

Received 14 February 2024, accepted 19 February 2024, date of publication 21 February 2024, date of current version 29 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368612

 SURVEY

Deep Learning in Virtual Try-On: A Comprehensive Survey

TASIN ISLAM[✉], ALINA MIRON[✉], XIAOHUI LIU, AND YONGMIN LI[✉], (Senior Member, IEEE)

Department of Computer Science, Brunel University London, London, UB8 3PH Uxbridge, U.K.

Corresponding author: Yongmin Li (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

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy[✉].

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 [140]. In order

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

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 Kalthor [132]	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. [59]	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. [61]	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.

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 [86], [101].

Virtual try-on allows for the seamless exploration of various items such as shoes [34], [196], upper clothing [67], [195], lower clothing [58], [189] and more [98], [127]. Some models even offer the convenience of modifying multiple items in a single image simultaneously [199]. 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], [59], [61], [132], 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 [67] and CP-VTON [181], 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 we 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

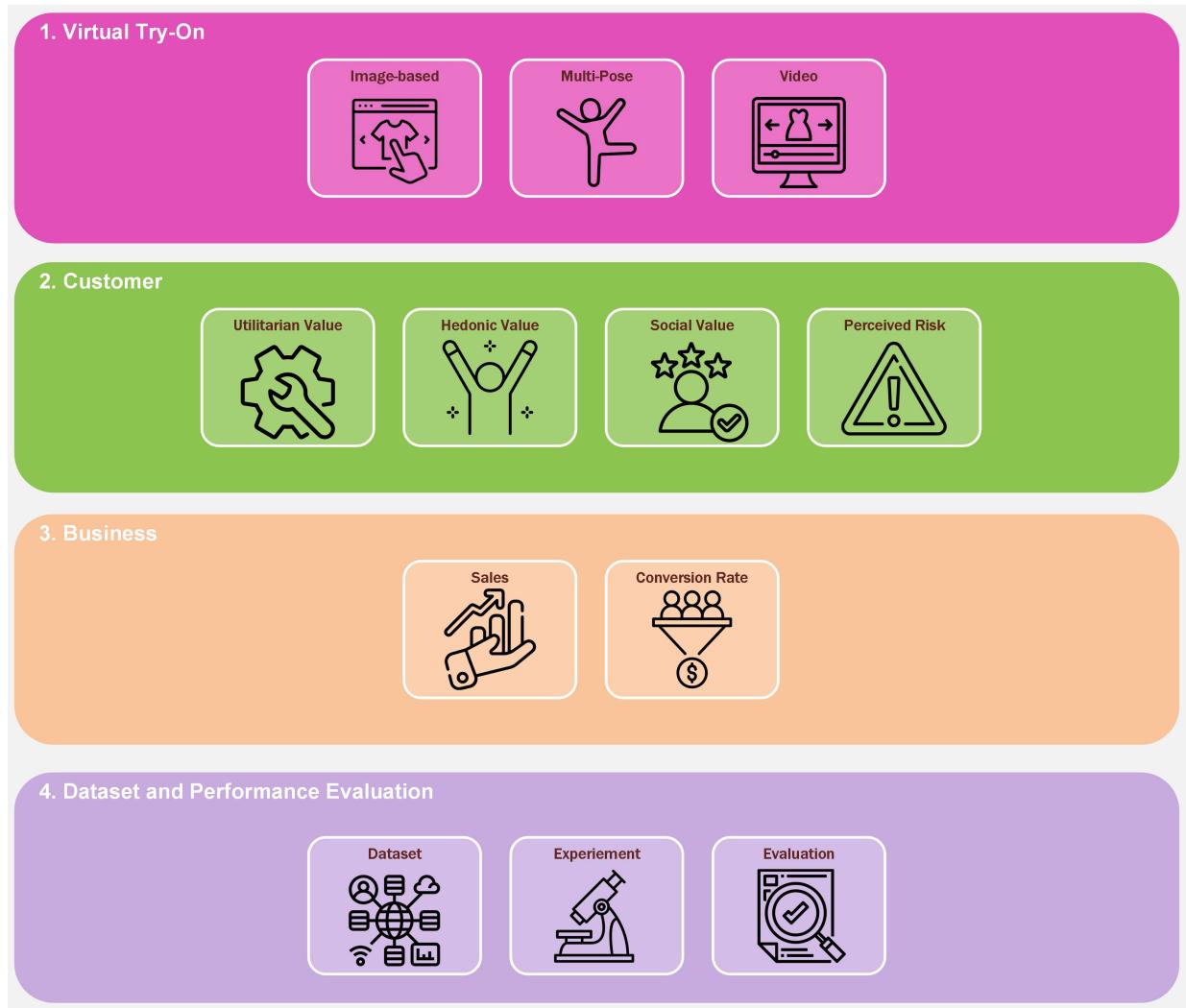


FIGURE 1. The scope of this study. Our survey paper explores image-based, multi-pose, and video virtual try-on models with a focus on technical details. We delve into how these models impact practical values like utility, enjoyment and social, as well as their perceived risks. We analyse how these models 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.

- 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

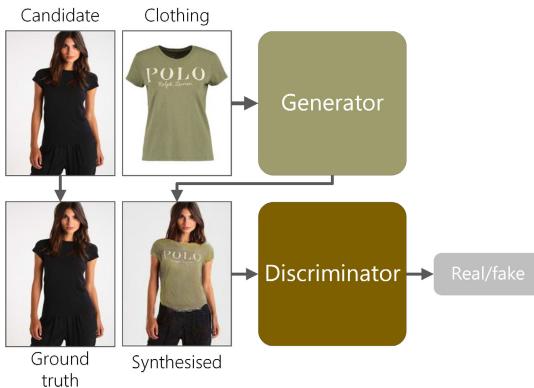


FIGURE 2. cGAN in virtual try-on. This variant of GAN takes in a conditional image to synthesise a desired output image. The discriminator is trained to distinguish real and fake samples, encouraging the generator to produce better results.

