulation obtained by ClothCap [105].

4. Benchmark datasets

Recent advances in virtual try-on systems have been driven by the construction of clothes datasets. Due to the large variations in different tasks, it is difficult to build a universal dataset to evaluate the whole methods of virtual try-on. 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.

4.1. Fashion datasets

Large number of benchmark datasets proposed to study fashion applications such as virtual try-on systems. In Table 1, we summarize some of these datasets.

As summarized in table 1, for each task there are specific datasets with according setting. Market-1501 [115] and DeepFashion [72] are the most popular datasets for virtual try-on. Fashion Landmark Dataset [70] is the most used dataset for fashion landmark detection. For fashion parsing task,

Table 1: Summary of the benchmark datasets for fashion task

Task	Dataset	Description	Number of photo	Publish Date
Virtual Try-On	LookBook [108]	Composed by 9,732 top product images and 75,016 fashion model images	84,748	2016
	DeepFashion [72]	Selected from the In-shop Clothes Benchmark and associated with several sentences as captions and a segmentation map.	78,979	2016
	VITON [81]	Contained around 19,000 frontal-view woman and top clothing image pairs, yielding 16,253 pairs.	32,506	2018
	FashionTryOn [90]	Comprising 28, 714 clothing person-person triplets with each consisting of a clothing item image and two model images in different poses.	28,714	2019
Fashion landmark detection	DeepFashion-C [72]	Annotated with clothing bounding box, pose variation type, landmark visibility, clothing type, category, and attributes.	289,222	2016
	Fashion Landmark Dataset (FLD) [70]	Annotated with clothing type, pose variation, landmark visibility, clothing bounding box, and human body joint.	123,016	2016
	Unconstrained Landmark Database (ULD)[73]	Collected from fashion blogs, forums and the consumer-to shop retrieval benchmark of DeepFashion [72] and contains substantial foreground and background clutters	30,000	2017
	DeepFashion2 [77]	Used in diverse tasks like fashion parsing, clothes detection, pose estimation, segmentation, and retrieval.	491,000	2019
Human Pose Estimation	MPII Human pose [25]	Data are from YouTube videos. It covers 410 human activities and each image is provided with activity label.	2.5104	2014
	MSCOCO [109]	Data are from Internet. It used for diverse activities.	328,000	2014
	AI Challenger [110]	Provide three sub-datasets for human keypoint detection, attribute based zero-shot recognition and image captioning.	300,000	2017
	PoseTrack [111]	Focusses on 3 aspects: (1) single-frame multi-person pose estimation. (2) multi-person pose estimation in videos. (3) multi-person articulated tracking.	550 video sequences	2017

Fashion Parsing	Fashionista [8]	Outfit information in the form of tags, comments, and links	158,235	2012
	Paper Doll [9, 10]	Annotated with metadata tags denoting characteristics, e.g., color, style, occasion, clothing type, brand	339,797	2013
	Chictopia10k [112]	Contains real-world annotated images in the wild with arbitrary postures, views and backgrounds	10,000	2015
	LIP [17, 18]	Focus on semantic understanding of person and contains images annotated with 19 semantic human part labels and 2D human poses with 16 key points.	28,714	2019
	Multi-Human Parsing: MHP [19]	Annotated images with 58 fine-grained semantic categories: 11 body parts and 47 clothes categories	25,403	2018
	CIHP [20]	Multi-person images and pixel-wise annotations in instance-level	28,280	2019
	ModaNet [113]	Annotated with pixel-level labels, bounding boxes, and polygons.	55,176	2019
	DeepFashion2 [77]	Diverse images of 13 popular clothing categories labeled with scale, occlusion, zoom-in, viewpoint, category, style, bounding box, dense landmarks and per-pixel mask	491,000	2019
Pose Transfer	Human3.6M [114]	Provides synchronized 2D and 3D data, accurate 3D human models and mixed reality settings	3.6M	2014
	Market-1501 [115]	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.	32,668	2015
	DeepFashion [72]	In-shop Clothes Retrieval Benchmark DeepFashion is used for pose transfer	52,712	2016

there are multiple datasets and the popular one is the LIP dataset [17, 18]. Datasets for physical simulation are different from other fashion tasks since the physical simulation is more related to computer graphics than computer vision. Physical simulation working within the fashion domain focus on clothing-body interactions, and datasets can be categorized into real data and created data.

