IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2023.0322000*

# Deep Learning in Virtual Try-On: A Comprehensive Survey

**TASIN ISLAM[1], ALINA MIRON[1], XIAOHUI LIU[1] and YONGMIN LI[1]***

[1]Brunel University London, Kingston Ln, London, Uxbridge UB8 3PH UK (e-mail:{tasin.islam2,alina.miron,xiaohui.liu,yongmin.li}@brunel.ac.uk)

Corresponding author: Yongmin (e-mail: yongmin.li@brunel.ac.uk).

**ABSTRACT** Virtual try-on technology has gained significant importance in the retail industry due to its potential to transform the way customers interact with products and make purchase decisions. It allows users to virtually try on clothing and accessories, providing a realistic representation of how the items would look and fit without the need for physical interaction. The ability to virtually try on products addresses common challenges associated with online shopping, such as uncertainty about fit and style, ultimately enhancing the overall customer experience and satisfaction. As a result, virtual try-on technology has the potential to reduce returns and optimise conversion rates for businesses, making it a valuable tool in the e-commerce landscape. In this paper, we provide a comprehensive review of deep learning based virtual try-on models, focusing on their functionality, technical details, dataset usage, weaknesses, and impact on customer satisfaction. The models are categorised into three main types: image-based, multi-pose, and video virtual try-on models, with detailed examples and technical summaries provided for each category. Additionally, we identify and discuss similarities and differences in these methods. Furthermore, we examine the datasets currently available for building and evaluating virtual try-on models, including the number of images/videos and their resolutions. We present the commonly used methods for both qualitative and quantitative evaluations, comparing synthesised images with previous work and performing quantitative evaluations across various metrics and benchmark datasets. We discuss the weaknesses of current deep learning based virtual try-on models, including challenges in preserving clothing characteristics and textures, the level of accuracy of applying the clothing to the person, and the preservation of facial identities. Additionally, we address dataset bias, particularly the domination of female models, limited diversity in clothing featured, and relatively simple and clean backgrounds in the datasets, which can negatively impact the model's ability to handle challenging situations. Moreover, we explore the impact of virtual try-ons on customer satisfaction, highlighting the benefits that customers can enjoy, which also reduces returns and optimises conversion rates for businesses.

**INDEX TERMS** Virtual Try-On (VTO), Deep Learning, Image Synthesis, Generative Adversarial Networks (GANs), Diffusion Models (DMs)

## I. INTRODUCTION

The online fashion industry is experiencing a remarkable boom, with a larger number of individuals embracing online shopping like never before [4]. Particularly during the COVID-19 pandemic, purchasing clothes online has become a widespread trend. This phenomenon reflects how garment shopping has evolved for numerous consumers. As the digital sphere continues to expand with the advent of e-commerce and online shopping, there has been a surge of interest in exploring methods to provide consumers with the same experience they would receive when shopping in physical stores [13].

The inability to try on garments poses a severe obstacle to online purchases as consumers would not know if the product would suit, fit and match them [142]. In order to succeed and boost sales, retailers must adapt and cater to the increasing number of online consumers. Virtual try-on provides a means for consumers to engage virtually with the product, facilitating a connection between businesses and online consumers. Embracing virtual try-on technology allows businesses to thrive in the online realm and effectively cater to the needs of their customer [87], [102].

Virtual try-on allows for the seamless exploration of various items such as shoes [34], [198], upper clothing [68], [197], lower clothing [59], [191] and more [99], [129]. Some models even offer the convenience of modifying multiple

items in a single image simultaneously [201]. This technology empowers consumers by providing an immersive experience and complete freedom in decision-making, enabling them to experiment with various products, resulting in a confident and personalised purchase. Using this technology can enhance the online shopping experience by resembling an atmosphere where consumers feel they are shopping in a physical store.

Several prominent brands have embraced virtual try-on technology as part of their online platforms. For instance, RayBan [1] enables consumers to try on glasses virtually using their webcam's live feed. Similarly, L'Oréal [3] allows consumers to experiment with different makeup options and change their hair colour by utilising a live feed from their webcam or by uploading an image. Hugo Boss [2] utilises a 3D avatar that allows customers to visualise how a desired garment would look when worn.

There are a few existing survey papers that discuss the potential use of artificial intelligence (AI) in supporting the fashion industry and facilitating fashion-related tasks [27], [60], [62], [134], as shown in TABLE 1. These literatures cover a wide range of topics, such as how AI can benefit the fashion industry, how fashion tools are developed and employed, which countries are leading in research on AI for fashion, how fashion data is utilised to enhance the effectiveness of AI models, and the classification of fashion-based AI tools.

However, there is a critical gap left by these existing reviews. They lack a focused examination of the technical details of deep learning based virtual try-on models and have not examined the effects and the impact that virtual try-on models have on a wider scale. Moreover, they only cover a few outdated papers, such as VITON [68] and CP-VTON [183], with brief explanations of how they work. None of the survey papers are dedicated solely to virtual try-ons. Instead, they incorporate a wider range of AI tools for the garment industry. We believe that virtual try-on models are among the most powerful AI fashion applications as they provide immediate personalised feedback to customers, and there is significant development in this field that is worth investigating. Therefore, we present a survey paper dedicated to these models and show their technical development and synthesise findings from various studies to offer a comprehensive understanding of how virtual try-on models can reshape online shopping experiences and drive business growth in the digital age.

We categorise virtual try-on methods based on common characteristics, providing a nuanced understanding of their functionality and efficacy. This in-depth categorisation sets our survey apart from existing works, which may not have delved as deeply into the specifics of virtual try-on technologies. Furthermore, our paper explores the impact of these models on consumers, shedding light on user experiences and perceptions—another dimension often left unexplored in prior surveys.

The key contributions of this study are as follows:

1) We fill a literature gap by conducting a comprehensive review of deep learning based virtual try-on models, focusing on three main types: image-based, multi-pose, and video virtual try-on models, with detailed examples and technical summaries provided for each category.
2) We examine the datasets currently available for building and evaluating virtual try-on models, considering the number of images/videos and their resolutions. Commonly used methods for both qualitative and quantitative evaluations are presented, including comparisons among models and quantitative assessments across various metrics and benchmark datasets.
3) We discuss the weaknesses of current deep learning based virtual try-on models, addressing challenges related to preserving clothing characteristics and textures, the accuracy of applying clothing to individuals, preserving facial identities, and dealing with dataset bias.
4) We explore the impact of virtual try-ons on customer satisfaction, emphasising the benefits that customers can enjoy.

## II. METHODOLOGY

We provide a detailed account of the methodology employed for conducting the Systematic Literature Review (SLR). The SLR aimed to comprehensively identify and analyse existing research studies relevant to deep learning-based virtual try-on models. The literature collection process was conducted in a systematic and rigorous manner to ensure the comprehensiveness of the search and the selection of relevant studies. A detailed search strategy was devised in consultation with subject matter experts to identify pertinent literature across multiple academic databases. The following steps outline the process:

1) **Identification of Databases:** We systematically searched key academic databases, including but not limited to IEEE Xplore Digital Library, ACM Digital Library, Scopus, Web of Science, and Google Scholar, to retrieve relevant publications before 2024.
2) **Search String Development:** A comprehensive search string was developed using a combination of relevant keywords and Boolean operators to capture a broad spectrum of literature related to deep learning-based virtual try-on models. The search string was iteratively refined through pilot searches to ensure its effectiveness in retrieving relevant studies.
3) **Inclusion and Exclusion Criteria:** Clear inclusion and exclusion criteria were established a priori to guide the selection of studies. Peer-reviewed and well-cited arXiv articles published in English were considered eligible for inclusion. Studies lacking empirical data or those not directly related to the scope of the research were excluded.
4) **Screening Process:** We have conducted the initial screening of retrieved studies based on title and abstract to assess their relevance to this topic. Then we conducted a thorough full-text assessment to determine

| Survey title | Authors | Year | Venue | Content | How it relates to virtual try-on |
|---|---|---|---|---|---|
| Smart fashion: a review of AI applications in virtual try-on & fashion synthesis | Mohammadi and Kalhor [134] | 2021 | Journal of Artificial Intelligence and Capsule Networks (AICN) | Classifies AI-based fashion applications into subcategories and notes that these tools are rapidly improving in this fast-growing area. | Lists some virtual try-on models without delving into technical details. They focus on the application side of it. |
| Fashion Meets Computer Vision: A Survey | Cheng et al. [27] | 2021 | ACM Computing Surveys (CSUR) | Highlights the abundance of online image data, which can be utilised to create smart fashion tools that provide suggestions and analysis for fashionable items. | Discusses a few virtual try-on models, their functionality and the weaknesses they face. |
| You Can Try Without Visiting: A Comprehensive Survey on Virtually Try-on Outfits | Ghodhbani et al. [60] | 2022 | Multimedia Tools and Applications | Discusses the technological advancements in the garment industry, briefly focusing on the development of virtual try-ons, challenges in modelling real-world problems, and the ongoing efforts to bridge the gap between research and industry demands in the online fashion industry. | Provides some examples of virtual try-on models, explains how they work and what researchers did to improve their model. |
| Augmented and virtual reality in apparel industry: A bibliometric review and future research agenda | Goel et al. [62] | 2023 | Foresight | Shows China leads in publishing papers on augmented and virtual reality in the apparel industry, followed by the USA and France, while the USA has received the highest number of citations. | Reveals that there is considerable interest in virtual try-ons and related domains, but the article does not specifically focus on the details of virtual try-ons. |

TABLE 1: Summary of related virtual try-on surveys.

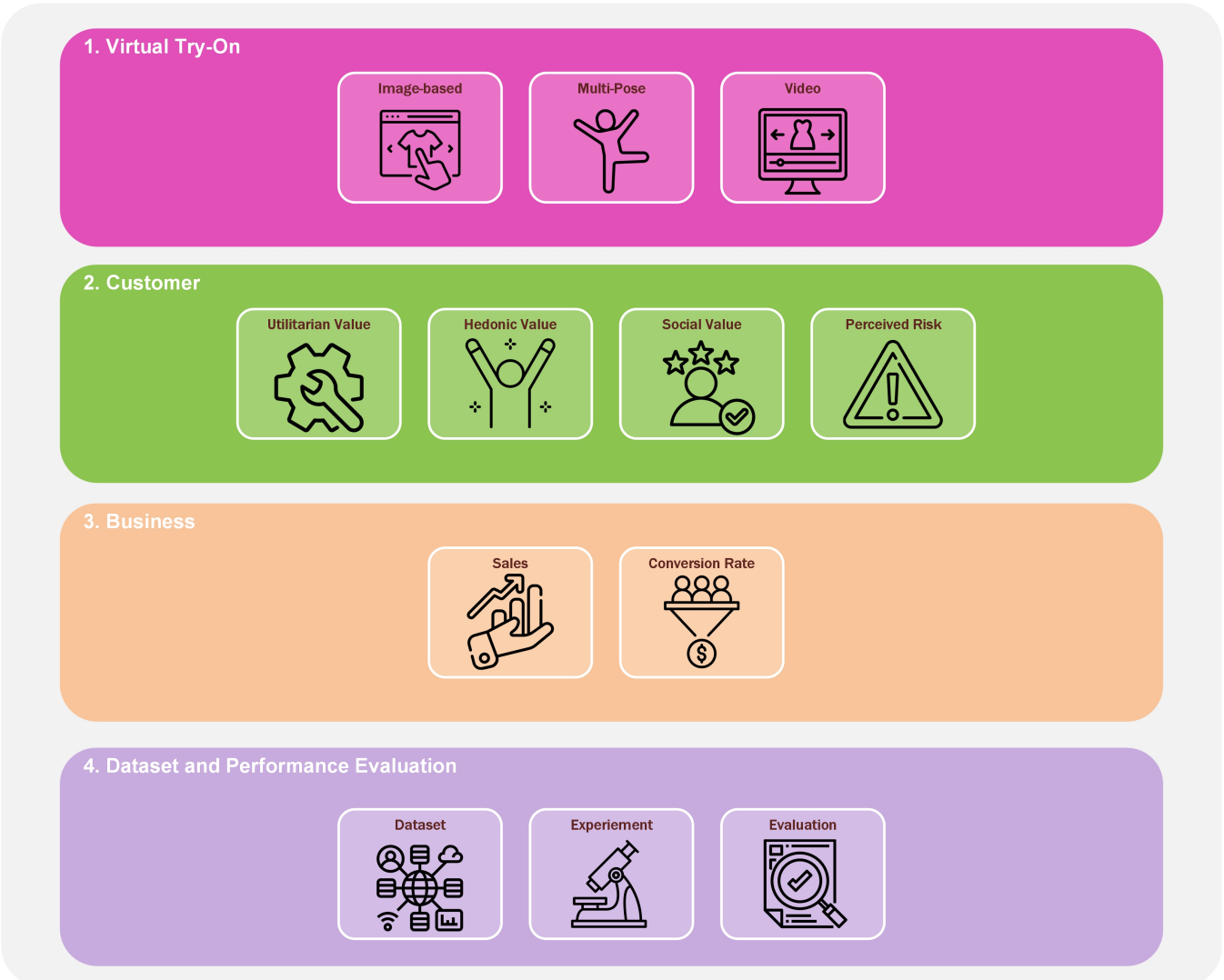



FIGURE 1: The scope of this study. Our survey paper explores image-based, multi-pose, and video virtual try-on models with a focus on technical details. We delve into how these models impact practical values like utility, enjoyment and social, as well as their perceived risks. We analyse how these models impact a business's sales and conversion rate. The paper also covers the datasets used, how experiments are conducted, and how these models are evaluated.

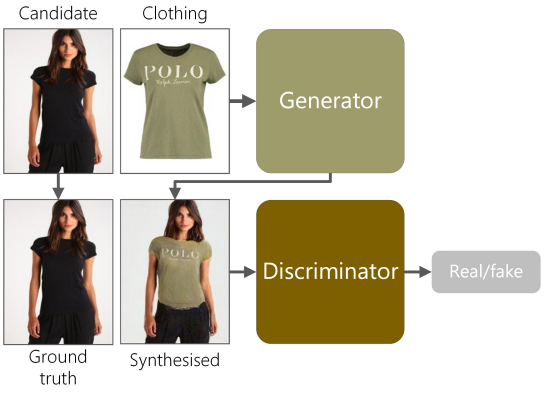



FIGURE 2: cGAN in virtual try-on. This variant of GAN takes in a conditional image to synthesise a desired output image. The discriminator is trained to distinguish real and fake samples, encouraging the generator to produce better results.

their suitability for inclusion based on our predefined criteria.

5) **Reference List Inspection:** The reference lists of included studies were manually inspected to identify additional relevant publications missed during the database search.

### A. SCOPE OF THE STUDY

FIGURE 1 illustrates the scope that will be included in this study. We focus on virtual try-on models that utilise deep learning methods. The deep learning based virtual try-on models are divided into image-based, multi-pose, and video virtual try-ons. We will investigate research papers examining how virtual try-on models influence utilitarian value, hedonic value, social value, and perceived risk. These factors determine and collectively shape customer satisfaction [52], [57]. We will investigate the impact it has on sales and conversion rates. Finally, we will explore the datasets used by virtual try-on models, their experiment procedures, and evaluation methods.

We will investigate how deep learning based virtual try-on systems can be categorised and what characteristics they have in common. We conduct a comprehensive review of existing literature related to virtual try-ons across the aforementioned categories. We will identify any patterns in the use of techniques within each category, which can help determine the direction of future development. After reviewing these studies, we discuss how these models measure performance and how they compare with one another.