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 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 [51], [56]. 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 [62] leverages two neural networks to achieve high-quality image synthesis [84], [95], [96], [143] and manipulation [43], [91], [106]. 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) [131] 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], [138], textual descriptions [141], [148], [151], [193], attributes [164], and sketches [84], [112]. 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 [121], [123], 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 [76], [142], [159], [195], [200]. 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 [181] 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) [74], noise conditioned score network (NCSN) [170] and stochastic differential equation (SDE)

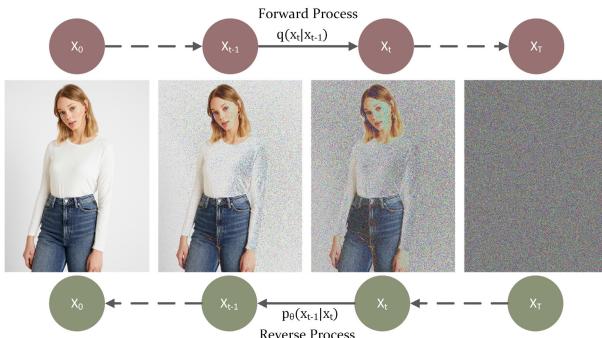


FIGURE 3. Diffusion models for image generation in virtual try-on. It involves two Markov chains: the forward process, which adds gradual noise to data, and the reverse process, which reverses the effects of the forward process using a neural network. x_t denotes the current timestep of noise applied 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.

[171]. 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 [154] or input images [157]. 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], [94], [133], [211]. 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 [64], [144], [146], [162]. Mostly mathematical models are used to manipulate clothing data and create realistic wrinkles [64], [146], [162], but newer models have started using neural networks [144]. The downside of these models is that physics-based algorithms are computationally expensive and difficult to control [144]. Additionally, they require either 3D scans of humans [162] or clothing [146], which is a time-consuming and impractical process.

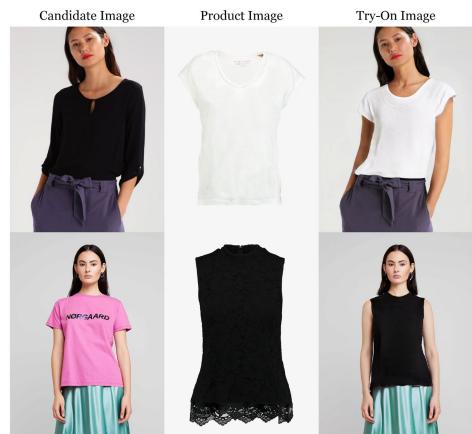


FIGURE 4. Results produced by the state-of-the-art HR-VTON [111], a single-pose image-based virtual try-on model that realistically replaces the candidate's garment with the product item.