Despite the rapid revolution on previous datasets which are based on 2D images like DeepFashion [72], DeepFashion2 [77] and FashionAI [116], the production of datasets basing on 3D clothing is almost rare or not sufficient for training like the digital wardrobe released by MGN [117]. In 2020, Heming et al. [118] develop a comprehensive dataset named Deep Fashion3D 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. 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 in order 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 [8] and LIP [17, 18] 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.

5.2.2. Human pose Estimation

Research in HPE has made significant progress during the last years. In this section, we present the most important evaluation metrics which are needed to measure the performance of human pose estimation models. Table 3 presents these different metrics used for comparisons of the existing state-of the-art approaches.

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. Table 4 and Fig. 12 presented examples of these performances results.

Table 2: Performance comparisons of fashion parsing methods (in %) [6].

Method	Dataaset	Evaluation Metrics							
		mIOU	aPA	mAGR	Acc	Fg.acc	Avg.prec	Avg.recall	AVG.F-1
Yamaguchi et al. [8]	SYSU-clothes [13]	-	85.97	51.25	-	-	-	-	-
CCP [13]		-	88.23	63.89	-	-	-	-	-
Yamaguchi et al. [8]	Fashionista [8]	-	89.00	64.37	-	-	-	-	-
CCP [13]		-	90.29	65.52	-	-	-	-	-
Yamaguchi et al. [8]	ATR [14]	-	-	-	88.96	62.18	52.75	49.43	-44.76
Liang et al. [17]		-	-	-	91.11	71.04	71.69	60.5	64.38
Co-CNN [14]		-	-	-	96.02	83.57	84.95	77.66	80.14
Yamaguchi et al. [8]	Fashionista [8]	-	-	-	89.98	65.66	54.87	51.16	46.80
Liang et al. [17]		-	-	-	92.33	76.54	73.93	66.49	69.30
Co-CNN [14]		-	-	-	97.06	89.15	87.83	81.73	83.78
MuLA [119]	LIP [18]	49.30	-	-	60.50	-	-	-	-
CE2P [21]		53.10	-	-	60.50	-	-	-	-
Wang et al. [23]		57.74	-	-	68.80	-	-	-	-
MuLA [119]	PASCAL-Person-	65.10	-	-	-	-	-	-	-
PGN [22]	Part [120]	68.40	-	-	-	-	-	-	-
Wang et al. [23]		70.76	-	-	-	-	-	-	-
Co-CNN [14]	ATR [14]	-	96.02	-	-	83.57	84.95	77.66	80.14
TGPNet [121]		-	96.45	-	-	87.91	83.36	80.22	81.76
Wang et al. [23]		-	96.26	-	-	87.91	84.62	86.41	85.51
Deeplab [23]	CFD+	-	87.68	-	-	56.08	35.35	39.00	37.09
TGPNet [121]	Fasfionista+CCP	-	91.25	-	-	66.37	50.71	53.18	51.92
Wang et al. [23]		-	92.20	-	-	68.59	56.84	59.47	58.12

Table 3: Summary of commonly used evaluation metrics for HPE [122]

Method	Meaning	Dataset	Description
Single Person			
PCP	Percentage of Correct Parts	LSP [26]	Percentage of correct predicted Parts which their end points fall within a threshold
PCK	Percentage of Correct Keypoints	LSP [26] MPII [25]	Percentage of correct predicted joints which fall within a threshold
Multiple Person			
AP	Average Precision	MPII [25]	mean AP (mAP) is reported by AP for each body part after assigning predicted pose to the ground truth pose by PCKh score.
		COCO [109]	- AP coco: at OKS=.50:.05:.95 (primary metric) - $AP_{coco}^{OKS=.50}$: $at\ OKS=.50(loose\ metric)$ - $AP_{coco}^{OKS=.75}$: $at\ OKS=.75(strict\ metric)$ - AP_{coco}^{medium} : $for\ medium\ objects: 32^2 < area < 96^2$ - AP_{coco}^{large} : $for\ large\ objects: area > 96^2$
AR	Average Recall	COCO [109]	- AR coco: at OKS=.50:.05:.95 - $AR_{coco}^{OKS=.50}$: $at\ OKS=.50$ - $AR_{coco}^{OKS=.75}$: $at\ OKS=.75$ - AR_{coco}^{medium} : $for\ medium\ objects: 32^2 < area < 96^2$. - AR_{coco}^{large} : $for\ large\ objects: area > 96^2$
OKS	Object Keypoint Similarity	COCO [109]	A similar role as the Intersection over Union (IoU) for AP/AR.