In the following subsections, we will first provide a brief review of the two fundamental generative models that have been widely used in many virtual try-ons applications, the generative adversarial networks (GANs) and diffusion models.

### B. GENERATIVE ADVERSARIAL NETWORKS

GANs [63] leverages two neural networks to achieve high-quality image synthesis [85], [96], [97], [145] and manipulation [44], [92], [108]. The fundamental principle of GANs involves a generator network that attempts to deceive a discriminator network, which, in turn, learns to distinguish between real and fake samples.

In order to control the generated output images in the realm of GANs, the adoption of Conditional GAN (cGAN) [133] emerges as a promising solution. Various methodologies exist for guiding the image generation process within cGANs. Notable examples encompass the utilisation of class labels [17], [140], textual descriptions [143], [150], [153], [195], attributes [166], and sketches [85], [114]. These techniques enable cGANs to produce images aligned with specific criteria or desired characteristics. Consequently, the applications of cGANs, particularly in the domains of virtual try-on and fashion-related contexts [123], [125], have gained significant relevance. We show an example in FIGURE 2 of how cGAN uses conditional data to produce the desired output.

A majority of virtual try-on models have incorporated the utilisation of the GAN mechanism either as a whole or within specific modules [77], [197], [144], [202], [161]. By integrating GANs into the virtual try-on framework, these models have demonstrated the ability to generate try-on images with exceptional fidelity.

However, it is crucial to acknowledge the challenges associated with cGAN-based methods when confronted with substantial spatial deformations between the target clothing and the pose of the individual. Notably, CP-VTON [183] has demonstrated instances where cGAN-based approaches can exhibit unstable image generation under such conditions. Consequently, it becomes imperative to develop prerequisite methods that effectively guide cGANs during the image synthesis process, mitigating potential issues arising from large spatial deformations.

### C. DIFFUSION MODELS

In recent studies, the performance of diffusion models has surpassed that of GANs in the domain of image synthesis [40]. There are three formulations of diffusion models [37], which are denoising diffusion probabilistic model (DDPM) [75], noise conditioned score network (NCSN) [172] and stochastic differential equation (SDE) [173]. In principle, they all involve two Markov chains, as shown in FIGURE 3. The forward process is usually designed manually to convert any data distribution into Gaussian noise in gradual steps, while the reverse process uses deep neural networks to undo the Gaussian noise to synthesise an image incrementally.

Like GANs, output images of diffusion models can be controlled through textual descriptions [156] or input images [159]. This flexibility opens up various possibilities for their application in the field of fashion.

Integrating diffusion models into fashion synthesis is still a relatively new research direction, and only a limited number of studies have been conducted thus far [11], [19], [95], [135],

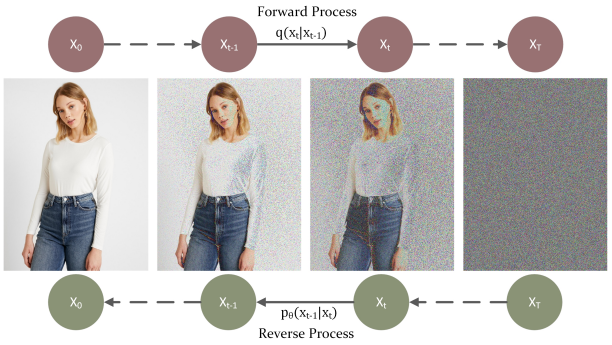



FIGURE 3: Diffusion models for image generation in virtual try-on. It involves two Markov chains: the forward process, which adds gradual noise to data, and the reverse process, which reverses the effects of the forward process using a neural network. $x_t$ denotes the current timestep of noise applied to an image. $x_0$ is the original image that a neural network is trained to gradually produce, while $x_T$ is the most perturbed image that the network starts with.


FIGURE 4: Results produced by the state-of-the-art HR-VTON [113], a single-pose image-based virtual try-on model that realistically replaces the candidate's garment with the product item.

[214]. However, given the promising results achieved and the potential of diffusion models in virtual try-on applications, it is plausible to expect that their utilisation in this area will become more prevalent in the near future.

## III. TYPES OF VIRTUAL TRY-ON

In this section, we organise virtual try-on models into distinct categories based on their common functionalities and characteristics. We discuss their respective architectures and provide a detailed explanation of how they operate. It is worth noting that certain models may exhibit overlapping functionalities and may be included in multiple categories to reflect their diverse traits.

### A. PHYSICS-BASED VIRTUAL TRY-ON

Virtual try-on models were first developed as physics-based simulations. These simulations involve using 3D data to create a virtual garment that fits onto a 3D avatar [65], [146], [148], [164]. Mostly mathematical models are used to manipulate clothing data and create realistic wrinkles [65], [148], [164], but newer models have started using neural networks [146]. The downside of these models is that physics-based algorithms are computationally expensive and difficult to control [146]. Additionally, they require either 3D scans of humans [164] or clothing [148], which is a time-consuming and impractical process.

In this paper, our emphasis will be on the deep learning-based models that have gained more popularity within the research community and shown greater promise in applications. Interested readers can refer to [27], [60], [62] for further details regarding the 3D virtual try-on models.

### B. IMAGE-BASED VIRTUAL TRY-ON

An image-based virtual try-on model takes input from images of both a user and a target clothing, and overlays virtual representations of the clothing onto the user's body. Typically all the input and output are 2D images. The criteria for an image-based virtual try-on model are:

- Preserving the posture and body shape of the person.
- Preserving clothing items that are not intended to be replaced.
- Ensuring the target clothing item fits well to the intended body part of the person.
- Retaining the texture and details of the garment.

We provide an example of an image-based virtual try-on in FIGURE 4, showcasing how these techniques can overlay virtual clothing onto individuals. In TABLE 2, we have listed all image-based virtual try-ons, highlighting the warping method they use and their contributions. Image-based virtual try-ons can be split into two categories: models that use a clean image of the clothing item and models that transfer a garment from one person to another.

CAGAN [89] is the first image-based virtual try-on model. This model employs a single network that combines various images to create a try-on image. Nonetheless, an important drawback of this method is its dependence on both the target clothing and the original garment images during the inference phase. Reliance on the original clothing image poses a significant constraint because users may not be able to provide it. VITON [68] has made this more practical by employing an approach that utilises only the image of target clothing.

Many researchers focus only on replacing a person's upper garment. This is because they use a commonly available dataset that only contains pairs of candidates with their upper clothes. By doing so, they can easily compare their work with previous research. However, some models are capable of changing any type of clothing, including trousers [112], [117], [118], [135]–[137].

Input data is crucial for virtual try-on models, which take into account not only the person and clothing images but also

pose maps [20], [168], though the type of pose map can vary. [67], [154], [183], [197] uses a multi-channel pose heatmap where each channel represents a key joint of the human body. Meanwhile, there are models that use an RGB skeleton image to show the spatial relation among the joints [9], [30], [83]. [6], [137], [161], [189] utilise DensePose [66] to extract texture from an image and apply to a UV parametrisation of a human model. The parser-free models use pose maps during training but not during inference [59], [86]. In FIGURE 6, we present the pose heatmaps.

The majority of virtual try-on models use GANs to create photorealistic images [77], [111], [197]. They use discriminators and GAN loss to encourage their modules to ensure accurate segmenting and photorealism of the image [48], [83], [136], [202]. Other models tend to use a generator only for synthesising the image [68], [193] and use the perceptual loss from a pre-trained VGG network [93] to ensure realism. More recently, researchers have started using diffusion models instead of the traditional GAN approach to synthesise virtual try-on images. Recent studies have shown that diffusion models can perform this task better than GAN-based methods [64], [135], [214].

In the early method of virtual try-on, they would employ a generic convolutional encoder-decoder architecture with skip connections [157] to synthesise a virtual try-on image. Over time, models have developed more sophisticated models of the architecture to improve the fidelity of images such as using a transformer [23], [154], creating normalisation layers or blocks [55], [112], [113], [177], [199] and using feature pyramid networks [67], [138], [190].

There have been challenges in maintaining the identity and details of clothes in virtual try-on models. Virtual try-on models have developed warping modules to help preserve vital clothing details and align with the person's body. There are two methods for warping the clothing image: thin-plate spline (TPS) [15] and appearance flow (AF) [212]. Models that use TPS determine their parameters by establishing correspondences between the target clothing image and the candidate's body key points. To prevent clothing distortion caused by TPS, some models induce constraints [111], [147], [194], [197]. On the other hand, some virtual try-on models have utilised AF to warp the garment image [33], [59], [67]. AF comprises a set of 2D coordinate vectors, where each vector indicates which pixels in the clothing image should be deformed to align with the candidate. The warping module is not commonly used in garment transfer models because they do not have a clean image of the clothing item [126], [151], [189].

Virtual try-on models use a segmentation module to predict how the semantic layout will look after a virtual try-on [137], [177], [197], [202]. This module helps the model determine which parts of the image should be generated or preserved to achieve high-quality results. However, some models are against using segmentation because it can sometimes generate inaccurate or incomplete segments, which can lead to disastrous outcomes. Instead, these models suggest


FIGURE 5: Transferring a garment using a virtual try-on. StylePoseGAN [161] demonstrates its capability to extract clothing from a given image and seamlessly apply it to a candidate.

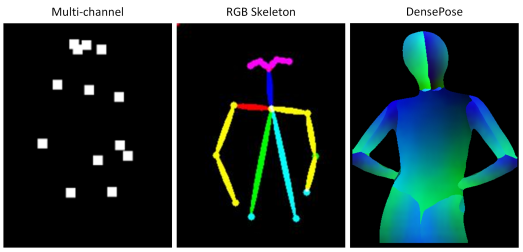



FIGURE 6: Variant of pose heatmaps used by virtual try-on models. Multi-channel heatmaps represent a single key joint in every channel, whereas the RGB skeleton shows how they are connected together. DensePose extracts texture from an image, which can be further manipulated.

using knowledge distillation [73] to train the student network to generate try-on images without relying on the predicted semantic layout [23], [59], [86], [138].

Alternative methods are needed in scenarios where clean images of clothes are absent. There are researchers who have proposed models capable of extracting the garment from one person and seamlessly applying it to another [6], [161], [189], [214]. As depicted in FIGURE 5, we showcase an illustrative example of how this garment transfer process can be accomplished.

Recently, there has been a notable surge of interest among researchers in enhancing the functionality of traditional image-based virtual try-on systems. A particular aspect that has captured attention is the incorporation of a posture modification feature.

TABLE 2: List of single-pose virtual try-on models.

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| CAGAN [89] | Jetchev and Bergmann | 2017 | ICCV[1] | GAN | N/A | Proposes to use images of a person in their current outfit and the desired outfit. Cycle loss is used in the training process by swapping the clothing item with a random item and then swapping it back to calculate the loss. |
| VITON [68] | Han et al. | 2018 | CVPR[2] | U-Net | N/A | Created an encoder-decoder network that generates a coarse sample from clothing-agnostic person images. Uses perceptual loss to improve image quality. |
| CP-VTON [183] | Wang et al. | 2018 | ECCV[3] | U-Net | TPS [15] | Uses Thin-Plate Spline (TPS) transformation to warp the garment before fusing with the person. |
| VITON-GAN [77] | Honda | 2019 | Arxiv | GAN | TPS | Improves CP-VTON [183] by training the model against a discriminator and uses improved loss function to improve the warping. |
| VTNFP [202] | Yu et al. | 2019 | ICCV[1] | GAN | TPS | Developed a network that predicts the segment of body parts of a person as if they are wearing the target clothing. Introduces non-local blocks to improve the accuracy of the segment. |
| LA-VITON [111] | Lee et al. | 2019 | ICCV[1] | GAN | TPS | Developed grid interval consistency loss and occlusion-handling techniques for improved geometric matching. |
| ClothFlow [67] | Han et al. | 2019 | ICCV[1] | Feature pyramid | OF [46], [174] | Utilises Optical Flows (OF) for warping the clothing image to fit it properly on the person. |
| CP-VTON+ [132] | Minar et al. | 2020 | CVPR[2] | U-Net | TPS | Improves CP-VTON [183] by fixing the erroneous clothing-agnostic human representations. |
| ACGPN [197] | Yang et al. | 2020 | CVPR[1] | GAN | TPS | Predicts segmentation to preserve non-targeted body parts. Added a constraint on the TPS to stabilise the way it warps the garment. |
| WUTON [86] | Issenhuth et al. | 2020 | ECCV[3] | GAN | TPS | During inference, the student network can synthesise images without the need for human parsing and pose estimation by using a pre-trained virtual try-on model to train via knowledge distillation. |
| CloTH-VTON [131] | Minar and Ahn | 2020 | ACCV[4] | GAN | TPS | Uses 3D reconstruction to convert 2D clothing images into 3D items, which are then accurately warped according to the person's depth and pose. |
| SieveNet [88] | Jandial et al. | 2020 | WACV[5] | U-Net | TPS | Introduces a perceptual geometric matching loss to capture detail from clothing and a triplet loss strategy for texture translation. |
| LG-VTON [193] | Xie et al. | 2020 | PRCV[6] | U-Net | TPS | Introduces clothes landmarks where the garments are separated into distinct regions and wrapped individually, which performs better than warping as a whole. |
| ZFlow [33] | Chopra et al. | 2021 | ICCV[1] | U-Net | AF [212] | Utilises a combination of gated aggregation of hierarchical appearance flow (AF) estimates and dense geometric priors to reduce undesirable output artefacts. |
| OVNet [117] | Li et al. | 2021 | CVPR[2] | GAN | TPS | Developed a warping module that can deform all garment types, which allows it to perform multi-garment synthesis. |
| DP-VTON [22] | Chang et al. | 2021 | ICASSP[8] | GAN | TPS | Proposing a multi-stage framework that first warps the clothing, then predicts the segments, generates the arms, and finally fuses them together. |
| M3D-VTON [208] | Zhao et al. | 2021 | ICCV[1] | U-Net | TPS | Performs standard 2D virtual try-on, followed by conversion into a 3D try-on mesh for multiple viewing angles. They use self-adaptive pre-alignment strategy to facilitate more accurate geometric matching. |
| DCTON [58] | Ge et al. | 2021 | CVPR[2] | GAN | TPS | Proposes to disentangle virtual try-ons into multiple modules within the cycle consistency framework [213]. Approach is similar to CAGAN [89]. |
| VTON-HF [48] | Du et al. | 2021 | ICTAI[7] | GAN | TPS | Proposes the Semantic Map-based Image Adjustment Network (SMIAN) for producing high-quality human body components by individually refining body parts without handling complexities arising from reliance on reference images. |

(Continued on next page)