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], [59], and [61] 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 item and overlays virtual representations of the clothing onto the user's body. 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 [88] 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 [67] has made this more practical by employing an approach that utilises only the image of target clothing.

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

Method	Authors	Year	Venues	Network type	Warping method	Contribution
CAGAN [88]	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 [67]	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 [181]	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 [76]	Honda	2019	Arxiv	GAN	TPS	Improves CP-VTON [181] by training the model against a discriminator and uses improved loss function to improve the warping.
VTNFP [200]	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 [109]	Lee et al.	2019	ICCV ¹	GAN	TPS	Developed grid interval consistency loss and occlusion-handling techniques for improved geometric matching.
ClothFlow [66]	Han et al.	2019	ICCV ¹	Feature pyramid	OF [45], [172]	Utilises Optical Flows (OF) for warping the clothing image to fit it properly on the person.
CP-VTON+ [130]	Minar et al.	2020	CVPR ²	U-Net	TPS	Improves CP-VTON [181] by fixing the erroneous clothing-agnostic human representations.
ACGPN [195]	Yang et al.	2020	CVPR ¹	GAN	TPS	Predicts segmentation to preserve non-targeted body parts. A constraint was added on the TPS to stabilise the way it warps the garment.
WUTON [85]	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 [129]	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 [87]	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 [191]	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 [209]	Utilises a combination of gated aggregation of hierarchical appearance flow (AF) estimates and dense geometric priors to reduce undesirable output artefacts.
OVNet [115]	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 [205]	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 a self-adaptive pre-alignment strategy to facilitate more accurate geometric matching.
DCTON [57]	Ge et al.	2021	CVPR ²	GAN	TPS	Proposes to disentangle virtual try-ons into multiple modules within the cycle consistency framework [210]. Approach is similar to CAGAN [88].
VTON-HF [47]	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.

TABLE 2. (Continued.) List of single-pose virtual try-on models.

Method	Authors	Year	Venues	Network type	Warping method	Contribution
PF-AFN [58]	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 [192]	Xie et al.	2021	MM ⁹	GAN	AF	Leverages neural architecture search (NAS) [50] to identify optimal neural network architecture to warp clothing in all categories.
NL-VTON [175]	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 [163]	Shen et al.	2021	ICCE-TW ¹¹	U-Net	TPS	Extended the CP-VTON [181] model by allowing it to work on clothes for the entire body rather than upper clothes only.
CIT [152]	Ren et al.	2021	TOMM ¹²	Transformer	TPS	Uses a transformer [179] 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 [70]	He et al.	2022	CVPR ²	StyleGAN	AF	Handles large misalignments between person and garment by using style-based [97] 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 [82], [83]	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 [194]	Yang et al.	2022	CVPR ²	Local gated attention	MLS [160]	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 [54]	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 [134]	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 [48]	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 [111]	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 [139] 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 [125] to extract and fuse global and local features to establish long-range dependencies.
POVNet [116]	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.

TABLE 2. (Continued.) List of single-pose virtual try-on models.

Method	Authors	Year	Venues	Network type	Warping method	Contribution
PG-VTON [53]	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 [97] to synthesise matching body parts.
GP-VTON [188]	Xie et al.	2023	CVPR ²	GAN	AF	Wraps 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 [197]	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 [150]	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 [110]	Lee et al.	2023	MAS ²³	GAN	AF	Wraps 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 [90]	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 [136]	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 [63]	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 [133]	Morelli et al.	2023	Arxiv	LDM	TPS	Uses LDM [154] 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 [149]	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 [124]	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 [187]	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 [135]	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 [156]	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 [159]	Sarkar et al.	2021	Arxiv	GAN	N/A	Modifies the StyleGAN [97] 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 [159], uses a neural network to inpaint appearance features using human body mirror-symmetry.

TABLE 2. (Continued.) List of single-pose virtual try-on models.

Method	Authors	Year	Venues	Network type	Warping method	Contribution
PASTA-GAN [190]	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 [113]	Lewis et al.	2021	TOG ²⁰	StyleGAN	N/A	Learns the internal interpolation coefficients per layer in the StyleGAN2 [97] architecture to synthesise an image. It allows the model to accurately preserve identity features like body shape and skin colour.
TryOn Diffusion [211]	Zhu et al.	2023	CVPR ²	DM	N/A	Proposes Parallel-UNet, which involves training two denoising U-Nets in parallel, with one sending information to the other through cross-attention [179].