Table 4: Performance comparisons of fashion landmark detection methods in terms of NE [73].

Method	Dataset	Evaluation Metrics								
		L.Collar	R.Collar	L.Sleeve	R.Sleeve	R.Waistline	R.Waistline	L.Hem	R.Hem	Avg.
DFA [70]	DeepFashion [72]	0.0628	0.0637	0.0658	0.0621	0.0726	0.0702	0.0658	0.0663	0.0660
DLAN [73]		0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
AttentiveNet [74]		0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	0.0551	0.0484
Global-Local [75]		0.0312	0.0324	0.0427	0.0434	0.0361	0.0373	0.0442	0.0475	0.0393
DFA [70]	FLD [70]	0.0480	0.0480	0.0910	0.0890	-	-	0.0710	0.0720	0.0680
DLAN [73]		0.0531	0.0547	0.0705	0.0735	0.0752	0.0748	0.0693	0.0675	0.0672
AttentiveNet [74]		0.0463	0.0471	0.0627	0.0614	0.0635	0.0692	0.0635	0.0527	0.0583
Global-Local [75]		0.0386	0.0391	0.0675	0.0672	0.0576	0.0605	0.0615	0.0621	0.0568

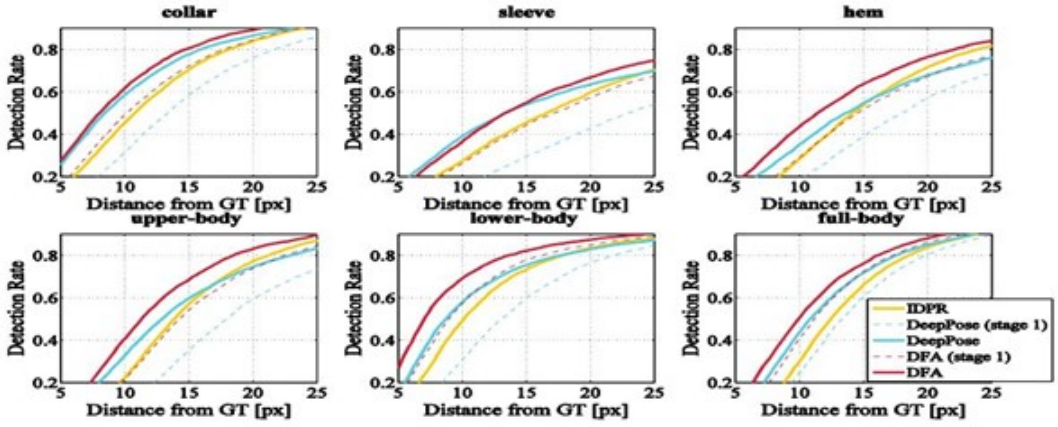

Fig. 12. Performance of fashion landmark detection: First row on different fashion landmarks and second row on different clothing types [70].

5.3. IQA for Fashion synthesis

The image quality evaluation is essential for image generation methods to synthesize desired outputs. Recent image synthesis research [96, 102, 123, 124] 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. Also, there are objective comparisons for virtual try-on, in terms of inception score (IS) or structural similarity (SSIM). IS [125] is used to evaluate the synthesis quality of images quantitatively. SSIM [126] 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 [126–129]. Table 5 shows evaluation metrics including Structural Similarity (SSIM) [126], Inception Score (IS) [125], masked version of Structural Similarity (mask-SSIM) [126], masked version of Inception Score (mask-IS) [125] and Detection Score (DS) [96] applied on Market-1501 dataset [115] and DeepFashion dataset [72].

Table 5: Results of different state-of-the-art methods for fashion parsing [129].

Model	Market-1501						DeepFashion [72]			
	SSIM	IS	Mask-SSIM	Mask-IS	DS	pSSIM	SSIM	IS	DS	pSSIM
PG2 [95]	0.261	3.495	0.782	3.367	0.390	-	0.773	3.163	0.951	-
Def-GAN [96]	0.291	3.230	0.807	3.502	0.720	-	0.760	3.362	0.976	-
PATN [128]	0.81	3.162	0.799	3.737	0.796	0.6186	0.771	3.201	0.976	0.799
Loss function [129]	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

5.3.2. 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. Fig. 13 presents an example of these comparisons.