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| PF-AFN [59] | Ge et al. | 2021 | CVPR[2] | Knowledge distillation | AF | Uses knowledge distillation scheme to generate try-on images, which allows for avoiding the utilisation of segmentation. |
| WAS-VTON [194] | Xie et al. | 2021 | MM[9] | GAN | AF | Leverages neural architecture search (NAS) [51] to identify optimal neural network architecture to warp clothing in all categories. |
| NL-VTON [177] | Tan et al. | 2021 | NSR[10] | U-Net | TPS | Proposes non-local feature attention, non-local grid regularisation loss and segmentation to alleviate blurriness in the image. |
| WBTP-VTON [165] | Shen et al. | 2021 | ICCE-TW[11] | U-Net | TPS | Extended the CP-VTON [183] model by allowing it to work on clothes for the entire body rather than upper clothes only. |
| CIT [154] | Ren et al. | 2021 | TOMM[12] | Transformer | TPS | Uses a transformer [181] to match the correlations when warping a cloth through TPS [15] and improves the rendering process. |
| VITON-HD [30] | Choi et al. | 2021 | CVPR[2] | GAN | TPS | Proposes alignment-aware segment (ALIAS) normalisation and ALIAS generator to handle misalignment between the warped clothes and the desired clothing regions. |
| FS-VTON [71] | He et al. | 2022 | CVPR[2] | StyleGAN | AF | Handles large misalignments between person and garment by using style-based [98] appearance flow to warp the garment more accurately. |
| ST-VTON [32] | Zheng and Lingfei | 2022 | IVC[13] | Transformer | TPS | Uses two-stage self-supervised training approach to allow the transformer model to merge the warped clothing with the person's body part and create a try-on image. |
| SVTON [83], [84] | Islam et al. | 2022 | ICMLA[14] | GAN | Affine | Developed a truncated U-Net that improves efficiency and uses an affine transform for warping the garment. Suggests that using an RGB skeleton image improves occlusion handling. |
| RT-VTON [196] | Yang et al. | 2022 | CVPR[2] | Local gated attention | MLS [162] | Proposes a mechanism that captures long-range dependency with local gated attention to predict segments and uses moving least squares (MLS) for clothes warping. |
| C-VTON [55] | Fele et al. | 2022 | WACV[5] | GAN | TPS | Makes use of contextual information by creating a Context-Aware (CA) generator that performs CA normalisation operations. |
| Dress Code [136] | Morelli et al. | 2022 | CVPR[2] | GAN | TPS | Created a new dataset with multiple clothing categories and introduced a new discriminator that improves the generation of high-quality images. |
| VTON-SCFA [49] | Du et al. | 2022 | TM[15] | U-Net | AF | Proposes a method that predicts a semantic map using global and local information. They also introduced the Appearance Flow Garment Alignment Network (AF-GAN), which aligns in-shop garments with the body using appearance flow estimation and feature-mapping structure to preserve maximum detail. |
| HR-VTON [113] | Lee et al. | 2022 | ECCV[3] | GAN | AF | Developed a model that simultaneously performs warping and segmentation prediction, allowing for misalignment-free images and handling of clothes occlusion by body parts. |
| SDAFN [9] | Bai et al. | 2022 | ECCV[3] | Feature pyramid | AF | Proposes a mechanism that fuses images in a single model using Deformable Attention Flow Network (DAFN) and Deformable Attention Warping (DAWarp). |
| PF-VTON [23] | Chang et al. | 2022 | MMM[16] | Knowledge distillation | TPS | Proposes a student network that uses a u-transformer [141] to synthesise images as it can deal with long-distance dependence. |
| UF-VTON [25] | Chang et al. | 2022 | ICMR[17] | U-Net | TPS | Combines convolutional neural network blocks with a swin-transformer [127] to extract and fuse global and local features to establish long-range dependencies. |
| POVNet [118] | Li et al. | 2023 | TPAMI[8] | GAN | TPS | Performs multiple warps on the garment image to ensure that the clothing texture always covers misaligned regions. |

(Continued on next page)

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| PG-VTON [54] | Fang et al. | 2023 | Arxiv | StyleGAN | TPS | Proposes a supervision strategy that enables the model to balance the warping of the garment and texture reservation. Uses StyleGAN2 [98] to synthesise matching body parts. |
| GP-VTON [190] | Xie et al. | 2023 | CVPR[2] | GAN | AF | Warps image parts separately and combine them based on the predicted garment segment. Uses gradient truncation for different samples to improve the training of the warping module. |
| LC-VTON [199] | Yao and Zheng | 2023 | IA[22] | GAN | AF | Uses clothing length value to control the generation of the try-on segmentation map, guiding the generation of length-controllable try-on results. |
| GC-VTON [152] | Rawal et al. | 2023 | Arxiv | GAN | AF | Predicts a visibility mask for body parts to avoid artefacts in occluded regions. Uses flow regularisation loss to prevent high appearance flow values that cause texture squeezing. |
| MT-VTON [112] | Lee et al. | 2023 | MAS[23] | GAN | AF | Warps clothing items using a pixel-level appearance flow. They use feature-level appearance flow at multiple levels, refining clothes from coarse to fine resolution, allowing them to preserve contextual features. |
| VTON-IT [5] | Adhikari et al. | 2023 | Arxiv | GAN | Pix2Pix | Proposes three components for image synthesis. Human detection, body part segmentation, and image translation network. |
| Versatile-VTON [91] | Jin and Kang | 2023 | ICCE-Asia[18] | U-Net | TPS | The model combines explicit clothing transformation networks and probabilistic models to enable versatile try-ons of various clothing items. They demonstrated that the TPS applied to long pants leads to inadequate warping. |
| DM-VTON [138] | Nguyen-Ngoc et al. | 2023 | ISMAR-Adjunct[19] | Knowledge distillation | AF | Develops a lightweight parser-free student network that achieves real-time image synthesis without compromising quality. |
| DCI-VTON [64] | Gou et al. | 2023 | MM[9] | DM | AF | Uses a warping network to predict the appearance flow, which warps the clothes. The warped clothes are added to the input of the DM, along with the global condition. They also introduce a new branch to better use the coarse synthesis results obtained in the previous step. |
| LaDI-VTON [135] | Morelli et al. | 2023 | Arxiv | LDM | TPS | Uses LDM [156] to synthesise virtual try-on images. They improved the image by adding skip connections to the autoencoder, preserving details outside the inpainting region and creating a forward-only textual inversion module to retain texture information. |
| Swapnet [151] | Raj et al. | 2018 | ECCV[3] | GAN | Semantic map | Proposes a two-stage training pipeline and uses split channel segmentation for garment transfer. Uses a weakly supervised training method to handle the absence of supervised data. |
| SwapGAN [126] | Liu et al. | 2019 | TM[15] | GAN | Semantic map | Combines three generators in a multistage process and employs GANs along with the mask-consistency loss to achieve better results. |
| M2E-TON [189] | Wu et al. | 2019 | MM[8] | GAN | Dense Pose | Aligns the model image with the target person's pose, enhances the clothes' characteristics and fits the desired clothes to the target person's image. |
| O-VITON [137] | Neuberger et al. | 2020 | CVPR[2] | GAN | N/A | Generate a geometrically-based segmentation map that deforms the shape of the selected garments to conform to the target person. |
| LGVTON [158] | Roy et al. | 2020 | Arxiv | GAN | TPS | Uses a self-supervised approach to transfer the garment. A mask generator is proposed to tackle noisy estimates of landmarks to avoid incorrect warping. |
| Style Pose GAN [161] | Sarkar et al. | 2021 | Arxiv | GAN | N/A | Modifies the StyleGAN [98] network to more control over the synthesised output by encoding source image into a global latent vector. |
| Pose with Style [6] | Albahar et al. | 2021 | TOG[20] | GAN | Affine | Based on StylePoseGAN [161], uses a neural network to inpaint appearance features using human body mirror-symmetry. |

(Continued on next page)

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| PASTA-GAN [192] | Xie et al. | 2021 | NeurIPS[21] | StyleGAN | N/A | Utilises a patch-routed disentanglement module and the spatially adaptive residual module to disentangle garment style and preserve spatial information. |
| TryOn GAN [115] | Lewis et al. | 2021 | TOG[20] | StyleGAN | N/A | Learns the internal interpolation coefficients per layer in the StyleGAN2 [98] architecture to synthesise an image. It allows the model to accurately preserve identity features like body shape and skin colour. |
| TryOn Diffusion [214] | Zhu et al. | 2023 | CVPR[2] | DM | N/A | Proposes Parallel-UNet, which involves training two de-noising U-Nets in parallel, with one sending information to the other through cross-attention [181]. |

[1] IEEE International Conference on Computer Vision
[2] IEEE Conference on Computer Vision and Pattern Recognition
[3] European Conference on Computer Vision
[4] Asian Conference on Computer Vision
[5] IEEE Winter Conference on Applications of Computer Vision
[6] Chinese Conference on Pattern Recognition and Computer Vision
[7] IEEE International Conference on Tools with Artificial Intelligence
[8] IEEE International Conference on Acoustics, Speech and Signal Processing
[9] ACM International Conference on Multimedia
[10] Nature Scientific Reports
[11] IEEE International Conference on Consumer Electronics-Taiwan
[12] ACM Transactions on Multimedia Computing, Communications, and Applications
[13] Image and Vision Computing
[14] IEEE International Conference on Machine Learning and Applications
[15] IEEE Transactions on Multimedia
[16] International Conference on Multimedia Modeling
[17] ACM International Conference on Multimedia Retrieval
[18] IEEE International Conference on Consumer Electronics-Asia
[19] IEEE International Symposium on Mixed and Augmented Reality Adjunct
[20] ACM Transactions on Graphics
[21] Advances in Neural Information Processing Systems
[22] IEEE Access
[23] MDI Applied Sciences

TABLE 3 shows a list of multi-pose virtual try-on models that aim to generate a realistic image of a person donning the desired garment in a predefined pose while maintaining the identity of the person and product. This task presents great challenges, such as preserving facial detail and clothing texture. FIGURE 7 demonstrates a multi-pose virtual try-on model that can change a candidate's garment and posture. In this section, we will provide an overview of these models and explain their operational mechanisms.

In summary, the evolution of image-based virtual try-on models has witnessed significant advancements, with each approach introducing novel techniques. While early models like CAGAN [89] pioneered the field, they faced limitations. The transition to models like VITON [68] and Swapnet [151] exemplifies a shift toward better practicality. The majority of the work is limited to upper garments. The incorporation of pose maps and segmentation modules provided vital information to networks to synthesise accurate virtual try-on synthesis. Many models employ warping modules, choosing between TPS and AF based to warp the garment. Using transformers [181], knowledge-distillation [73] and creating normalisation layers have played a crucial role in refining image fidelity. The adoption of diffusion models over GANs in recent studies highlights the ongoing exploration of alternative synthesis methods for photorealistic outcomes.

### C. MULTI-POSE VIRTUAL TRY-ON

While the image-based virtual try-on models normally attempt to overlay the targeted clothing item onto the user's body without changing the posture (in this sense, it is also referred to as single-pose virtual try-on), the multi-pose virtual try-on models offer more flexibility by generating images with different clothing and onto different postures simultaneously. The criteria for a multi-pose virtual try-on model are:

- Transferring the facial identity of the person to the desired pose.
- Transferring clothing items that are not intended to be replaced to the desired pose.
- Ensuring the target clothing item fits well to the intended body part of the person in the desired pose.
- Retaining the texture and details of the garment.

MG-VTON [41] represents a pioneering step towards enabling virtual try-on models to synthesise images in new postures. The approach is similar to that employed by conventional virtual try-on models, with modules for generating body segments and warping the target garment. However, the novelty of the MG-VTON lies in its ability to generate images of candidates in diverse poses, which was previously not possible with traditional virtual try-on systems.

Most virtual try-on models have three major modules: segmentation, warping, and try-on synthesis [41], [50], [184]. The segmentation module is responsible for generating a semantic layout that aligns with the desired target pose. There are two ways in which a multi-pose virtual try-on model can warp a garment: either through a commonly used TPS [41], [209] or appearance flow [50], [200]. With both methods, the garment is warped and transformed to align with the target pose. The try-on module then combines all images and synthesises the remaining parts of the try-on image.

Multi-pose virtual try-on models often use a multi-channel pose heatmap to guide the synthesis of an image that matches the desired posture. In contrast, there are exceptions, such as AB-GAN [209], which uses an RGB skeleton image. The models listed in TABLE 3 utilised GANs; however, it is expected that diffusion models will be used in the future since single-pose virtual try-on models are already utilising them [64], [135], [214].

To summarise, MG-VTON [41] is a significant advancement in generating multi-pose virtual try-on images, which offers more benefits than traditional single-pose models. Like the single-pose variant, multi-pose also consists of three main modules: segmentation, warping, and try-on synthesis. The segmentation module creates a semantic layout corresponding to the desired pose, while the warping module adjusts the garment using techniques like TPS or appearance flow. The try-on synthesis module combines these elements to generate the final try-on image. The majority of multi-pose virtual try-on models use a multi-channel pose heatmap for image synthesis. Notably, GANs are widely employed in existing models. However, we predict that diffusion models will be used in the future since single-pose virtual try-on models are already utilising them [64], [135], [214].

### D. VIDEO VIRTUAL TRY-ON

Different to the image-based and multi-pose virtual try-on models, the video virtual try-on models aim to produce a continuous video of a user wearing a new clothing from the input of a target clothing and reference video of the user, or in some models, a single image of the user instead of a reference video. The criteria for a video virtual try-on model are:

- The facial identity of the person needs to remain consistent throughout the video.
- Preserving clothing items that are not intended to be replaced.
- Ensure the clothing item is consistently fitting, accurately positioned, and smooth in the video.
- The texture and details of the garment should be maintained throughout the video.

A video virtual try-on model uses an image of the target garment to dress a person in a video with spatiotemporal consistency, as demonstrated in FIGURE 8. This approach offers a wider range of viewing angles for the clothing product and illustrates how it moves with the human body. There are challenges presented in this category, such as ensuring the clothing is warped and deformed accordingly throughout the video and eliminating inconsistency between frames. TABLE 4 provides an overview of the methods used by video virtual try-on. This table reveals how researchers explore new and innovative ways of improving video virtual try-on models.

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| MG-VTON [41] | Dong et al. | 2019 | ICCV[1] | GAN | TPS | Proposes multiple modules to produce segment, manipulate the garment and transfer the person's image to a desired posture in a try-on image. |
| AB-GAN [209] | Zheng et al. | 2019 | MM[2] | GAN | TPS | Similar to CAGAN [89], the training process involves changing the clothing and poses and then swapping it back to the original to ensure natural alignment with the target body and consistency. |
| FashionOn [78] | Hsieh et al. | 2019 | MM[2] | GAN | N/A | Uses segmentation to guide its refinement network to preserve or generate realistic clothing details for the person's face. A perceptual loss function is used to synthesise high-quality faces. |
| TB-VTON [184] | Wang et al. | 2020 | MM[2] | GAN | TPS | Developed a generator that fuses the cloth with the person image via tree dilated fusion block (Tree-Block). |
| PT-VTON [211] | Zhou et al. | 2021 | Arxiv | GAN | TPS | Uses pose-attentional transfer network (PATN) [215] and two-stage training procedures for accurate texture transfer and better results. |
| 3D-MPVTON [180] | Tuan et al. | 2021 | IA[3] | GAN | TPS | Extends the capability of CloTH-VTON [131] and proposes a method for mapping the clothing texture that reduces artefacts at clothing boundaries. |
| MV-TON [50] | Du et al. | 2022 | ICASSP[4] | U-Net | AF | Use a deformation module for warped garment generation and preliminary pose transfer. A pose transfer network captures valuable conditions and enhances spatial transformation learning. |
| SS-VTON [70] | He et al. | 2022 | Arxiv | GAN | N/A | Proposing a single network that utilises the StyleGAN [97] architecture and SPADE [145] layers to generate images. |
| SPG-VTON [79] | Hu et al. | 2022 | TM[5] | GAN | TPS | Uses cyclic consistency loss to align clothing warping and global/local adversarial losses and face identity loss to improve overall quality. |
| DO-VTON [24] | Chang et al. | 2022 | MMM[7] | GAN | TPS | This process predicts the semantic map of a person wearing specific clothes in a certain pose. It has two stages: spatial alignment and generating fine-grained details using a multi-scale dilated convolution U-Net to alleviate artefacts. |
| VTON-MP [200] | Yu et al. | 2023 | TCE[8] | GAN | AF | Enhances detail preservation by using appearance flow for garment and pose manipulation and using an enhancement network to refine the image further. |
| CF-VTON [47] | Du and Xiong | 2023 | ICASSP[4] | GAN | N/A | Uses bi-directional feature matching for precise garment and body posture alignment, along with a network that captures vital information and enhances spatial transformation. |

[1] IEEE International Conference on Computer Vision
[2] ACM International Conference on Multimedia
[3] IEEE Access
[4] IEEE International Conference on Acoustics, Speech and Signal Processing
[5] IEEE Transactions on Multimedia
[7] International Conference on Multimedia Modeling
[8] IEEE Transactions on Consumer Electronics

TABLE 3: List of multi-pose virtual try-on models.