¹ IEEE International Conference on Computer Vision² IEEE Conference on Computer Vision and Pattern Recognition³ European Conference on Computer Vision⁴ Asian Conference on Computer Vision⁵ IEEE Winter Conference on Applications of Computer Vision⁶ Chinese Conference on Pattern Recognition and Computer Vision⁷ IEEE International Conference on Tools with Artificial Intelligence⁸ IEEE International Conference on Acoustics, Speech and Signal Processing⁹ ACM International Conference on Multimedia¹⁰Nature Scientific Reports¹¹IEEE International Conference on Consumer Electronics-Taiwan¹²ACM Transactions on Multimedia Computing, Communications, and Applications¹³Image and Vision Computing¹⁴IEEE International Conference on Machine Learning and Applications¹⁵IEEE Transactions on Multimedia¹⁶International Conference on Multimedia Modeling¹⁷ACM International Conference on Multimedia Retrieval¹⁸IEEE International Conference on Consumer Electronics-Asia¹⁹IEEE International Symposium on Mixed and Augmented Reality Adjunct²⁰ACM Transactions on Graphics²¹Advances in Neural Information Processing Systems²²IEEE Access²³MDI Applied Sciences

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 [110], [115], [116], [133], [134], [135].

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], [166], though the type of pose map can vary. References [66], [181], [195], and [152] 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], [82]. References [6], [135], [159], and [187] utilise DensePose [65] 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 [58], [85]. In FIGURE 6, we present the pose heatmaps.

The majority of virtual try-on models use GANs to create photorealistic images [76], [109], [195]. They use discriminators and GAN loss to encourage their modules to ensure accurate segmenting and photorealism of the

image [47], [82], [134], [200]. Other models tend to use a generator only for synthesising the image [67], [191] and use the perceptual loss from a pre-trained VGG network [92] 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 [63], [133], [211].

In the early method of virtual try-on, they would employ a generic convolutional encoder-decoder architecture with skip connections [155] 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], [152], creating normalisation layers or blocks [54], [110], [111], [175], [197] and using feature pyramid networks [66], [136], [188].

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) [209]. Models that use TPS determine their parameters by establishing correspondences between the target clothing image and the



FIGURE 5. Transferring a garment using a virtual try-on. StylePoseGAN [159] demonstrates its capability to extract clothing from a given image and seamlessly apply it to a candidate.

candidate's body key points. To prevent clothing distortion caused by TPS, some models induce constraints [109], [145], [192], [195]. On the other hand, some virtual try-on models have utilised AF to warp the garment image [33], [58], [66]. 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 [124], [149], [187].

Virtual try-on models use a segmentation module to predict how the semantic layout will look after a virtual try-on [135], [175], [195], [200]. 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 using knowledge distillation [72] to train the student network to generate try-on images without relying on the predicted semantic layout [23], [58], [85], [136].

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], [159], [187], [211]. 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.

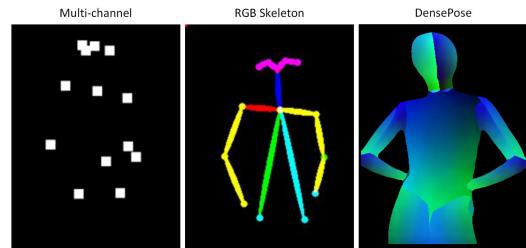


FIGURE 6. Variant of pose heatmaps used by virtual try-on models. Multi-channel heatmaps represent a single key joint in every channel, whereas the RGB skeleton shows how they are connected together. DensePose extracts texture from an image, which can be further manipulated.

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 [88] pioneered the field, they faced limitations. The transition to models like VITON [67] and Swapnet [149] 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 [179], knowledge-distillation [72] 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.