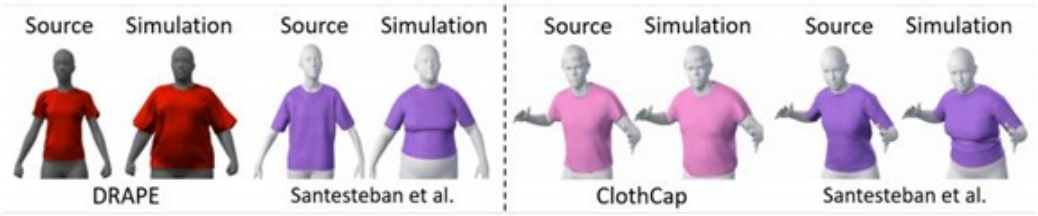

Fig. 13. Evaluation of the work of Santesteban et al. [130] compared with DRAPE [104] et ClothCap [105].

6. Application and future work

Automate the manual processes is a great achievement ensured by new technologies such as computer vision. One of the popular industries that is influenced by technology advancement at a much faster speed than ever before is Fashion. Due to computer vision powered tools, a great experience can be born for both retailers and consumers. In the following, we discuss emerging uses of fashion technology in some application areas and present future works needed to achieve the promised benefits.

6.1. Application

Apparel industry is all about visual and computer vision can recognize images just as we do by making computers understand images. Thus, creating AI systems that can understand fashion in images, can have a big impact on the industry and create a next-level customer experience like online fashion shopping. Here is where the future research work, in this area, will bring value and become useful for fashion business by making smart shopping.

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. Large fashion brands implemented online virtual fitting rooms in a bid 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 especially when we use an image-based approach to transfer clothes.

6.2. Future Directions

Despite the great development of image-based fitting systems, there remain some unresolved challenges and gap between research and practical applications such as the influence of body part occlusion and crowded people. Therefore, there are still many challenges in adopting fashion technologies in industry because real-world fashion is much more complex than in the experiments.

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.

Network efficiency 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. Data collection for specific complex scenarios is an option and there are other ways to extend existing datasets. Synthetic technology can theoretically generate unlimited data while there is a domain gap between synthetic data and real data. Cross-dataset supplementation, especially to supplement 3D datasets with 2D datasets, can mitigate the problem of insufficient diversity of training data. Transfer learning proves to be useful in this application.

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 that can understand fashion in images. 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 clothes between different images of persons and deal with the large variations on body poses and shapes.

7. Conclusion

With the explosive growth of clothing images, its study has attracted more attention of researchers to develop applications based on clothing models. The future directions must bridge the gap between research and real industry demand. Given the huge profit potential in the fashion industry, the representative intelligent fashion analysis techniques surveyed here are just the beginning of this expanding research field because up to now, enormous research efforts have been spent on these tasks. In this direction, our future work will exploit the use of AI to develop virtual try-on system and overcome the challenges ranging from most of the important topics in computer vision domain, especially, the techniques used in virtual fitting like fashion detection and fashion synthesis.

References

[1] Statista, Global fashion industry statistics - international apparel, https://fashionunited.com/global-fashion-industry-statistics/ (2019) (accessed 04 January 2019).

[2] Statista, Fashion - worldwide, https://www.statista.com/outlook/244/100/fashion/worldwide (2019) (accessed 04 January 2019).

[3] S. Liu, L. Liu, S. Yan, Fashion analysis: Current techniques and future directions, IEEE MultiMedia 21 (2) (2014) 72–79.

[4] S. Song, T. Mei, When multimedia meets fashion, IEEE MultiMedia 25 (3) (2018) 102–108.

[5] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, J. Liu, Fashion meets computer vision: A survey, arXiv preprint arXiv:2003.13988.

[6] T. E. Johnsen, J. Miemczyk, M. Howard, A systematic literature review of sustainable purchasing and supply research: Theoretical perspectives and opportunities for imp-based research, Industrial Marketing Management 61 (2017) 130–143.

[7] W. Ji, X. Li, Y. Zhuang, O. Bourahla, Y. Ji, S. Li, J. Cui, Semantic locality-aware deformable network for clothing segmentation, 2018, pp. 764–770. doi:10.24963/ijcai.2018/106.

[8] K. Yamaguchi, M. Kiapour, L. Ortiz, T. Berg, Parsing clothing in fashion photographs, 2012, pp. 3570–3577. doi:10.1109/CVPR.2012.6248101.

[9] K. Yamaguchi, M. Kiapour, T. Berg, Paper doll parsing: Retrieving similar styles to parse clothing items, 2013, pp. 3519–3526. doi:10.1109/ICCV.2013.437.