GR-VTON [155] was the first to create virtual try-on videos. This model utilised image-based and 3D techniques to apply dresses to a person in a video. However, relying on 3D data is difficult to gather and work with [43], [149]. The first deep learning based approach for producing video virtual try-on is called image2Video [149]. Their method uses images of segmentation, clothing, and video to generate virtual try-on videos.

Some of the video virtual try-on models employ TPS to warp the garment [43], [90], [106] whilst others perform garment transfer [45], [210]. Optical flow (OF) [46], [174] is a common technique utilised by video virtual try-ons [26], [43], [45]. OF allows the model to calculate offsets between adjacent frames in videos and helps to warp and align the garment and the person throughout the video to ensure consistency.

The earlier models for video synthesis used variants of the standard U-Net architecture [157]. These models added extra components like a memory module and made use of optical flow to guide the U-Net [43], [149]. However, recent models have achieved better performance by using the transformer model [181] to create try-on videos [81], [90], [178].

The discriminator has an important role in some video generation models to improve the temporal consistency of generated try-on videos [43], [90], [149]. Moreover, it can also be used to enhance the overall quality of the video by improving the sharpness, clarity, level of detail and reducing visual artefacts or distortion that may appear in the generated video [43], [149], [210].

The evolution of video virtual try-on models has witnessed significant advancements, beginning with GR-VTON [155] as the pioneer, leveraging image-based and 3D techniques to apply dresses in videos. Despite its innovation, reliance on 3D data posed challenges. Subsequent models like image2Video [149] emerged, employing deep learning to generate virtual try-on videos by utilising images only. Techniques such as TPS are used for garment warping. Many models use OF to ensure alignment and consistency between garments and individuals throughout the video. Early models incorporated

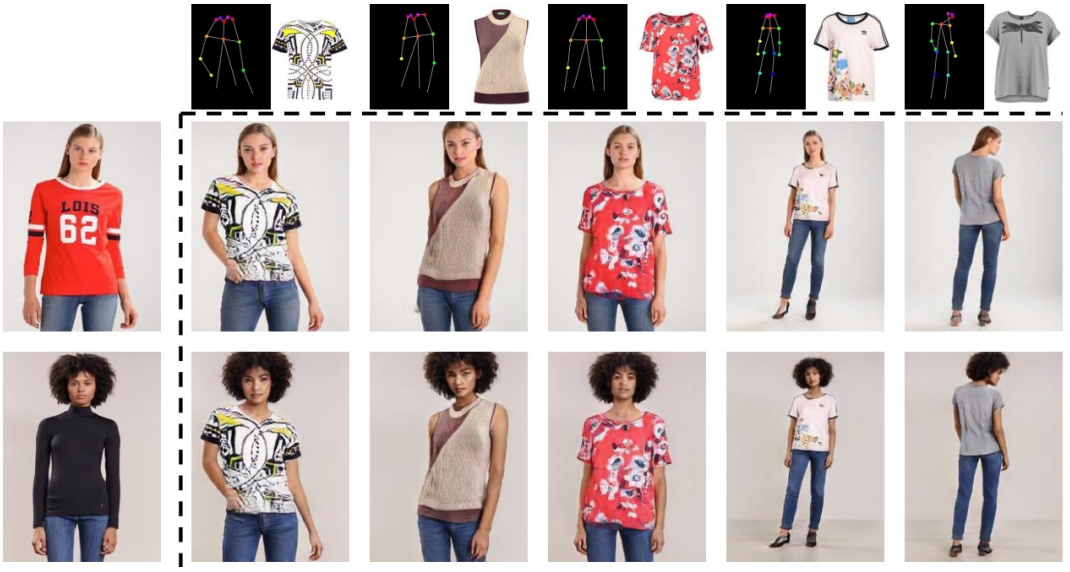


FIGURE 7: Examples of multi-pose virtual try-on images illustrated by He et al. [70]. The aim of this model is to synthesise a try-on image and change the posture simultaneously.

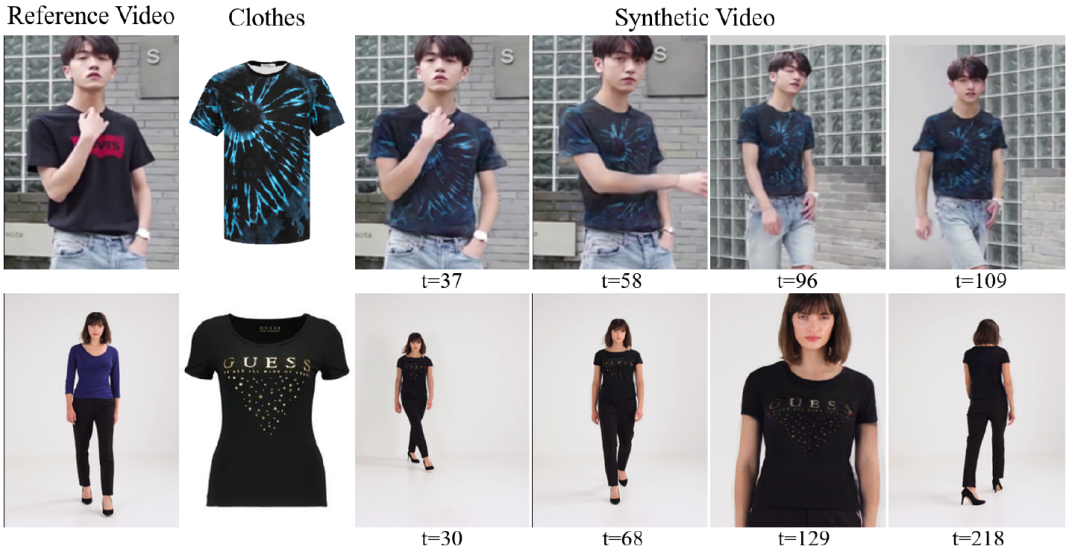



FIGURE 8: Examples of virtual try-on videos produced by ClothFormer [90]. The model aims to synthesise a try-on video using data from the garment.

U-Net architectures [157] with additional components. Recent advancements have embraced transformer models [181], achieving superior performance. The discriminator's role is pivotal, enhancing temporal consistency and overall video quality by refining sharpness, clarity, detail, and reducing visual artefacts and distortions in generated try-on videos.

## IV. DATASETS AND PERFORMANCE EVALUATION

In this section, we will discuss the dataset commonly used to train and evaluate virtual try-on models. Datasets serve as the foundational building blocks for the development and evaluation of such technologies. These datasets play a pivotal role in training algorithms, enabling virtual try-on algorithms to simulate realistic clothing interactions and user experiences. The diversity and quality of the dataset directly influence the accuracy and generalisation capabilities of virtual try-on systems.

It is crucial to demonstrate the quantitative and qualitative performance of virtual try-on models because they can significantly influence the shopping experience of customers [102], [207]. The accuracy of a virtual try-on model in generating content impacts the confidence of customers believing that the clothing item will fit and suit them in real life.

| Method | Authors | Year | Venues | Network type | Warping method | Contribution |
|---|---|---|---|---|---|---|
| GR-VTON [155] | Rogge et al. | 2014 | TOG[1] | N/A | Radial basis function | Uses a combination of 3D/image-based techniques to apply virtual garments to a video convincingly. |
| image2Video [149] | Pumarola et al. | 2019 | ICCV[2] | GAN | N/A | Proposes a GAN with a memory module that progressively refines the texture map to transfer the clothing to the target person. |
| FW-GAN [43] | Dong et al. | 2019 | ICCV[2] | GAN | TPS and OF | Redesigned discriminator specifically assessing video's temporal flow, with new loss function promoting model to synthesise results under various poses and clothes. |
| ShineOn [106] | Kuppa et al. | 2021 | WACV[3] | Transformer | TPS | Uses a U-Net consisting of attention layers to synthesise a try-on video frame and mask [36], [80] and accompanied by a warping module and optical flow module. |
| MV-TON [210] | Zhong et al. | 2021 | MM[4] | GAN | 3D AF [119] | Incorporates a memory refinement operation and a flow consistency loss to enhance and regularise OF in generated videos. |
| FashionMirror [26] | Chen et al. | 2021 | ICCV[2] | U-Net | OF | Proposes a co-attention feature-remapping in its model to produce a smooth video via OF and uses a parsing-free co-attention mask mechanism for efficiency. |
| wFlow [45] | Dong et al. | 2022 | CVPR[5] | GAN | OF | Integrates 2D and 3D body information to map garment textures between different persons and uses cyclic optimisation to refine the video. |
| ClothFormer [90] | Jiang et al. | 2022 | CVPR[5] | Transformer + GAN | TPS and AF | Combines TPS and AF to improve warping, along with using ridge regression to temporally smooth warped clothing sequence. Then, the Multi-scale Patch-based Dual-stream Transformer (MPDT) generator uses the input sequences to synthesize the final realistic video. |
| MMTrans [81] | Hu et al. | 2022 | VRCAI[6] | Transformer | TPS and OF | Uses a cross-modal transformer [179] to capture dependencies between clothing and person representations, thus enabling it to identify important areas and generate coherent videos. |
| DPV-VTON [170] | SK et al. | 2023 | ICMIP[7] | U-Net | TPS | Performs virtual try-on by capturing a clothing item from a video. Developed a module to select the best frames to build detailed clothing profiles. |
| AV-VTON [178] | Tsai et al. | 2023 | ICMR[8] | Knowledge distillation | AF | Proposes knowledge distillation to train a student model to generate video without segmentation. Uses a transformer [181] to enhance the quality of the generated video frame, refine the warped clothing and achieve temporal consistency. |

[1] ACM Transactions on Graphics
[2] IEEE International Conference on Computer Vision
[3] IEEE Winter Conference on Applications of Computer Vision
[4] ACM International Conference on Multimedia
[5] IEEE Conference on Computer Vision and Pattern Recognition
[6] ACM International Conference on Virtual-Reality Continuum and its Applications in Industry
[7] ACM International Conference on Multimedia and Image Processing
[8] ACM International Conference on Multimedia Retrieval

TABLE 4: List of video virtual try-on models.

### A. DATASETS

As outlined in TABLE 5, the datasets presented are highly valuable for training various groups of virtual try-on models. These datasets have gained significant recognition and are widely utilised within the research community.

All datasets listed for image-based virtual try-ons have a comprehensive collection of candidate images along with their corresponding clothing counterparts. The VITON [68], FashionTryOn [209] and MPV [42] datasets possess a resolution of 256 pixels for height and 192 pixels for width, while the VITON-HD dataset [30] has been crafted to yield superior quality try-on images by employing significantly higher resolutions, with a height of 1024 pixels and a width of 768 pixels.

The FashionTryOn and MPV datasets introduces an additional component wherein each candidate is presented with a different pose, further enriching the dataset for multi-pose virtual try-on research. This dataset maintains a resolution of 256 pixels for height and 192 pixels for width.

The authors of the MPV dataset have made the decision to withdraw it from public availability. The specific reasons behind the withdrawal have not been disclosed.

The MVC dataset introduced [124] offers distinct characteristics compared to the FashionTryOn and MPV datasets. MVC provides four different views (front, back, left, and right views) for each candidate, enabling a comprehensive multi-view analysis of clothing. The dataset has 161260 images and a resolution of 2240 pixels for height and 1920 pixels for width.

The video virtual try-on (VVT) dataset [43] consists of 791 videos. These videos were recorded at a rate of 30 frames per second, with a resolution of 256 pixels in height and 192 pixels in width. Each video has a different length, ranging from 250 to 300 frames. For training and testing purposes, the authors of the ShineOn framework [106] divided the dataset into a training set with 159170 frames and a testing set with 30931 frames. Each video in the dataset is paired with a clean image of the upper clothing item, like shirts or blouses.

The Dwnet dataset [204] comprises 500 videos in the training set and 100 videos in the testing set. These videos consist

| Name | #Training Set | # Testing set | Resolution (HxW) |
|---|---|---|---|
| **Single-pose** | | | |
| VITON [68] | 14221 | 2032 | 256x192 |
| VITON-HD [30] | 11647 | 2032 | 1024x768 |
| Dress Code [136] | 53000 | - | 1024×768 |
| **Multi-pose** | | | |
| FashionTryOn [209] | 21558 | 7156 | 256x192 |
| MPV [42] | 52236 | 10544 | 256x192 |
| MVC [124] | 161260 | - | 2240x1920 |
| DeepFashion [128] | 101966 | 8750 | 256x256 |
| **Video** | | | |
| VVT [43] | 791 | - | 256x192 |
| Dwnet [204] | 500 | 100 | 940x720 |

TABLE 5: Common datasets used on virtual try-on models.

of approximately 350 frames each and are recorded at 30 frames per second. In these videos, a female model is depicted wearing a dress and engaging in simple movements, such as shifting from side to side. This allows for a continuous and varied view of the dress from different angles. The average resolution of the videos is 720 pixels wide and 940 pixels tall, although this varies. It is worth noting that, unlike the VVT dataset, the Dwnet dataset does not include a clean image of the dress.

In TABLE 5, most of the datasets listed show an adult female standing upright against a white background. In these images and videos, the model is wearing various upper garments. This enables virtual try-on models to handle all sorts of cases regardless of the person's posture and the type of garment.

### B. PAIRED V UNPAIRED SETTINGS

Typically, a virtual try-on model uses two types of data: an image or video of a candidate and an image of the clothing product. The models are trained using the *paired settings*. This means that the candidate is paired with a clothing item that they are already wearing. By doing this, the synthesised content can be compared to the original candidate image or video without needing to modify it. This type of learning is called supervised learning.

Evaluating and comparing work is another reason for using paired settings. TABLE 6, 7, 8 show quantitative comparison in terms of quality of synthesised images/videos in paired settings.

*Unpaired settings* is employed for real-life scenarios. This entails pairing the candidate's image with clothing items that the person desires. This enables consumers to explore and experiment with various garments virtually. Once a model is trained in *paired settings* and has learned how to map the original clothing onto the person, it should be able to work on *unpaired settings*.

There are only a few studies that use *unpaired settings* to evaluate the accuracy of models that generate try-on images [83], [84], [192]. These studies have demonstrated that some models produce inaccurate results, for example, transforming a long-sleeved target garment into a short-sleeved one [83]. To demonstrate their performance, they use *unpaired settings* to show that they perform better. However, since there is no ground truth for *unpaired settings*, these models use pseudo ground-truth to show how a garment is meant to fit a person, even if the person's appearance and pose in the synthesised image are different.