TABLE 3. List of multi-pose 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 [206]	Zheng et al.	2019	MM ²	GAN	TPS	Similar to CAGAN [88], 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 [77]	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 [182]	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 [208]	Zhou et al.	2021	Arxiv	GAN	TPS	Uses pose-attentional transfer network (PATN) [212] and two-stage training procedures for accurate texture transfer and better results.
3D-MPVTON [178]	Tuan et al.	2021	IA ³	GAN	TPS	Extends the capability of CloTH-VTON [129] and proposes a method for mapping the clothing texture that reduces artefacts at clothing boundaries.
MV-TON [49]	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 [69]	He et al.	2022	Arxiv	GAN	N/A	Proposing a single network that utilises the StyleGAN [96] architecture and SPADE [143] layers to generate images.
SPG-VTON [78]	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 ⁶	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 [198]	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 [46]	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 Vision² ACM International Conference on Multimedia³ IEEE Access⁴ IEEE International Conference on Acoustics, Speech and Signal Processing⁵ IEEE Transactions on Multimedia⁷ International Conference on Multimedia Modeling⁸ IEEE Transactions on Consumer Electronics

- 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], [49], [182]. 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], [206] or appearance flow [49], [198]. 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 [206], 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 [63], [133], [211].

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



FIGURE 7. Examples of multi-pose virtual try-on images illustrated by He et al. [69]. The aim of this model is to synthesise a try-on image and change the posture simultaneously.

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 [63], [133], [211].

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 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.

GR-VTON [153] 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 [42], [147]. The first deep learning based approach for producing video virtual try-on is called image2Video [147]. 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 [42], [89], [104] whilst others perform garment transfer [44], [207]. Optical flow (OF) [45], [172] is a common technique utilised by video virtual try-ons [26], [42], [44]. 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 [155]. These models added extra components like a memory module and made use of optical flow to guide the U-Net [42], [147]. However, recent models have achieved better performance by using the transformer model [179] to create try-on videos [80], [89], [176].

The discriminator has an important role in some video generation models to improve the temporal consistency of generated try-on videos [42], [89], [147]. 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 [42], [147], [207].

The evolution of video virtual try-on models has witnessed significant advancements, beginning with GR-VTON [153]

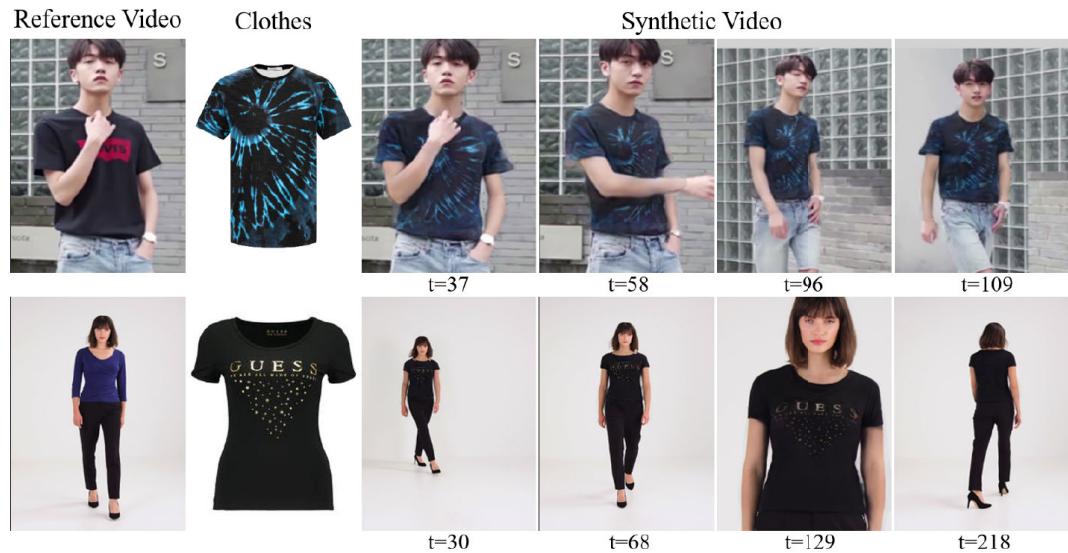


FIGURE 8. Examples of virtual try-on videos produced by ClothFormer [89]. The model aims to synthesise a try-on video using data from the garment.

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 [147] 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 U-Net architectures [155] with additional components. Recent advancements have embraced transformer models [179], 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 [101], [204]. 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.