[10] K. Yamaguchi, M. Kiapour, L. Ortiz, T. Berg, Retrieving similar styles to parse clothing, Pattern Analysis and Machine Intelligence, IEEE Transactions on 37 (2015) 1028–1040. doi:10.1109/TPAMI.2014.2353624.

[11] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, S. Yan, Fashion parsing with weak color-category labels, Multimedia, IEEE Transactions on 16 (2014) 253–265. doi:10.1109/TMM.2013.2285526.

[12] J. Dong, Q. Chen, W. Xia, Z. Huang, S. Yan, A deformable mixture parsing model with parselets, 2013, pp. 3408–3415. doi:10.1109/ICCV.2013.423.

[13] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, S. Yan, Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval, IEEE Transactions on Multimedia 18 (2016) 1–1. doi:10.1109/TMM.2016.2542983.

[14] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, S. Yan, Human parsing with contextualized convolutional neural network, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2016) 1–1. doi:10.1109/TPAMI.2016.2537339.

[15] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE transactions on pattern analysis and machine intelligence 37 (12) (2015) 2402–2414.

[16] L. Liao, X. He, B. Zhao, C.-W. Ngo, T.-S. Chua, Interpretable multimodal retrieval for fashion products, 2018, pp. 1571–1579. doi:10.1145/3240508.3240646.

[17] K. Gong, X. Liang, D. Zhang, X. Shen, L. Lin, Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing, 2017, pp. 6757–6765. doi:10.1109/CVPR.2017.715.

[18] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: Joint body parsing pose estimation network and a new benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (2018) 1–1. doi:10.1109/TPAMI.2018.2820063.

[19] J. Zhao, J. Li, H. Liu, S. Yan, J. Feng, Fine-grained multi-human parsing, International Journal of Computer Vision 128 (8) (2020) 2185–2203.

[20] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, J. Feng, Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing, 2018, pp. 792–800. doi:10.1145/3240508.3240509.

[21] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, Y. Zhao, Devil in the details: Towards accurate single and multiple human parsing, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 4814–4821. doi:10.1609/aaai.v33i01.33014814.

[22] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, L. Lin, Instance-level human parsing via part grouping network, 2018.

[23] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, L. Shao, Learning compositional neural information fusion for human parsing, 2020.

[24] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, L. Lin, Graphonomy: Universal human parsing via graph transfer learning, 2019, pp. 7442–7451. doi:10.1109/CVPR.2019.00763.

[25] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, 2014. doi:10.1109/CVPR.2014.471.

[26] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, 2010, pp. 1–11. doi:10.5244/C.24.12.

[27] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Human pose estimation using body parts dependent joint regressors, 2013, pp. 3041–3048. doi:10.1109/CVPR.2013.391.

[28] X. Chen, A. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, arXiv preprint arXiv:1407.3399.

[29] G. Gkioxari, B. Hariharan, R. Girshick, J. Malik, Using k-poselets for detecting people and localizing their keypoints, 2014, pp. 3582–3589. doi:10.1109/CVPR.2014.458.

[30] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, C. Bregler, Learning human pose estimation features with convolutional networks, arXiv preprint arXiv:1312.7302.

[31] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, 2014, pp. 2337–2344. doi:10.1109/CVPR.2014.299.

[32] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[33] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (6) (2017) 84–90.

[34] T. Pfister, K. Simonyan, J. Charles, A. Zisserman, Deep convolutional neural networks for efficient pose estimation in gesture videos, 2014. doi:10.1007/978-3-319-16865-4_35.

[35] D. Luvizon, H. Tabia, D. Picard, Human pose regression by combining indirect part detection and contextual information, Computers Graphicsdoi:10.1016/j.cag.2019.09.002.

[36] A. Nibali, Z. He, S. Morgan, L. Prendergast, Numerical coordinate regression with convolutional neural networks, arXiv preprint arXiv:1801.07372.

[37] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2602–2611.

[38] Z.-Q. Liu, Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network, 2014.

[39] G. Gkioxari, B. Hariharan, R. Girshick, J. Malik, R-cnns for pose estimation and action detection, arXiv preprint arXiv:1406.5212.

[40] D. C. Luvizon, D. Picard, H. Tabia, 2d/3d pose estimation and action recognition using multitask deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5137–5146.

[41] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, Vol. 9912, 2016, pp. 483–499. doi:10.1007/978-3-319-46484-8_29.

[42] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, arXiv preprint arXiv:1406.2984.