### C. QUALITATIVE EVALUATION

Facilitating a direct comparison of try-on images among virtual try-on methods is a crucial aspect of evaluating their performance. Researchers assess and contrast the outcomes of different techniques. FIGURE 9, extracted from a popular virtual try-on research paper [55], serves as an illustrative example of this approach.

Many works present specific selections of candidate images with varying poses being paired with garments with varying sleeve lengths. By doing so, researchers aim to comprehensively demonstrate how their model performs across a variety of real-world situations.

The purpose of the qualitative comparison is to highlight the strengths and limitations of the presented virtual try-on model in relation to its competitors. By placing the visual outputs of different methods side by side, researchers can effectively showcase the unique features, accuracy, and realism achieved by their approach. This visual analysis provides valuable insights into the model's ability to accurately synthesise the try-on image and adapt to various garment types.

Another way researchers conduct qualitative evaluation is to carry out a user study. This method gathers volunteers and asks them to choose the best images generated from different virtual try-on models based on specific criteria such as photorealism and accuracy. ACGPN [197], WUTON [86], PF-AFN [59], StylePoseGAN [161], HR-VTON [113], DCI-VTON [64], MG-VTON [42], FashionOn [78], TB-VTON [184], FW-GAN [43], MV-TON [210] and video attention-based method [178] all use user studies to demonstrate their superiority over their predecessors.

### D. QUANTITATIVE EVALUATION

There are various methods available to evaluate the quality of synthesised try-on images on a test dataset. Some popular quantitative methods include the Structural Similarity Index (SSIM), Inception Score (IS), Fréchet Inception Distance

(FID), and learned perceptual image patch similarity (LPIPS). The majority of researchers have used paired settings for evaluation, which means that the person in the image is paired with a clothing image that they are already wearing.

The SSIM metric [186] gauges the resemblance between the synthesised image and the corresponding ground truth by evaluating the luminance, contrast, and structural similarities. The magnitude of the SSIM index directly reflects the level of concordance between the two images, with larger values indicating superior correspondence. The formula for SSIM is:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

The Structural Similarity Index (SSIM) formula compares the structural similarity between two input images, denoted as $x$ and $y$. It incorporates various variables to measure the average intensity $(\mu_x, \mu_y)$, spread or variability of pixel intensities $(\sigma_x^2, \sigma_y^2)$, and linear relationship $(\sigma_{xy})$ of the pixel values. Additionally, small constants $(C_1, C_2)$ are included to ensure stability and prevent division by zero. By evaluating the luminance similarity, contrast similarity, and structure similarity, the SSIM metric produces a similarity value ranging from 0 to 1, where 1 represents a perfect match or similarity between the images.

The Inception Score (IS) [160] is a metric for evaluating the quality of generative models. It measures the diversity and visual appeal of the generated images by feeding them through a pre-trained classifier and computing the score based on the output probabilities. Specifically, the IS is calculated as the exponential of the expected value of the KL divergence between the class distribution of the generated images and the class distribution of a large set of real images. A higher IS indicates that the generated images are more diverse and visually appealing.

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim P_G}\left[D_{\text{KL}}(P(y|x)\|P(y))\right]\right) \quad (2)$$

IS measures the quality and diversity of generated images. In the formula, $P_G$ represents the distribution of generated images. The notation $\mathbb{E}_{x\sim P_G}$ indicates that $x$ is sampled from this distribution. The conditional distribution $P(y|x)$ represents the distribution of labels $y$ (value from pre-trained classifier) given a generated image $x$. The marginal distribution $P(y)$ represents the overall distribution of labels. The Kullback-Leibler divergence, denoted as $D_{\text{KL}}$, quantifies the dissimilarity between these two distributions. The formula calculates the expected value, denoted by $\mathbb{E}$, of the Kullback-Leibler divergence overall generated images. Finally, the exponential function $\exp$ is applied to the expected Kullback-Leibler divergence to obtain the Inception Score. This score provides a measure of the quality and diversity of the generated images based on the discrepancy between the conditional and marginal label distributions.

The Fréchet Inception Distance (FID) metric [72], [163] leverages the widely used Inception network [176] to extract feature representations from both real and synthesised images. Subsequently, it quantifies the divergence between the two distributions of features by computing the Fréchet distance. Notably, a lower FID score implies that the feature distributions of the generated images are more closely aligned with those of the real images.

$$\text{FID}(P, Q) = \|\mu_P - \mu_Q\|_2^2 + \text{Tr}(C_P + C_Q - 2\sqrt{C_P C_Q}) \quad (3)$$

The FID formula compares the similarity between two probability distributions, P and Q, typically representing real and generated data distributions. The mean vectors, $\mu_P$ and $\mu_Q$, represent the average feature vectors of the data in each distribution, indicating their central tendencies. The covariance matrices, $C_P$ and $C_Q$, provide information about the spread or variability of the feature vectors. The squared Euclidean distance, $\|\cdot\|_2^2$, measures the dissimilarity between the mean feature vectors. The term $\text{Tr}(C_P + C_Q - 2\sqrt{C_P C_Q})$, which involves the trace operator (sum of diagonal elements), captures the difference between the covariances of the two distributions. By combining these components, the FID formula quantifies the discrepancy between the real and generated data distributions, providing a metric for evaluating the quality of generated data.

There are two types of Video Frechet Inception Distance (VFID) that can be utilised to assess the quality of synthesised videos, namely VFID I3D [21] and VFID ResNeXt101 [69]. Both I3D and ResNeXt101 are deep learning models that have achieved state-of-the-art performance in video object detection and action recognition tasks. They can extract temporal and spatial features from real and synthesised videos and use the equation from Eq. 3 to determine their performance scores.

The Learned Perceptual Image Patch Similarity (LPIPS) [205] metric employs a pre-trained deep neural network that has been fine-tuned to assess the perceptual similarity between images. The network is trained to capture human perception in the context of image quality. To determine the perceptual distance between two images, LPIPS calculates the dissimilarity between their respective feature maps of deep convolutional networks across multiple spatial scales and computes the average of these values to yield an overall score. A lower LPIPS score indicates that the generated images exhibit higher perceptual similarity to the real images.

Metrics such as IS and FID may not always accurately assess the output of a generative model [16], [31]. The sample size used to calculate FID should be sufficiently large; otherwise, smaller sample sizes can lead to an overestimation of the actual FID [31]. Also, IS and FID may not be ideal for diagnosing specific issues related to diversity [16]. SSIM was initially designed to evaluate the video compression ability of a model, and it is not a suitable metric for assessing the model's distortion performance, such as garment warping [131], [139]. To mitigate these issues, researchers conducting


FIGURE 9: A qualitative comparison was carried out by C-VTON [55]. This is a common approach researchers use to demonstrate their model's superiority over previous work on the same pair.

various types of evaluations, including comparisons of quantitative, qualitative, and user studies, can provide researchers with a clearer understanding of how their model performs.

In TABLE 6, 7 and 8, we have used FID, IS, LPIPS, SSIM, VFID I3D, and VFID ResNeXt101 as evaluation metrics to compare various models. These are the most commonly used metrics in the field. However, researchers could have used more advanced metrics, such as MS-SSIM [187] or KID [12], which could have provided a more accurate depiction of their model's performance. Due to the lack of consistency in reporting with these metrics by different researchers, comparing the results has become more challenging.

### *E. COMPARISON*

In this section, we aim to demonstrate the quantitative comparison made amongst virtual try-on models, assessed in terms of the quality of the synthesised try-on content. TABLE 6, 7 and 8 present published results of all categories of virtual try-on models, showcasing the scores achieved across various metrics.

TABLE 6 shows the results reported by researchers from different datasets. Most of these scores were evaluated on the VITON dataset [68]. The mean FID score for each year is as follows: 21.957 in 2018, 14.922 in 2020, 12.925 in 2021, and 11.563 in 2022. Furthermore, there is one model published in 2023 that scored 8.82. This suggests that every year, the FID score decreases, which indicates that the image quality gradually improves and these models are making advancements. The increase in the SSIM score over the years indicates promising improvements. The mean SSIM score for each year is as follows: 0.761 in 2018, 0.81 in 2020, 0.852 in 2021, and 0.893 in 2022. One model published in 2023 has achieved an impressive SSIM score of 0.918. The highest performing models for the FID metric in each year are ACGPN [197] in 2020, DP-VTON [22] in 2021, and VTON-SCFA [49] in 2022. As for SSIM, the highest performers are ACGPN [197] in 2020, Zflow [33] in 2021, and ST-VTON [32] in 2022.

Modern image-based virtual try-on models are evaluated on the VITON-HD dataset [30] due to their higher resolution and better quality. The mean FID score for each year is 13.645 in 2021, 10.91 in 2022 (based on a single model), and 8.606 in 2023. The decreasing FID score indicates an improvement in the performance of the models. Similarly, the mean SSIM score for each year is 0.8525 in 2021, 0.892 in 2022 (based on a single model), and 0.883 in 2023. The increasing SSIM score shows that the models are getting better at synthesising virtual try-on images. The highest performing models for the FID metric in each year are VITON-HD [30] in 2021 and GC-VTON [152] in 2023. As for SSIM, the highest performers are VITON-HD [30] in 2021 and GP-VTON [190].

TABLE 7 displays the quantitative performance of multi-pose virtual try-on models that have been reported by various researchers from two datasets. Most of these scores were evaluated on the MPV dataset [42]. However, when it comes to examining the metrics, it is challenging to identify whether multi-pose virtual try-on models are making any quantitative improvements. This is mainly due to the lack of adequate results to draw a conclusion. That being said, the models that performed best in the FID metric are SS-VTON [70] for 2022 and VTON-MP [200] for 2023. The models that performed

| Method | Year Published | FID ↓ | IS ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|---|
| **VITON dataset** | | | | | |
| CAGAN [89] | 2017 | $47.34^{\delta}$ | $2.56^{\delta}$ | - | $0.74^{\delta}$ |
| VITON [68] | 2018 | 19.463 | 2.514 | $0.0818^{\alpha}$ | $0.801^{\alpha}$ |
| CP-VTON [183] | 2018 | $24.45^{\beta}$ | $2.59^{\beta}$ | - | $0.720^{\beta}$ |
| VTNFP [202] | 2019 | - | $2.78^{\delta}$ | - | $0.80^{\delta}$ |
| ACGPN [197] | 2020 | 13.318 | 2.829 | $0.0737^{\alpha}$ | 0.845 |
| CloTH-VTON [131] | 2020 | - | 3.111 | - | 0.813 |
| CP-VTON+ [132] | 2020 | 16.800 | 3.105 | $0.117^{\gamma}$ | 0.816 |
| SieveNet [88] | 2020 | 14.65 | 2.82 | - | 0.766 |
| ZFlow [33] | 2021 | 15.17 | - | - | 0.885 |
| OVNet [117] | 2021 | - | 2.846 | - | 0.852 |
| DP-VTON [22] | 2021 | 8.726 | 3.044 | - | 0.871 |
| DCTON [58] | 2021 | 14.82 | 2.85 | - | 0.83 |
| VTON-HF [48] | 2021 | 12.63 | 2.86 | - | 0.87 |
| PF-AFN [59] | 2021 | 10.09 | - | - | - |
| WAS-VTON [194] | 2021 | 13.83 | - | $0.0959^{\gamma}$ | 0.8430 |
| NL-VTON [177] | 2021 | 14.163 | - | - | 0.843 |
| CIT [154] | 2021 | 13.97 | 3.060 | - | 0.827 |
| FS-VTON [71] | 2022 | 8.89 | - | - | 0.91 |
| ST-VTON [32] | 2022 | - | 2.9854 | 0.0725 | 0.9112 |
| RT-VTON [196] | 2022 | 11.66 | - | - | - |
| SVTON [83] | 2022 | 15.662 | 2.719 | 0.0647 | 0.854 |
| C-VTON [55] | 2022 | 19.535 | - | - | - |
| VTON-SCFA [49] | 2022 | 7.82 | 2.80 | - | 0.901 |
| SDAFN [9] | 2022 | 9.46 | - | - | - |
| PF-VTON [23] | 2022 | 9.628 | 3.003 | - | 0.889 |
| UF-VTON [25] | 2022 | 9.854 | 2.967 | - | - |
| POVNet [118] | 2023 | 8.82 | 2.92 | - | 0.918 |
| **VITON-HD dataset** | | | | | |
| VITON-HD [30] | 2021 | 11.74 | - | 0.053 | 0.895 |
| DCTON [58] | 2021 | 15.55 | 2.84 | - | 0.81 |
| HR-VTON [113] | 2022 | 10.91 | - | 0.065 | 0.892 |
| GP-VTON [190] | 2023 | 9.197 | - | 0.0799 | 0.8939 |
| LC-VTON [199] | 2023 | 8.21 | - | 0.073 | 0.858 |
| GC-VTON [152] | 2023 | 7.888 | - | 0.0831 | 0.887 |
| DCI-VTON [64] | 2023 | 9.13 | - | 0.053 | 0.892 |
| **MPV dataset** | | | | | |
| WUTON [86] | 2020 | 7.927 | 3.154 | 0.101 | 0.799 |
| PF-AFN [59] | 2021 | 6.429 | - | - | - |
| M3D-VTON [208] | 2021 | 20.04 | - | - | 0.8804 |
| C-VTON [55] | 2022 | 4.846 | - | 0.073 | - |
| SDAFN [9] | 2022 | 5.805 | - | - | - |

TABLE 6: Performance comparison of image-based virtual try-on models. Higher values indicate better results for SSIM and IS, while lower values are desirable for FID and LPIPS. Some scores have been reported in papers other than the original publication: $\alpha$: SVTON [84]; $\beta$: DCTON [58], $\gamma$: ST-VTON [32], $\delta$: VTON-HF [48]

| Method | Year Published | FID ↓ | IS ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|---|
| **MPV dataset** | | | | | |
| MG-VTON [42] | 2019 | $22.418^{\alpha}$ | 3.154 | $0.202^{\alpha}$ | 0.744 |
| TB-VTON [184] | 2020 | 16.006 | 3.193 | 0.187 | 0.723 |
| SS-VTON [70] | 2022 | 9.34 | - | 0.170 | 0.761 |
| MV-TON [50] | 2022 | 11.986 | 2.981 | $0.143^{\beta}$ | 0.793 |
| SPG-VTON [79] | 2022 | - | 3.243 | - | 0.752 |
| DO-VTON [24] | 2022 | 14.699 | - | - | 0.739 |
| VTON-MP [200] | 2023 | 11.986 | 2.982 | 0.273 | 0.7842 |
| CF-VTON [47] | 2023 | 15.371 | - | 0.227 | 0.794 |
| **DeepFashion dataset** | | | | | |
| MG-VTON [42] | 2019 | - | 3.030 | - | - |
| FashionOn [78] | 2019 | - | 3.191 | - | 0.894 |
| SPG-VTON [79] | 2022 | - | 3.124 | - | - |

TABLE 7: Performance comparison of multi-pose virtual try-on models. Higher values indicate better results for SSIM and IS, while lower values are desirable for FID and LPIPS. Some scores have been reported in papers other than the original publication: $\alpha$: TB-VTON [184], $\beta$: CF-VTON [47]

best in the SSIM metric are MV-TON [50] for 2022 and CF-VTON [47] for 2023.