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 [67], FashionTryOn [206] and MPV [41] 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 introduce 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 [122] 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 [42] 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

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

Method	Authors	Year	Venues	Network type	Warping method	Contribution
GR-VTON [153]	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 [147]	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 [42]	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 [104]	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], [79] and accompanied by a warping module and optical flow module.
MV-TON [207]	Zhong et al.	2021	MM ⁴	GAN	3D AF [117]	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 [44]	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 [89]	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 [80]	Hu et al.	2022	VRCAI ⁶	Transformer	TPS and OF	Uses a cross-modal transformer [177] to capture dependencies between clothing and person representations, thus enabling it to identify important areas and generate coherent videos.
DPV-VTON [168]	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 [176]	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 [179] to enhance the quality of the generated video frame, refine the warped clothing and achieve temporal consistency.

¹ ACM Transactions on Graphics² IEEE International Conference on Computer Vision³ IEEE Winter Conference on Applications of Computer Vision⁴ ACM International Conference on Multimedia⁵ IEEE Conference on Computer Vision and Pattern Recognition⁶ ACM International Conference on Virtual-Reality Continuum and its Applications in Industry⁷ ACM International Conference on Multimedia and Image Processing⁸ ACM International Conference on Multimedia Retrieval**TABLE 5.** Common datasets used on virtual try-on models.

Name	#Training Set	# Testing set	Resolution (HxW)
Single-pose			
VITON [67]	14221	2032	256x192
VITON-HD [30]	11647	2032	1024x768
Dress Code [134]	53000	-	1024×768
Multi-pose			
FashionTryOn [206]	21558	7156	256x192
MPV [41]	52236	10544	256x192
MVC [122]	161260	-	2240x1920
DeepFashion [126]	101966	8750	256x256
Video			
VVT [42]	791	-	256x192
Dwnet [202]	500	100	940x720

from 250 to 300 frames. For training and testing purposes, the authors of the ShineOn framework [104] 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 [202] comprises 500 videos in the training set and 100 videos in the testing set. These videos consist 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



FIGURE 9. A qualitative comparison was carried out by C-VTON [54]. This is a common approach researchers use to demonstrate their model's superiority over previous work on the same pair.

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 [82], [83], [190]. These studies have demonstrated that some models produce inaccurate results, for example, transforming a long-sleeved target garment into a short-sleeved one [82]. 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 [54], 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 [195], WUTON [85], PF-AFN [58], StylePoseGAN [159], HR-VTON [111], DCI-VTON [63], MG-VTON [41], FashionOn [77], TB-VTON [182], FW-GAN [42], MV-TON [207] and video attention-based method [176] 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 [184] 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 (μ_x, μ_y), spread or variability of pixel intensities (σ_x^2, σ_y^2), and linear relationship (σ_{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) [158] 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.

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

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_{KL} , quantifies the dissimilarity between these two distributions. The formula calculates the expected value, denoted by \mathbb{E} , of the Kullback-Leibler divergence over 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 [71], [161] leverages the widely used Inception network [174] 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, μ_P and μ_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 [68]. 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) [203] 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 [129], [137]. To mitigate these issues, researchers conducting 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 [185] or KID [12], which could have provided a more accurate depiction of their model's performance. Due to the lack of consistency in reporting 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 [67]. 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 [195] in 2020, DP-VTON [22] in 2021, and

VTON-SCFA [48] in 2022. As for SSIM, the highest performers are ACGPN [195] 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 [150] in 2023. As for SSIM, the highest performers are VITON-HD [30] in 2021 and GP-VTON [188].

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 [41]. 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 [69] for 2022 and VTON-MP [198] for 2023. The models that performed best in the SSIM metric are MV-TON [49] for 2022 and CF-VTON [46] 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 [89] 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 [51], [56]. 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], [75], provides a significant source of pleasure and motivation for customers, which leads to a positive shopping experience. Several studies have demonstrated that

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: α : SVTON [83]; β : DCTON [57], γ : ST-VTON [32], δ : VTON-HF [47].