[43] A. Jain, J. Tompson, Y. Lecun, C. Bregler, Modeep: A deep learning framework using motion features for human pose estimation, 2014. doi:10.1007/978-3-319-16808-1_21.

[44] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, 2017, pp. 3711–3719. doi:10.1109/CVPR.2017.395.

[45] U. Rafi, B. Leibe, J. Gall, I. Kostrikov, An efficient convolutional network for human pose estimation, 2016, pp. 109.1–109.11. doi:10.5244/C.30.109.

[46] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, 2018. doi:10.1007/978-3-030-01231-1_29.

[47] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 648–656.

[48] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, Vol. 9911, 2016. doi:10.1007/978-3-319-46478-7_44.

[49] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, 2016. doi:10.1109/CVPR.2016.511.

[50] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: proceedings of the IEEE international conference on computer vision, 2017, pp. 1281–1290.

[51] V. Belagiannis, A. Zisserman, Recurrent human pose estimation, 2017, pp. 468–475. doi:10.1109/FG.2017.64.

[52] S. ke, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, 2019, pp. 5686–5696. doi:10.1109/CVPR.2019.00584.

[53] W. Yang, W. Ouyang, H. Li, X. Wang, End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, 2016, pp. 3073–3082. doi:10.1109/CVPR.2016.335.

[54] G. Gkioxari, A. Toshev, N. Jaitly, Chained predictions using convolutional neural networks, in: European Conference on Computer Vision, Springer, 2016, pp. 728–743.

[55] X. Chu, W. Ouyang, H. Li, X. Wang, Structured feature learning for pose estimation, 2016, pp. 4715–4723. doi:10.1109/CVPR.2016.510.

[56] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831–1840.

[57] G. Ning, Z. Zhang, Z. He, Knowledge-guided deep fractal neural networks for human pose estimation, IEEE Transactions on Multimedia 20 (5) (2017) 1246–1259.

[58] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 713–728.

[59] W. Tang, P. Yu, Y. Wu, Deeply Learned Compositional Models for Human Pose Estimation: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III, 2018, pp. 197–214. doi:10.1007/978-3-030-01219-9_12.

[60] W. Tang, Y. Wu, Does learning specific features for related parts help human pose estimation?, 2019, pp. 1107–1116. doi:10.1109/CVPR.2019.00120.

[61] C.-J. Chou, J.-T. Chien, H.-T. Chen, Self adversarial training for human pose estimation, 2018, pp. 17–30. doi:10.23919/APSIPA.2018.8659538.

[62] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, 2017, pp. 1221–1230. doi:10.1109/ICCV.2017.137.

[63] X. Peng, Z. Tang, F. Yang, R. Feris, D. Metaxas, Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation, 2018. doi:10.1109/CVPR.2018.00237.

[64] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, L. Lin, Lstm pose machines, 2018, pp. 5207–5215. doi:10.1109/CVPR.2018.00546.

[65] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, D. Metaxas, Quantized densely connected u-nets for efficient landmark localization (08 2018).

[66] F. Zhang, X. Zhu, M. Ye, Fast human pose estimation, 2019, pp. 3512–3521. doi:10.1109/CVPR.2019.00363.

[67] J. Li, W. Su, Z. Wang, Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 11354–11361. doi:10.1609/aaai.v34i07.6797.

[68] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, 2017, pp. 1302–1310. doi:10.1109/CVPR.2017.143.

[69] X. Nie, J. Feng, J. Xing, S. Yan, Pose Partition Networks for Multi-person Pose Estimation: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V, 2018, pp. 705–720. doi:10.1007/978-3-030-01228-1_42.

[70] Z. Liu, S. Yan, P. Luo, X. Wang, X. Tang, Fashion landmark detection in the wild, Vol. 9906, 2016, pp. 229–245. doi:10.1007/978-3-319-46475-6_15.

[71] Y. Li, S. Tang, Y. Ye, J. Ma, Spatial-aware non-local attention for fashion landmark detection, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 820–825.

[72] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, 2016, pp. 1096–1104. doi:10.1109/CVPR.2016.124.

[73] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Unconstrained fashion landmark detection via hierarchical recurrent transformer networks, 2017, pp. 172–180. doi:10.1145/3123266.3123276.

[74] W. Wang, Y. Xu, J. Shen, S. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, 2018. doi:10.1109/CVPR.2018.00449.

[75] S. Lee, S. Oh, C. Jung, C. Kim, A global-local embedding module for fashion landmark detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[76] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (2018) 1–1. doi:10.1109/TPAMI.2018.2844175.