TABLE 8 presents the quantitative evaluation of video virtual try-on models. It is worth mentioning that research in this category is not as active as it is with image-based and multi-pose virtual try models. Due to the limited amount of scores, it is not possible to draw a conclusion about whether newer models are making actual improvements. The ClothFormer [90] model performed the best out of all models based on all metrics.

## V. IMPACTS AND APPLICATIONS

In this section, we will scrutinise the influence of virtual try-on models on both customers and businesses. Our exploration will encompass the assessment of hedonic, utilitarian, perceived risk, and social values associated with virtual try-on technology, providing insights into its profound effects on customer satisfaction [52], [57]. Additionally, we will delve into the strategic advantages of this technology, explaining how it augments sales and optimises conversion rates.

### A. HEDONIC VALUE

The concept of hedonic value pertains to the enjoyment and satisfaction customers feel when they engage in a certain task [8]. Shopping, whether in physical stores [8], [14] or online [29], [76], provides a significant source of pleasure and motivation for customers, which leads to a positive shopping experience. Several studies have demonstrated that using virtual try-on technology results in a positive hedonic experience for customers [28], [102], [109].

In the fashion industry, customers have a strong desire for exceptional shopping experiences [13], and virtual try-ons allow it to happen. If customers' shopping expectations are not met, they may choose to engage in other leisure activities. Therefore, businesses need to ensure that their virtual try-on models are enjoyable to use [110]. This claim is also supported by the technology acceptance model (TAM) [38], which states that the adoption of technology is influenced by its perceived hedonic value.

Customers who perceive that their body is relatable to the virtual model are more likely to enjoy using virtual-try ons, leading to increased hedonic value [130]. This phenomenon is known as self-congruity, which refers to the tendency to compare oneself with other objects and stimuli [122].

### B. UTILITARIAN VALUE

The utilitarian value of a tool or technology refers to how useful customers perceive it to be and how effective they believe it will be in helping them achieve their goals [182]. Virtual try-on technology, for example, enables customers to quickly and conveniently try on numerous fashion items from any location, helping them to assess the size and fit of apparel [87]. Traditionally, customers need to physically see, feel, touch and try on clothing products before they can make a purchase [35]. This is difficult in the online realm; however, virtual try-on technology has the potential to reduce the strain.

In many studies, it has been shown that customers found virtual try-ons to have a high utilitarian value [10], [28], [102], [110]. They praised its ability to allow them to try on various fashion products and assess how well the items complimented their skin tone, hair colour, and other personal attributes. Furthermore, virtual try-ons' utilitarian values are significantly influenced by customer motivation, potentially affecting adoption intention [110].

The customers' perception of the utilitarian value towards virtual try-on models is a critical factor in determining their acceptance and usage [38], [130]. Virtual try-on models have to be user-friendly and straightforward to navigate; otherwise, individuals will be reluctant to utilise them and will diminish their utilitarian value [130].

Another important factor in increasing the utilitarian value of virtual try-ons is self-congruity. If customers perceive that their bodies are compatible or similar to the virtual models on the website, they are more likely to assume that the way a clothing product fits a virtual model will fit them in real life [130]. It is worth mentioning that customers with higher body esteem tend to perceive the virtual model as more self-congruent. This suggests that a virtual try-on model's utilitarian value can be influenced by the customer's perception of their own body [130].

It is important to note that not all customers find virtual try-on models to be very functional. In particular, those who have a low degree of self-congruity with virtual models may not see the value in using a virtual try-on model [130]. Additionally, some customers may be sceptical about how accurately the model can predict how well a garment will fit, which also reduces its utilitarian value [102], [207].

### C. PERCEIVED RISK

It is possible that some customers may be hesitant to use virtual try-on technology due to the chance that the virtual model's fit may not accurately reflect their own (meaning the clothing could appear to fit well on the model but not on the customer in real life) [102], [207]. The extent to which customers perceive risk plays a significant role in their willingness to use virtual try-ons [207].

When using virtual try-ons, customers may have concerns about the safety of their personal information, such as facial image, height, weight, bust size, waist size, and body shape. This information is transferred to other parties and may be at risk of being leaked, which can create security risks [107], [207]. Many mature users are hesitant to input sensitive information online, making them less likely to use tools that require it [107].

The level of perceived risk affects the connection between the utilitarian value and customers' overall opinion about virtual try-on technology [28]. Therefore, it is essential for businesses to help customers build confidence in using virtual try-on models to overcome any perceived risks.

Age is an important factor in determining the level of risk perception for customers. Minors (under the age of 18) are more interested in trying out new technologies like virtual

| Method | Year Published | VFID I3D ↓ | VFID ResNeXt101↓ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|---|
| **VVT dataset** | | | | | |
| FW-GAN [43] | 2019 | 7.052 | 23.94 | 0.283$^{\alpha}$ | 0.675$^{\alpha}$ |
| MV-TON [210] | 2021 | 8.367$^{\alpha}$ | 9.702$^{\alpha}$ | 0.233$^{\alpha}$ | 0.853$^{\alpha}$ |
| ClothFormer [90] | 2022 | 3.967 | 5.048 | 0.081 | 0.921 |
| MMTrans [81] | 2022 | 7.922 | 9.897 | 0.155 | 0.877 |
| **FashionVideo dataset [204]** | | | | | |
| FashionMirror [26] | 2021 | 3.097 | - | 0.057 | 0.923 |
| **Dance50k dataset [45]** | | | | | |
| wFlow [45] | 2022 | - | - | 0.090 | 0.920 |
| **DeepFashion dataset** | | | | | |
| wFlow [45] | 2022 | - | - | 0.187 | 0.844 |

TABLE 8: Performance comparison of video virtual try-on models. Higher values indicate better results for SSIM, while lower values are desirable for VFID I3D, VFID ResNeXt101, and LPIPS. Some scores have been reported in papers other than the original publication: $\alpha$: ClothFormer [90]

try-ons and perceive these tools to be less risky compared to adults [203], [207]. Minors may be more eager to use virtual try-ons due to their lower perceived risk or lack of concern about sharing personal information to interact with the technology [121].

### D. SOCIAL VALUE

In the context of online shopping, social value refers to how a product affects a customer's perception of themselves and their status to others. Not only do customers evaluate the hedonic or utilitarian value of the product or service but also how it influences their image to others [175]. Virtual try-on models can help customers evaluate the social value of clothing products by allowing them to see how they look on themselves and consider how others would perceive them.

Having a positive body image makes customers more inclined to share pictures of themselves, including virtual try-on images, on social media [53]. For this reason, body esteem significantly influences the perceived social value that virtual try-ons can provide [87]. When customers feel confident in their own bodies, they can use the try-on result to assess how a clothing item looks on them and how others perceive it.

Virtual try-on technology allows customers to generate and share previews of themselves wearing trendy clothing, increasing social value [87], [94]. Some customers prefer socially connected online businesses. This means making it practical to share content with friends and family [39], making virtual try-on a valuable asset.

### E. MULTI-GROUP DIFFERENCES

There are a few studies that have investigated differences in attitudes among various social identity categories towards virtual fitting rooms. There are no notable variations between genders when it comes to their preference towards virtual try-on systems [102], [207]. This means that all genders are equally likely to utilise virtual fitting room tools in the same manner. Customers under the age of 18 exhibit a more positive attitude towards virtual fitting tools compared to those above the age of 18 [207].

In a utilitarian sense, male interviewees indicated that virtual try-on models would be useful for buying suits and jeans, whereas female interviewees believed that models would be helpful when shopping for underwear, bathing suits, or dresses, as it allowed them to see how the item would look like on a body [102].

### F. IMPACT ON BUSINESS

The use of deep learning/generative AI models in the business field is a relatively new concept and has not yet been extensively explored in academic research. As a result, there is limited information on the impact of these models on business activities, outcomes and financial performance [104]. However, recent studies have suggested that AI tools can help businesses provide personalised, timely and relevant content to customers, resulting in greater customer satisfaction [18]. Additionally, the use of deep learning and generative AI models can enable businesses to perform tasks at a lower cost and more effectively than human teams and even tackle tasks that are impossible for humans to perform [104], [105]. Virtual try-on models, in particular, have many benefits for businesses, such as promoting personalised marketing and performing tasks that are not possible for humans.

Virtual try-on technology has been shown to have several positive outcomes for businesses. These outcomes include increased visitor numbers, longer time spent on the site, higher spending, and greater sales, as well as an increase in the conversion rate of customers [103]. By employing virtual try-on models, it is observed that the return rate was significantly decreased [82], and the customers were willing to spend more [60]. Research has indicated that the adoption of virtual try-on technology for online apparel shopping can significantly reduce the risk regarding apparel fit perceived by consumers when shopping online while also increasing the enjoyment of the shopping experience [100]. The perceived entertainment value of virtual try-on technology has been found to have a positive influence on attitudes towards the technology, as consumers tend to enjoy immersing themselves in the virtual simulation [206]. Additionally, a study focusing on the US department store industry during the pandemic highlighted the success of virtual try-on technology as an innovative strategy, with potential for success in the post-pandemic world [185].

Moreover, offering virtual fit information has been shown to increase conversion rates and order value [7], while reducing fulfilment costs arising from returns and home try-on behaviour [56]. It is also suggested that the direction of technology acceptance model-related research should be drawn by the functional or hedonic purpose of the technology/system [101]. The effectiveness of 3D virtual fitting technology has been investigated, showing its potential to create well-fitting clothes efficiently, particularly for tailored suits and jackets [171]. Furthermore, the application of virtual try-on technology has been explored in the context of reducing consumers' perceived risk about apparel fit, which can positively impact their purchasing experiences [167].

## VI. CHALLENGES AND FUTURE DIRECTIONS

In this section, we will delve into the obstacles that virtual try-on models encounter. These challenges include issues related to performance, accuracy, and user experience. Additionally, we will examine the latest advancements in virtual try-on technology and discuss potential future directions for research and development in this field.

### A. ACCURACY

Virtual try-on models can sometimes inaccurately deform the garment, resulting in the item appearing as short-sleeved instead of long-sleeved and not aligning with the person's body properly. This is because the segmentation module responsible for predicting the semantic layout is not always accurate [83], [86]. The garment is distorted based on the predicted semantic layout, so when the prediction is incorrect, the garment will also be distorted inaccurately.

A limitation of multi-pose virtual try-on models is that they may not accurately preserve the body size of the input image when generating a new try-on image in a different posture [70]. This means that if the person in the original image is wider, the model may generate a try-on image that makes them appear skinnier. He et al. [70] argue that datasets like MPV [42] do not include enough people with diverse body shapes, which limits the model's ability to capture the body shape of the input image.

FIGURE 10a illustrates how the models [183], [197] failed to maintain the shape of the clothing and align it correctly on the person, resulting in inaccurate images. Multi-pose virtual try-on does not maintain the same level of accuracy as single-pose models. All elements have to be transferred to the desired pose, which can lead to inaccuracies in synthesising images that do not complement the candidate's body shape. This is demonstrated in the third row of FIGURE 10a.

### B. QUALITY

Virtual try-on models sometimes produce low-quality content. For instance, they may fail to accurately preserve the logo or capture textures of the target garment. Occlusion handling is another issue that can degrade the quality of virtual try-on images. For example, if a person crosses their arms, some models may struggle to handle this situation and end up producing unrealistic try-on images. It is possible that the image generator may not be capable enough to handle certain pairs of images in some cases. To address such issues, newer researchers have attempted to use more complex loss functions or add more powerful components to the generator [54], [55], [113].

In FIGURE 10b, it is evident that some models have generated low-quality virtual try-on images. The first row [43] shows that the model failed to incorporate the bike that appears in the target garment in the synthesised image. The second row [183] displays poor occlusion handling, as the exposed arms have unnaturally blended behind the garment. The last row [135] displays a result generated using a diffusion model. While diffusion models are known to be superior to GANs [40], they still have limitations. The result shows that the model did not preserve the logo adequately, as the text is not readable.

### C. FACIAL PRESERVATION

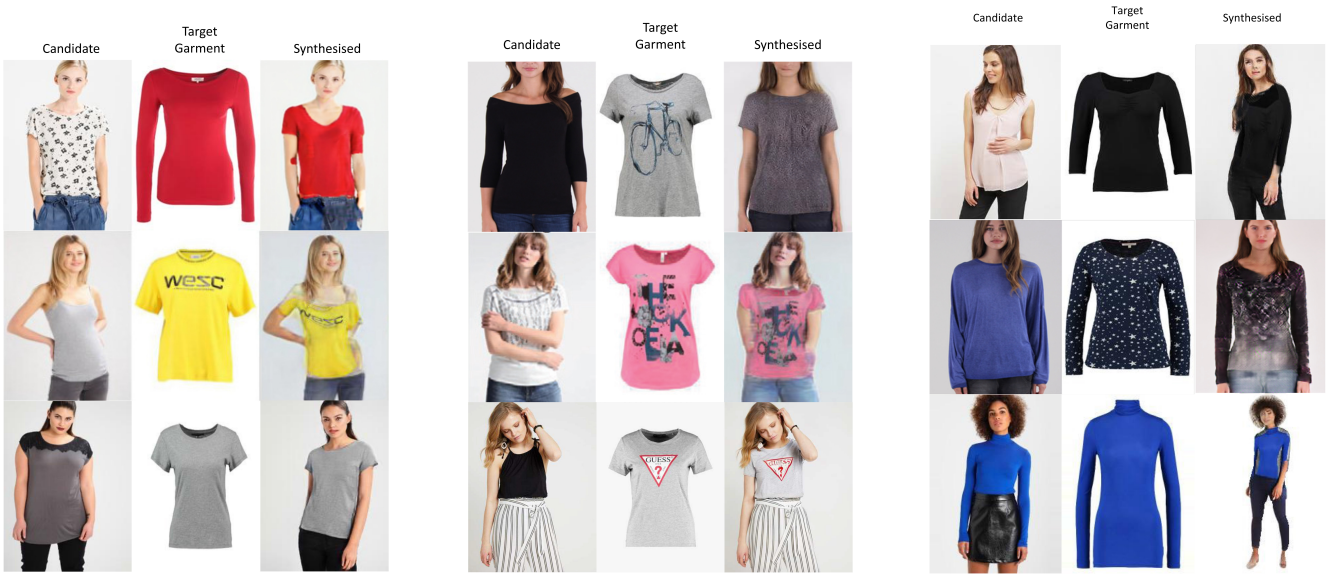
Image-based virtual try-on models have an easier task because they can copy the candidate's face and place it in the desired position without requiring much additional alteration. However, multi-pose and some video virtual try-on models face a more difficult challenge of maintaining the candidate's facial identity. This is because the face must be transferred to a new position while preserving its unique characteristics. Accomplishing this requires the generator to handle significant spatial deformation, which is an extremely challenging task [183]. These models ultimately generate facial images that do not resemble the person.

For example, FIGURE 10c shows the outcomes generated by multi-pose and video virtual try-on models [41], [43], [184]. These models do not preserve the faces properly and produce a generic face that only captures the skin colour, hair colour, and length.

### D. DATASETS

It is essential to include a wider and more diverse range of images in virtual try-on model datasets. The current datasets mainly consist of adult women who are mostly Caucasians and wearing Western-specific clothing, which underrepresents other ethnic groups. The datasets have less representation of people who are Asian and African, which can lead to biased models and poor performance for customers who fall under those races. Moreover, underprivileged countries with underrepresented cultures will face limited accuracy and reduced compatibility when trying on cultural clothing. Therefore, it is necessary to incorporate a more diverse range of images in the datasets to ensure fair and equal representation for all customers.