Method	Year Published	FID \downarrow	IS \uparrow	LPIPS \downarrow	SSIM \uparrow
VITON dataset					
CAGAN [88]	2017	47.34 δ	2.56 δ	-	0.74 δ
VITON [67]	2018	19.463	2.514	0.0818 α	0.801 α
CP-VTON [181]	2018	24.45 β	2.59 β	-	0.720 β
VTNFP [200]	2019	-	2.78 δ	-	0.80 δ
ACGPN [195]	2020	13.318	2.829	0.0737 α	0.845
CloTH-VTON [129]	2020	-	3.111	-	0.813
CP-VTON+ [130]	2020	16.800	3.105	0.117 γ	0.816
SieveNet [87]	2020	14.65	2.82	-	0.766
ZFlow [33]	2021	15.17	-	-	0.885
OVNet [115]	2021	-	2.846	-	0.852
DP-VTON [22]	2021	8.726	3.044	-	0.871
DCTON [57]	2021	14.82	2.85	-	0.83
VTON-HF [47]	2021	12.63	2.86	-	0.87
PF-AFN [58]	2021	10.09	-	-	-
WAS-VTON [192]	2021	13.83	-	0.0959 γ	0.8430
NL-VTON [175]	2021	14.163	-	-	0.843
CIT [152]	2021	13.97	3.060	-	0.827
FS-VTON [70]	2022	8.89	-	-	0.91
ST-VTON [32]	2022	-	2.9854	0.0725	0.9112
RT-VTON [194]	2022	11.66	-	-	-
SVTON [82]	2022	15.662	2.719	0.0647	0.854
C-VTON [54]	2022	19.535	-	-	-
VTON-SCFA [48]	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 [116]	2023	8.82	2.92	-	0.918
VITON-HD dataset					
VITON-HD [30]	2021	11.74	-	0.053	0.895
DCTON [57]	2021	15.55	2.84	-	0.81
HR-VTON [111]	2022	10.91	-	0.065	0.892
GP-VTON [188]	2023	9.197	-	0.0799	0.8939
LC-VTON [197]	2023	8.21	-	0.073	0.858
GC-VTON [150]	2023	7.888	-	0.0831	0.887
DCI-VTON [63]	2023	9.13	-	0.053	0.892
MPV dataset					
WUTON [85]	2020	7.927	3.154	0.101	0.799
PF-AFN [58]	2021	6.429	-	-	-
M3D-VTON [205]	2021	20.04	-	-	0.8804
C-VTON [54]	2022	4.846	-	0.073	-
SDAFN [9]	2022	5.805	-	-	-

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: α : TB-VTON [182], β : CF-VTON [46].

Method	Year Published	FID \downarrow	IS \uparrow	LPIPS \downarrow	SSIM \uparrow
MPV dataset					
MG-VTON [41]	2019	22.418 α	3.154	0.202 α	0.744
TB-VTON [182]	2020	16.006	3.193	0.187	0.723
SS-VTON [69]	2022	9.34	-	0.170	0.761
MV-TON [49]	2022	11.986	2.981	0.143 β	0.793
SPG-VTON [78]	2022	-	3.243	-	0.752
DO-VTON [24]	2022	14.699	-	-	0.739
VTON-MP [198]	2023	11.986	2.982	0.273	0.7842
CF-VTON [46]	2023	15.371	-	0.227	0.794
DeepFashion dataset					
MG-VTON [41]	2019	-	3.030	-	-
FashionOn [77]	2019	-	3.191	-	0.894
SPG-VTON [78]	2022	-	3.124	-	-

using virtual try-on technology results in a positive hedonic experience for customers [28], [101], [107].

In the fashion industry, customers have a strong desire for exceptional shopping experiences [13], and virtual try-ons

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: α : ClothFormer [89].

Method	Year Published	VFID I3D \downarrow	VFID ResNeXt101 \downarrow	LPIPS \downarrow	SSIM \uparrow
VVT dataset					
FW-GAN [42]	2019	7.052	23.94	0.283 α	0.675 α
MV-TON [207]	2021	8.367 α	9.702 α	0.233 α	0.853 α
ClothFormer [89]	2022	3.967	5.048	0.081	0.921
MMTrans [80]	2022	7.922	9.897	0.155	0.877
FashionVideo dataset [202]					
FashionMirror [26]	2021	3.097	-	0.057	0.923
Dance50k dataset [44]					
wFlow [44]	2022	-	-	0.090	0.920
DeepFashion dataset					
wFlow [44]	2022	-	-	0.187	0.844

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 [108]. 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 [128]. This phenomenon is known as self-congruity, which refers to the tendency to compare oneself with other objects and stimuli [120].

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 [180]. 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 [86]. 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-on models to have a high utilitarian value [10], [28], [101], [108]. 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 [108].