[77] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, 2019, pp. 5332–5340. doi:10.1109/CVPR.2019.00548.

[78] P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, Image-to-image translation with conditional adversarial networks, 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632.

[79] S. Jiang, Y. Fu, Fashion style generator, 2017, pp. 3721–3727. doi:10.24963/ijcai.2017/520.

[80] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, J. Hays, Texturegan: Controlling deep image synthesis with texture patches, 2018, pp. 8456–8465. doi:10.1109/CVPR.2018.00882.

[81] X. Han, Z. Wu, Z. Wu, R. Yu, L. Davis, Viton: An image-based virtual try-on network, 2018, pp. 7543–7552. doi:10.1109/CVPR.2018.00787.

[82] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, Toward characteristic-preserving image-based virtual try-on network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 589–604.

[83] X. Han, Z. Wu, W. Huang, M. R. Scott, L. S. Davis, Compatible and diverse fashion image inpainting, arXiv preprint arXiv:1902.01096.

[84] S. Zhu, S. Fidler, R. Urtasun, D. Lin, C. C. Loy, Be your own prada: Fashion synthesis with structural coherence, 2017, pp. 1689–1697. doi:10.1109/ICCV.2017.186.

[85] Z. Wu, G. Lin, Q. Tao, J. Cai, M2e-try on net: Fashion from model to everyone, 2019, pp. 293–301. doi:10.1145/3343031.3351083.

[86] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, W.-H. Cheng, Fit-me: Image-based virtual try-on with arbitrary poses, 2019, pp. 4694–4698. doi:10.1109/ICIP.2019.8803681.

[87] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, W.-H. Cheng, Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information, 2019, pp. 275–283. doi:10.1145/3343031.3351075.

[88] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, J. Lu, SwapNet: Image Based Garment Transfer: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII, 2018, pp. 679–695. doi:10.1007/978-3-030-01258-8_41.

[89] R. Yu, X. Wang, X. Xie, Vtnfp: An image-based virtual try-on network with body and clothing feature preservation, 2019, pp. 10510–10519. doi:10.1109/ICCV.2019.01061.

[90] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, L. Nie, Virtually trying on new clothing with arbitrary poses, 2019, pp. 266–274. doi:10.1145/3343031.3350946.

[91] X. Han, W. Huang, X. Hu, M. Scott, Clothflow: A flow-based model for clothed person generation, 2019, pp. 10470–10479. doi:10.1109/ICCV.2019.01057.

[92] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, J. Yin, Fw-gan: Flow-navigated warping gan for video virtual try-on, 2019, pp. 1161–1170. doi:10.1109/ICCV.2019.00125.

[93] F. Sun, J. Guo, Z. Su, C. Gao, Image-based virtual try-on network with structural coherence, 2019, pp. 519–523. doi:10.1109/ICIP.2019.8803811.

[94] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, S. Alpert, Image based virtual try-on network from unpaired data, 2020, pp. 5183–5192. doi:10.1109/CVPR42600.2020.00523.

[95] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, arXiv preprint arXiv:1705.09368.

[96] A. Siarohin, E. Sangineto, S. Lathuiliere, N. Sebe, Deformable gans for pose-based human image generation, 2018, pp. 3408–3416. doi:10.1109/CVPR.2018.00359.

[97] G. Balakrishnan, A. Zhao, A. Dalca, F. Durand, J. Guttag, Synthesizing images of humans in unseen poses, 2018, pp. 8340–8348. doi:10.1109/CVPR.2018.00870.

[98] C. Si, W. Wang, L. Wang, T. Tan, Multistage adversarial losses for pose-based human image synthesis, 2018, pp. 118–126. doi:10.1109/CVPR.2018.00020.

[99] A. Pumarola, A. Agudo, A. Sanfeliu, F. Moreno-Noguer, Unsupervised person image synthesis in arbitrary poses, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8620–8628.

[100] J.-Y. Zhu, T. Park, P. Isola, A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017, pp. 2242–2251. doi:10.1109/ICCV.2017.244.

[101] P. Esser, E. Sutter, A variational u-net for conditional appearance and shape generation, 2018, pp. 8857–8866. doi:10.1109/CVPR.2018.00923.

[102] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, M. Fritz, Disentangled person image generation, 2018, pp. 99–108. doi:10.1109/CVPR.2018.00018.

[103] S. Song, W. Zhang, J. Liu, T. Mei, Unsupervised person image generation with semantic parsing transformation, 2019, pp. 2352–2361. doi:10.1109/CVPR.2019.00246.