Non-diverse datasets can also cause models to exhibit bias against other underrepresented groups such as different genders, body types, ages, disabilities, and locations. This can negatively affect the popularity of virtual try-on models. Research shows that there is no significant difference in the attitude towards virtual try-ons between males and females




(a) (b) (c)

FIGURE 10: Weaknesses presented by virtual try-on models. (a) Some models find it difficult to accurately apply clothing on a person. The first [197] row shows the sleeve has been ruined, and the second [183] row depicts the clothing did not align with the body properly. The third [70] row shows the body shape has inaccurately shrunken in the synthesised image. (b) Capturing and merging content from conditional data is complex. The first [43] row demonstrates that the model was unable to capture logos, patterns, and textures accurately. Occlusion handling is also a significant challenge and can lead to synthesising low-quality images, as shown in the second [183] row. Even the diffusion model can struggle to capture text and logo from a conditional image, as seen in the third [135] row. (c) Preserving facial identity is extremely challenging for multi-pose and video virtual try-on models. The first [41] and third [184] rows are multi-pose models that did not preserve the facial identity, while the second row [43] is a video model that also failed to preserve the facial identity.

[102], [207] and that minors (under 18) are more likely to engage with and utilise virtual try-on systems compared to adults [207]. Therefore, it is crucial to include as many diverse groups as possible in the dataset to make virtual try-ons applicable to everyone.

Researchers assume that the test set has the same has the same distribution as the train set, which does not reflect the real-world scenario [113]. In reality, people can take pictures or videos from various angles or even capture images without the full body of humans, which can affect the performance of all categories of virtual try-ons. Therefore, it is essential to consider real-world scenarios when developing and evaluating these tools.

### E. ECONOMICAL IMPACT

Several companies have implemented virtual try-on models on their platforms. These models allow customers to experiment with products virtually. Some examples of such companies are Ray-Ban, L'Oreal Paris, and Hugo Boss [1]–[3]. However, these companies have not provided any information regarding the benefits for themselves or their customers or whether this tool has had an economic impact. Researchers have pointed out the lack of reported outcomes and financial performance when employing deep learning/generative AI models, as they are a new concept in the business domain [104]. Therefore, research is needed to determine if virtual try-on models have a financial impact on businesses. Currently, there are no clear studies exploring this topic.

### F. EMOTIONAL CHALLENGES

Virtual try-on models are designed to address the uncertainty that customers may have when purchasing clothing online without the ability to physically try it on. However, some customers may be hesitant to trust the accuracy of these models, fearing that they may not accurately reflect the real-life fit or look of the product [102], [207]. This presents a challenge for researchers, who must work to find a solution that will increase customer confidence in the accuracy of virtual try-on models and demonstrate that they effectively reduce the uncertainty associated with online clothing shopping. The development progress of virtual try-on models is significant, but research is needed to indicate whether these models perform sufficiently to the degree that they earn customers' trust in the accuracy and are willing to use them when shopping for clothes online.

### G. CODE AVAILABILITY

We have discussed several models in TABLE 2, 3, and 4, but only a few of these models have shared their code publicly. The implementation of models developed by others can be

a challenging task due to unclear implementation details, and results may vary from those reported by the original authors. This poses a problem as it reduces the scope for reproducibility and makes it difficult for other researchers to compare their work with these models.

### H. FUTURE DIRECTIONS

The image-based virtual try-on category is the most competitive and popular among other categories. Every year, new studies are published showing how image-based virtual try-ons have surpassed previous work. This trend is shown in TABLE 2 where researchers keep making progress and finding weaknesses in prior work, promptly coming up with solutions. It has been observed that researchers are moving away from using GANs and focusing more on using powerful generative models such as the diffusion model [19], [64], [95], [120], [214].

Researchers have proposed various approaches to synthesise multi-pose try-on images, as shown in TABLE 3. These approaches generally consist of three stages: first, generating a semantic map of the target pose; second, warping the garment; and finally, re-rendering the person in the target pose while fusing with the warped garment. However, the single-stage model [70] accomplishes all these steps in a single network. The three-stage trend may continue in future developments. Similarly to single-pose virtual try-on, researchers may use a diffusion model to synthesise multi-pose virtual try-on images.

The development of video virtual try-on is progressing towards using attention mechanism [181] and transformer model [181], as demonstrated in TABLE 4. These methods have exhibited promising results and enabled models to produce impressive outcomes. It is possible that future researchers will use the diffusion model to synthesise virtual try-on videos because there is active research on diffusion models generating videos [61], [74], [169], [188].

Virtual try-ons have been introduced as an alternative to physical fitting rooms, but not many customers have embraced this technology yet [87], [116]. This could be because many customers or businesses are not aware of the existence of virtual try-on models. However, with the advancements in technology, improvements in the accuracy of virtual try-ons, and gradual recognition by the masses, it is possible that the adoption of this technology may increase over time.

## VII. CONCLUSIONS

We have reviewed deep learning based virtual try-on models into three categories based on their functionality: image-based, multi-pose, and video. For each of these categories, we have provided comprehensive examples of models and summarised their technical details and contributions. We have also identified similarities in terms of methods used by researchers.

Furthermore, we have examined the datasets used by these models, including the number of images/videos and their resolutions. We have also observed that researchers tend to conduct qualitative comparisons by comparing their synthesised images with previous work. Additionally, they perform quantitative evaluations across various metrics and benchmark datasets.

We discussed weaknesses of deep learning based virtual try-on models, such as being unable to preserve clothing characteristics and textures and sometimes having difficulty applying the clothing to the person. Furthermore, we discuss dataset bias which mainly consists of images featuring women posing against a white background with limited clothing diversity. This can negatively impact the model's ability to handle less common garment types.

Our research has revealed that a number of enterprises have already integrated virtual fitting rooms into their platforms. These cutting-edge tools have been proven to offer a variety of benefits, which we anticipate will encourage more companies to adopt this technology in order to enhance the decision-making process for their customers and provide a highly positive experience for all involved.

Finally, we have examined how virtual try-ons affect the attributes and factors that lead to customer satisfaction. We have shown that researchers highlight the benefits that customers can enjoy, which also reduces returns and optimises conversion rates for businesses.

## REFERENCES

[1] https://www.ray-ban.com. [Accessed 11-12-2023].
[2] https://www.hugoboss.com/uk/all-brands/men/new-in/virtual-try-on/. [Accessed 11-12-2023].
[3] Virtual Try On for Hair Colour & Makeup | L'Oréal Paris UK — loreal-paris.co.uk. https://www.loreal-paris.co.uk/virtual-try-on. [Accessed 11-12-2023].
[4] Fashion e-commerce worldwide - statistics & facts. https://www.statista.com/topics/9288/fashion-e-commerce-worldwide/#topicOverview, Jun 2023.
[5] S. Adhikari, B. Bhusal, P. Ghimire, and A. Shrestha. Vton-it: Virtual try-on using image translation. *arXiv preprint arXiv:2310.04558*, 2023.
[6] B. Albahar, J. Lu, J. Yang, Z. Shu, E. Shechtman, and J.-B. Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021.
[7] A. Ayanso and R. Yoogalingam. Profiling retail web site functionalities and conversion rates: A cluster analysis. *International Journal of Electronic Commerce*, 14(1):79–114, 2009.
[8] B. J. Babin, W. R. Darden, and M. Griffin. Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of consumer research*, 20(4):644–656, 1994.
[9] S. Bai, H. Zhou, Z. Li, C. Zhou, and H. Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022.
[10] F. Baytar, T.-l. D. Chung, and E. Shin. Can augmented reality help e-shoppers make informed purchases on apparel fit, size, and product performance? In *International Textile and Apparel Association Annual Conference Proceedings*, volume 73. Iowa State University Digital Press, 2016.
[11] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023.
[12] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.
[13] M. Blázquez. Fashion shopping in multichannel retail: The role of technology in enhancing the customer experience. *International Journal of Electronic Commerce*, 18(4):97–116, 2014.
[14] P. H. Bloch, D. L. Sherrell, and N. M. Ridgway. Consumer search: An extended framework. *Journal of consumer research*, 13(1):119–126, 1986.

[15] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989.

[16] A. Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[17] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. sep 2018.

[18] F. Buttle and S. Maklan. *Customer relationship management: concepts and technologies*. Routledge, 2019.

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

[20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[21] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[22] Y. Chang, T. Peng, R. He, X. Hu, J. Liu, Z. Zhang, and M. Jiang. Dp-vton: toward detail-preserving image-based virtual try-on network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2295–2299. IEEE, 2021.

[23] Y. Chang, T. Peng, R. He, X. Hu, J. Liu, Z. Zhang, and M. Jiang. Pf-vton: Toward high-quality parser-free virtual try-on network. In *International Conference on Multimedia Modeling*, pages 28–40. Springer, 2022.

[24] Y. Chang, T. Peng, R. He, X. Hu, J. Liu, Z. Zhang, and M. Jiang. Toward detail-oriented image-based virtual try-on with arbitrary poses. In *International Conference on Multimedia Modeling*, pages 82–94. Springer, 2022.

[25] Y. Chang, T. Peng, R. He, X. Hu, J. Liu, Z. Zhang, and M. Jiang. Uf-vton: Toward user-friendly virtual try-on network. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 313–321, 2022.

[26] C.-Y. Chen, L. Lo, P.-J. Huang, H.-H. Shuai, and W.-H. Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13809–13818, 2021.

[27] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021.

[28] V. Chidambaram, N. P. Rana, and S. Parayitam. Antecedents of consumers' online apparel purchase intention through virtual try on technology: A moderated moderated-mediation model. *Journal of Consumer Behaviour*, 2023.

[29] T. L. Childers, C. L. Carr, J. Peck, and S. Carson. Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of retailing*, 77(4):511–535, 2001.

[30] S. Choi, S. Park, M. Lee, and J. Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021.

[31] M. J. Chong and D. Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[32] Z. Chong and L. Mo. St-vton: Self-supervised vision transformer for image-based virtual try-on. *Image and Vision Computing*, 127:104568, 2022.

[33] A. Chopra, R. Jain, M. Hemani, and B. Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442, 2021.

[34] C.-T. Chou, C.-H. Lee, K. Zhang, H.-C. Lee, and W. H. Hsu. Pivtons: Pose invariant virtual try-on shoe with conditional image completion. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 654–668. Springer, 2019.

[35] A. V. Citrin, D. E. Stem Jr, E. R. Spangenberg, and M. J. Clark. Consumer need for tactile input: An internet retailing challenge. *Journal of Business research*, 56(11):915–922, 2003.

[36] J.-B. Cordonnier, A. Loukas, and M. Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[37] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[38] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. Extrinsic and intrinsic motivation to use computers in the workplace 1. *Journal of applied social psychology*, 22(14):1111–1132, 1992.

[39] C. Dennis, A. Morgan, L. T. Wright, and C. Jayawardhena. The influences of social e-shopping in enhancing young women's online shopping behaviour. *Journal of customer behaviour*, 9(2):151–174, 2010.

[40] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[41] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019.

[42] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin. Towards multi-pose guided virtual try-on network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9025–9034, 2019.

[43] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019.

[44] H. Dong, X. Liang, Y. Zhang, X. Zhang, Z. Xie, B. Wu, Z. Zhang, X. Shen, and J. Yin. Fashion Editing with Adversarial Parsing Learning. jun 2019.

[45] X. Dong, F. Zhao, Z. Xie, X. Zhang, D. K. Du, M. Zheng, X. Long, X. Liang, and J. Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022.

[46] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[47] C. Du and S. Xiong. Cf-vton: Multi-pose virtual try-on with cross-domain fusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[48] C. Du, F. Yu, Y. Chen, M. Jiang, X. Wei, T. Peng, and X. Hu. Vton-hf: High fidelity virtual try-on network via semantic adaptation. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 224–231. IEEE, 2021.

[49] C. Du, F. Yu, M. Jiang, A. Hua, X. Wei, T. Peng, and X. Hu. Vton-scfa: A virtual try-on network based on the semantic constraints and flow alignment. *IEEE Transactions on Multimedia*, 25:777–791, 2022.

[50] C. Du, F. Yu, M. Jiang, X. Wei, T. Peng, and X. Hu. Multi-pose virtual try-on via self-adaptive feature filtering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2544–2548. IEEE, 2022.

[51] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[52] T. Y. Evelina, A. Kusumawati, U. Nimran, et al. The influence of utilitarian value, hedonic value, social value, and perceived risk on customer satisfaction: survey of e-commerce customers in indonesia. *Business: Theory and Practice*, 21(2):613–622, 2020.

[53] X. Fan, X. Jiang, N. Deng, X. Dong, and Y. Lin. Does role conflict influence discontinuous usage intentions? privacy concerns, social media fatigue and self-esteem. *Information Technology & People*, 34(3):1152–1174, 2021.

[54] N. Fang, L. Qiu, S. Zhang, Z. Wang, and K. Hu. Pg-vton: A novel image-based virtual try-on method via progressive inference paradigm. *arXiv preprint arXiv:2304.08956*, 2023.

[55] B. Fele, A. Lampe, P. Peer, and V. Struc. C-vton: Context-driven image-based virtual try-on network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022.

[56] S. Gallino and A. Moreno. The value of fit information in online retail: Evidence from a randomized field experiment. *Manufacturing & Service Operations Management*, 2018.

[57] C. Gan and W. Wang. The influence of perceived value on purchase intention in social commerce context. *Internet research*, 27(4):772–785, 2017.

[58] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu, and P. Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16928–16937, 2021.

[59] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.

[60] H. Ghodhbani, M. Neji, I. Razzak, and A. M. Alimi. You can try without visiting: a comprehensive survey on virtually try-on outfits. *Multimedia Tools and Applications*, 81(14):19967–19998, 2022.

[61] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[62] P. Goel, K. Mahadevan, and K. K. Punjani. Augmented and virtual reality in apparel industry: A bibliometric review and future research agenda. *foresight*, 25(2):167–184, 2023.

[63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[64] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607, 2023.

[65] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012.

[66] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[67] X. Han, X. Hu, W. Huang, and M. R. Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

[68] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[69] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[70] S. He, Y.-Z. Song, and T. Xiang. Single stage multi-pose virtual try-on. *arXiv preprint arXiv:2211.10715*, 2022.

[71] S. He, Y.-Z. Song, and T. Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[72] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[73] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[74] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[75] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[76] D. L. Hoffman and T. P. Novak. Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of marketing*, 60(3):50–68, 1996.

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

[78] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM international conference on multimedia*, pages 275–283, 2019.

[79] B. Hu, P. Liu, Z. Zheng, and M. Ren. Spg-vton: Semantic prediction guidance for multi-pose virtual try-on. *IEEE Transactions on Multimedia*, 24:1233–1246, 2022.

[80] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[81] X. Hu, Z. Zhang, R. Luo, J. Huang, J. Liang, J. Huang, T. Peng, and H. Cai. Mmtrans: Multimodal transformer for realistic video virtual try-on. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–8, 2022.

[82] H. Hwangbo, E. H. Kim, S.-H. Lee, and Y. J. Jang. Effects of 3d virtual "try-on" on online sales and customers' purchasing experiences. *IEEE Access*, 8:189479–189489, 2020.

[83] T. Islam, A. Miron, X. Liu, and Y. Li. Svton: Simplified virtual try-on. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 369–374. IEEE, 2022.

[84] T. Islam, A. Miron, X. Liu, and Y. Li. Image-based virtual try-on: Fidelity and simplification. *Available at SSRN 4342099*, 2023.

[85] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[86] T. Issenhuth, J. Mary, and C. Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020.

[87] A. Ivanov, Y. Mou, and L. Tawira. Avatar personalisation vs. privacy in a virtual try-on app for apparel shopping. *International Journal of Fashion Design, Technology and Education*, 16(1):100–109, 2023.

[88] S. Jandial, A. Chopra, K. Ayush, M. Hemani, B. Krishnamurthy, and A. Halwai. Sievenet: A unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2182–2190, 2020.

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

[90] J. Jiang, T. Wang, H. Yan, and J. Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10799–10808, 2022.

[91] H.-W. Jin and D.-O. Kang. Versatile-vton: A versatile virtual try-on network. In *2023 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4. IEEE, 2023.

[92] Y. Jo and J. Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1745–1753, 2019.

[93] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[94] J.-Y. M. Kang and K. K. Johnson. How does social commerce work for apparel shopping? apparel social e-shopping with social network storefronts. *Journal of Customer Behaviour*, 12(1):53–72, 2013.

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

[96] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, oct 2018.

[97] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[98] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[99] K. Kikuchi, K. Yamaguchi, E. Simo-Serra, and T. Kobayashi. Regularized adversarial training for single-shot virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[100] D.-E. Kim. Psychophysical testing of garment size variation using three-dimensional virtual try-on technology. *Textile Research Journal*, 2015.

[101] J. Kim and S. Forsythe. Hedonic usage of product virtualization technologies in online apparel shopping. *International Journal of Retail & Distribution Management*, 2007.

[102] J. Kim and S. Forsythe. Adoption of virtual try-on technology for online apparel shopping. *Journal of interactive marketing*, 22(2):45–59, 2008.

[103] J. Kim and S. Forsythe. Adoption of virtual try-on technology for online apparel shopping. *Journal of Interactive Marketing*, 2008.

[104] N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli. Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda, 2023.

[105] W. H. Kunz and J. Wirtz. Corporate digital responsibility (cdr) in the age of ai: implications for interactive marketing. *Journal of Research in Interactive Marketing*, 18(1):31–37, 2024.

[106] G. Kuppa, A. Jong, X. Liu, Z. Liu, and T.-S. Moh. Shineon: Illuminating design choices for practical video-based virtual clothing try-on. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 191–200, 2021.

[107] M. Law and M. Ng. Age and gender differences: Understanding mature online users with the online purchase intention model. *Journal of Global Scholars of Marketing Science*, 26(3):248–269, 2016.

[108] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[109] H. Lee and Y. Xu. Classification of virtual fitting room technologies in the fashion industry: from the perspective of consumer experience. *International Journal of Fashion Design, Technology and Education*, 13(1):1–10, 2020.

[110] H. Lee and Y. Xu. Influence of motivational orientations on consumers' adoption of virtual fitting rooms (vfrs): Moderating effects of fashion leadership and technology visibility. *International Journal of Fashion Design, Technology and Education*, 15(3):297–307, 2022.

[111] H. J. Lee, R. Lee, M. Kang, M. Cho, and G. Park. LA-VITON: A Network for Looking-Attractive Virtual Try-On. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3129–3132. IEEE, oct 2019.

[112] J. Lee, M. Lee, and Y. Kim. Mt-vton: Multilevel transformation-based virtual try-on for enhancing realism of clothing. *Applied Sciences*, 13(21):11724, 2023.

[113] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 204–219. Springer, 2022.

[114] Y. Lei, W. Du, and Q. Hu. Face sketch-to-photo transformation with multi-scale self-attention gan. *Neurocomputing*, 396, 2020.

[115] K. M. Lewis, S. Varadharajan, and I. Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021.

[116] A. Li and Y. Xu. A study of chinese consumers' adoption behaviour toward virtual fitting rooms. *International Journal of Fashion Design, Technology and Education*, 13(2):140–149, 2020.

[117] K. Li, M. J. Chong, J. Zhang, and J. Liu. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15546–15555, 2021.

[118] K. Li, J. Zhang, and D. Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[119] Y. Li, C. Huang, and C. C. Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[120] Z. Li, P. Wei, X. Yin, Z. Ma, and A. C. Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22788–22797, 2023.

[121] J.-W. Lian and D. C. Yen. Online shopping drivers and barriers for older adults: Age and gender differences. *Computers in human behavior*, 37:133–143, 2014.

[122] F. Liu, J. Li, D. Mizerski, and H. Soh. Self-congruity, brand attitude, and brand loyalty: a study on luxury brands. *European Journal of Marketing*, 46(7/8):922–937, 2012.

[123] J. Liu, X. Song, Z. Chen, and J. Ma. MGCM: Multi-modal generative compatibility modeling for clothing matching. *Neurocomputing*, 414, 2020.

[124] K.-H. Liu, T.-Y. Chen, and C.-S. Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 313–316, 2016.

[125] L. Liu, H. Zhang, Y. Ji, and Q. M. Jonathan Wu. Toward AI fashion design: An Attribute-GAN model for clothing match. *Neurocomputing*, 341, 2019.

[126] Y. Liu, W. Chen, L. Liu, and M. S. Lew. Swapgan: A multistage generative approach for person-to-person fashion style transfer. *IEEE Transactions on Multimedia*, 21(9):2209–2222, 2019.

[127] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[128] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[129] D. Marelli, S. Bianco, and G. Ciocca. Designing an ai-based virtual try-on web application. *Sensors*, 22(10):3832, 2022.

[130] A. Merle, S. Senecal, and A. St-Onge. Whether and how virtual try-on influences consumer responses to an apparel web site. *International Journal of Electronic Commerce*, 16(3):41–64, 2012.

[131] M. R. Minar and H. Ahn. Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. In *Asian Conference on Computer Vision (ACCV)*, 2020.

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

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

[134] S. O. Mohammadi and A. Kalhor. Smart fashion: a review of ai applications in virtual try-on & fashion synthesis. *Journal of Artificial Intelligence*, 3(4):284, 2021.

[135] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023.

[136] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022.

[137] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5184–5193, 2020.

[138] K.-N. Nguyen-Ngoc, T.-T. Phan-Nguyen, K.-D. Le, T. V. Nguyen, M.-T. Tran, and T.-N. Le. Dm-vton: Distilled mobile real-time virtual try-on. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 695–700. IEEE, 2023.

[139] J. Nilsson and T. Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.

[140] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[141] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[142] I. Pachoulakis and K. Kapetanakiss. Augmented reality platforms for virtual fitting rooms. *The International Journal of Multimedia & Its Appliications*, 4(4):35, 2012.

[143] S. Pande, S. Chouhan, R. Sonavane, R. Walambe, G. Ghinea, and K. Kotecha. Development and deployment of a generative model-based framework for text to photorealistic image generation. *Neurocomputing*, 463, 2021.

[144] N. Pandey and A. Savakis. Poly-gan: Multi-conditioned gan for fashion synthesis. *Neurocomputing*, 414, 2020.

[145] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[146] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[147] D. L. Pham, N. T. Ngyuen, and S.-T. Chung. Keypoints-based 2d virtual try-on network system. *Journal of Korea Multimedia Society*, 23(2):186–203, 2020.

[148] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally.

[149] A. Pumarola, V. Goswami, F. Vicente, F. De la Torre, and F. Moreno-Noguer. Unsupervised image-to-video clothing transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[150] Z. Qi, J. Sun, J. Qian, J. Xu, and S. Zhan. PCCM-GAN: Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model. *Neurocomputing*, 449:330–341, 2021.

[151] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays. Swapnet: Garment transfer in single view images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018.

[152] H. Rawal, M. J. Ahmad, and F. Zaman. Gc-vton: Predicting globally consistent and occlusion aware local flows with neighborhood integrity preservation for virtual try-on. *arXiv preprint arXiv:2311.04932*, 2023.

[153] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. may 2016.

[154] B. Ren, H. Tang, F. Meng, R. Ding, P. H. Torr, and N. Sebe. Cloth interactive transformer for virtual try-on. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2021.

[155] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor. Garment replacement in monocular video sequences. *ACM Transactions on Graphics (TOG)*, 34(1):1–10, 2014.

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

[157] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[158] D. Roy, S. Santra, and B. Chanda. Lgvton: A landmark guided approach to virtual try-on. *arXiv preprint arXiv:2004.00562*, 2020.

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

[160] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[161] K. Sarkar, V. Golyanik, L. Liu, and C. Theobalt. Style and pose control for image synthesis of humans from a single monocular view, 2021.

[162] S. Schaefer, T. McPhail, and J. Warren. Image deformation using moving least squares. In *ACM SIGGRAPH 2006 Papers*, pages 533–540. 2006.

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

[164] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014.

[165] H.-W. Shen, T.-J. Liu, C.-M. Fan, and K.-H. Liu. Wbtp-vton: Whole body and texture preservation based virtual try-on network. In *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE, 2021.

[166] W. Shen and R. Liu. Learning Residual Images for Face Attribute Manipulation. dec 2016.

[167] S. S. Shim and Y. Lee. Consumer's perceived risk reduction by 3d virtual model. *International Journal of Retail & Distribution Management*, 2011.

[168] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[169] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[170] R. SK, J. AB, V. SD, K. TS, and P. Agarwal. Detail-preserving video-based virtual try-on (dpv-vton). In *Proceedings of the 2023 8th International Conference on Multimedia and Image Processing*, pages 57–66, 2023.

[171] J. M. Sohn, S. J. Lee, and D.-E. Kim. An exploratory study of fit and size issues with mass customized men's jackets using 3d body scan and virtual try-on technology. *Textile Research Journal*, 2020.

[172] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[173] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[174] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[175] J. C. Sweeney and G. N. Soutar. Consumer perceived value: The development of a multiple item scale. *Journal of retailing*, 77(2):203–220, 2001.

[176] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[177] Z. L. Tan, J. Bai, S. M. Zhang, and F. W. Qin. Nl-vton: a non-local virtual try-on network with feature preserving of body and clothes. *Scientific Reports*, 11(1):19950, 2021.

[178] W.-J. Tsai and Y.-C. Tien. Attention-based video virtual try-on. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 209–216, 2023.

[179] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[180] T. T. Tuan, M. R. Minar, H. Ahn, and J. Wainwright. Multiple pose virtual try-on based on 3d clothing reconstruction. *IEEE Access*, 9:114367–114380, 2021.

[181] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[182] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

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

[184] J. Wang, T. Sha, W. Zhang, Z. Li, and T. Mei. Down to the last detail: Virtual try-on with fine-grained details. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 466–474, 2020.

[185] R. Wang, M. Dresner, and X. Pan. Strategic responses to the pandemic: A case study of the us department store industry. *International Journal of Physical Distribution & Logistics Management*, 2022.

[186] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[187] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[188] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[189] Z. Wu, G. Lin, Q. Tao, and J. Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM international conference on multimedia*, pages 293–301, 2019.

[190] Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023.

[191] Z. Xie, Z. Huang, F. Zhao, H. Dong, M. Kampffmeyer, X. Dong, F. Zhu, and X. Liang. Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv preprint arXiv:2207.13475*, 2022.

[192] Z. Xie, Z. Huang, F. Zhao, H. Dong, M. Kampffmeyer, and X. Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021.

[193] Z. Xie, J. Lai, and X. Xie. Lg-vton: Fashion landmark meets image-based virtual try-on. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 286–297. Springer, 2020.

[194] Z. Xie, X. Zhang, F. Zhao, H. Dong, M. C. Kampffmeyer, H. Yan, and X. Liang. Was-vton: Warping architecture search for virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3350–3359, 2021.

[195] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[196] H. Yang, X. Yu, and Z. Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2022.

[197] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.

[198] Y.-I. Yang, C.-K. Yang, and C.-H. Chu. A virtual try-on system in augmented reality using rgb-d cameras for footwear personalization. *Journal of Manufacturing Systems*, 33(4):690–698, 2014.

[199] J. Yao and H. Zheng. Lc-vton: Length controllable virtual try-on network. *IEEE Access*, 2023.

[200] F. Yu, A. Hua, C. Du, M. Jiang, X. Wei, T. Peng, L. Xu, and X. Hu. Vton-mp: Multi-pose virtual try-on via appearance flow and feature filtering. *IEEE Transactions on Consumer Electronics*, 2023.

[201] L. Yu, Y. Zhong, and X. Wang. Inpainting-based virtual try-on network for selective garment transfer. *IEEE Access*, 7:134125–134136, 2019.

[202] R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10511–10520, 2019.

[203] Y. Y. Yuen, K. Narayanasamy, Y. L. Cheng, D. Rasiah, and S. Ramasamy. Consumer's perception towards real-time virtual fitting system. 2017.

[204] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

[205] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[206] T. Zhang, W. Y. Chung Wang, L. Cao, and Y. Wang. The role of virtual try-on technology in online purchase decision from consumers' aspect. *Internet Research*, 2019.

[207] T. Zhang, W. Y. C. Wang, L. Cao, and Y. Wang. The role of virtual try-on technology in online purchase decision from consumers' aspect. *Internet Research*, 29(3):529–551, 2019.

[208] F. Zhao, Z. Xie, M. Kampffmeyer, H. Dong, S. Han, T. Zheng, T. Zhang, and X. Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021.

[209] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM international conference on multimedia*, pages 266–274, 2019.

[210] X. Zhong, Z. Wu, T. Tan, G. Lin, and Q. Wu. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 908–916, 2021.

[211] H. Zhou, T. Lan, and G. Venkataramani. Pt-vton: an image-based virtual try-on network with progressive pose attention transfer. *arXiv preprint arXiv:2111.12167*, 2021.

[212] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016.

[213] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on computer vision*, pages 2223–2232, 2017.

[214] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023.

[215] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai. Progressive Pose Attention Transfer for Person Image Generation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2342–2351. IEEE, jun 2019.

**TASIN ISLAM** is a PhD student in the Department of Computer Science at Brunel University London, UK. He received his B.Sc. in Computer Science from Brunel University London in 2020. His research interests lie in the fields of machine learning, artificial intelligence, image processing, and computer vision, with a particular focus on their applications.

**ALINA MIRON** is a lecturer at Brunel University London. She has a PhD in machine learning in the field of autonomous vehicles from INSA de Rouen and MEng and BEng from Babes-Bolyai University, Romania. She is an artificial intelligence researcher, developer and educator. Her current research interests include computer vision and machine learning applied to medical imagining, analysis of human behaviour from videos and other sensors, action recognition and object detection.

**XIAOHUI LIU** is a Professor of Computing at Brunel University London where he conducts research in artificial intelligence, data science and optimisation, with applications in diverse areas including biomedicine and engineering. Professor Liu has been named as a Highly Cited Researcher since 2014 for nine consecutive years in three different areas: Computer Science, Engineering, or Cross-Field (Clarivate/Web of Science).

**YONGMIN LI (SENIOR MEMBER, IEEE)** received his Ph.D. from Queen Mary, University of London, MEng and BEng from Tsinghua University, China. Before joining Brunel University London, he worked as a research scientist in the British Telecom Laboratories. His research interest covers the areas of data science, machine learning, artificial intelligence, image processing, computer vision, video analysis, medical imaging, bio-imaging, biomedical engineering, healthcare technologies, automatic control and nonlinear filtering. Together with his colleagues, he has won the Most Influential Paper over the Decade Award at MVA 2019 and Best Paper Awards at Bioimaging 2018, HIS 2012, BMVC 2007, BMVC 2001 and RATFG 2001. Dr. Li is a Senior Member of the IEEE, and Senior Fellow of the Higher Education Academy.