[104] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, M. Black, Drape: Dressing any person, ACM Transactions on Graphics 31 (2012) 1–10. doi:10.1145/2185520.2335386.

[105] G. Pons-Moll, S. Pujades, S. Hu, M. Black, Clothcap: Seamless 4d clothing capture and retargeting, ACM Transactions on Graphics 36 (2017) 1–15. doi:10.1145/3072959.3073711.

[106] H. Wang, J. O'Brien, R. Ramamoorthi, Data-driven elastic models for cloth: Modeling and measurement, ACM Trans. Graph. 30 (2011) 71. doi:10.1145/2010324.1964966.

[107] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. Berg, M. C. Lin, Detailed garment recovery from a single-view image, arXiv preprint arXiv:1608.01250.

[108] D. Yoo, N. Kim, S. Park, A. Paek, I. Kweon, Pixel-level domain transfer, Vol. 9912, 2016, pp. 517–532. doi:10.1007/978-3-319-46484-8_31.

[109] D. Puri, Coco dataset stuff segmentation challenge, 2019, pp. 1–5. doi:10.1109/ICCUBEA47591.2019.9129255.

[110] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al., Ai challenger: A large-scale dataset for going deeper in image understanding, arXiv preprint arXiv:1711.06475.

[111] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, B. Schiele, Posetrack: A benchmark for human pose estimation and tracking, 2018, pp. 5167–5176. doi:10.1109/CVPR.2018.00542.

[112] A. Dalmia, S. Joshi, R. Singh, V. Raykar, Styling with attention to details, arXiv preprint arXiv:1807.01182.

[113] S. Zheng, F. Yang, M. Kiapour, R. Piramuthu, Modanet: A large-scale street fashion dataset with polygon annotations, 2018, pp. 1670–1678. doi:10.1145/3240508.3240652.

[114] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, 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) (2013) 1325–1339.

[115] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, 2015, pp. 1116–1124. doi:10.1109/ICCV.2015.133.

[116] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, Y. Cao, Fashionai: A hierarchical dataset for fashion understanding, 2019, pp. 296–304. doi:10.1109/CVPRW.2019.00039.

[117] B. Bhatnagar, G. Tiwari, C. Theobalt, G. Pons-Moll, Multi-garment net: Learning to dress 3d people from images, 2019, pp. 5419–5429. doi:10.1109/ICCV.2019.00552.

[118] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, X. Han, Deep Fashion3D: A Dataset and Benchmark for 3D Garment Reconstruction from Single Images, 2020, pp. 512–530. doi:10.1007/978-3-030-58452-8_30.

[119] X. Nie, J. Feng, S. Yan, Mutual Learning to Adapt for Joint Human Parsing and Pose Estimation: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V, 2018, pp. 519–534. doi:10.1007/978-3-030-01228-1_31.

[120] F. Xia, P. Wang, X. Chen, A. L. Yuille, Joint multi-person pose estimation and semantic part segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6769–6778.

[121] X. Luo, Z. Su, J. Guo, G. Zhang, X. He, Trusted guidance pyramid network for human parsing, 2018, pp. 654–662. doi:10.1145/3240508.3240634.

[122] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Computer Vision and Image Understanding 192 (2020) 102897.

[123] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, J. Yin, Towards multi-pose guided virtual try-on network, 2019, pp. 9025–9034. doi:10.1109/ICCV.2019.00912.

[124] P. Costa, A. Galdran, M. I. Meyer, M. D. Abramoff, M. Niemeijer, A. M. Mendonça, A. Campilho, Towards adversarial retinal image synthesis, arXiv preprint arXiv:1701.08974.

[125] A. Hore, D. Ziou, Image quality metrics: Psnr vs. ssim, in: 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 2366–2369.

[126] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[127] A. Siarohin, E. Sangineto, S. Lathuiliere, N. Sebe, Deformable gans for pose-based human image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3408–3416.

[128] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, X. Bai, Progressive pose attention transfer for person image generation, 2019, pp. 2342–2351. doi:10.1109/CVPR.2019.00245.

[129] H. Shi, L. Wang, W. Tang, N. Zheng, G. Hua, Loss functions for person image generation, in: BMVC, 2020.

[130] I. Santesteban, M. Otaduy, D. Casas, Learning-based animation of clothing for virtual try-on, Computer Graphics Forum 38 (2019) 355–366. doi:10.1111/cgf.13643.