The customers' perception of the utilitarian value towards virtual try-on models is a critical factor in determining their acceptance and usage [38], [128]. 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 [128].

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 [128]. 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 [128].

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 [128]. 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 [101], [204].

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) [101], [204]. The extent to which customers perceive risk plays a significant role in their willingness to use virtual try-ons [204].

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 [105], [204]. Many mature users are hesitant to input sensitive information online, making them less likely to use tools that require it [105].

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 try-ons and perceive these tools to be less risky compared

to adults [201], [204]. 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 [119].

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 [173]. 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 [52]. For this reason, body esteem significantly influences the perceived social value that virtual try-ons can provide [86]. 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 [86], [93]. 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 [101], [204]. 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 [204].

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 [101].

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 [102]. 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 [102], [103]. 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 [101]. By employing virtual try-on models, it is observed that the return rate was significantly decreased [81], and the customers were willing to spend more [59]. 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 [99]. 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 [204]. 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 [183].

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 [55]. 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 [100]. 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 [169]. 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 [165].

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

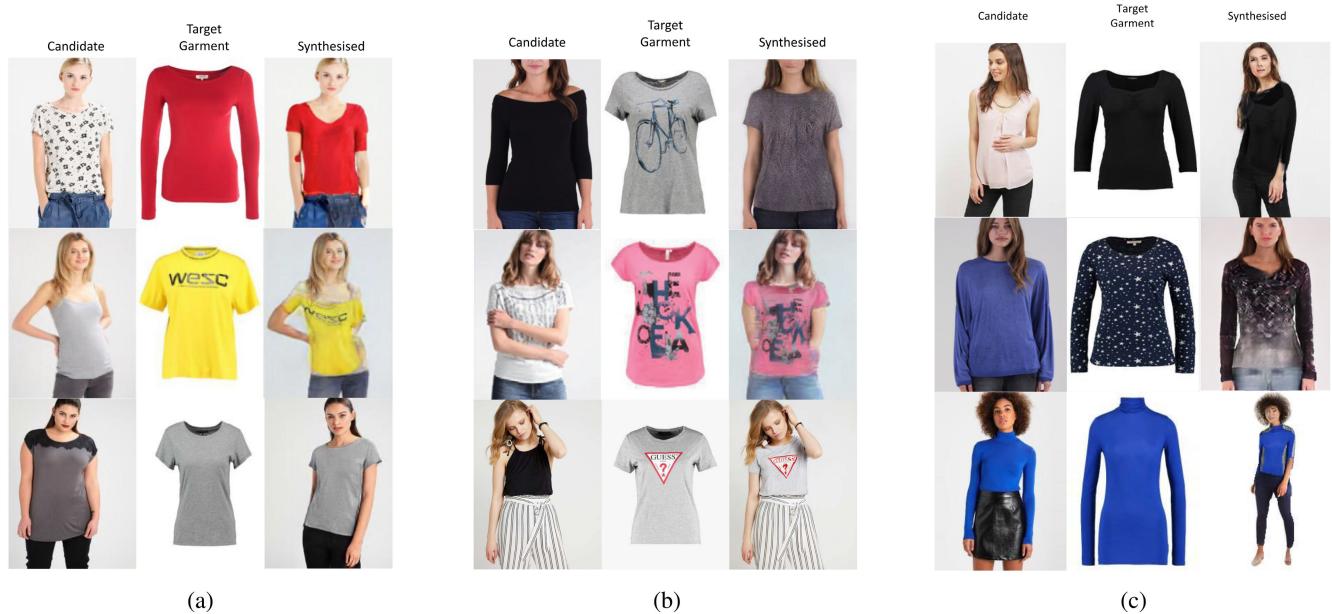


FIGURE 10. Weaknesses presented by virtual try-on models. (a) Some models find it difficult to accurately apply clothing to a person. The first [195] row shows the sleeve has been ruined, and the second [181] row depicts the clothing did not align with the body properly. The third [69] row shows the body shape has inaccurately shrunken in the synthesised image. (b) Capturing and merging content from conditional data is complex. The first [42] 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 [181] row. Even the diffusion model can struggle to capture text and logo from a conditional image, as seen in the third [133] row. (c) Preserving facial identity is extremely challenging for multi-pose and video virtual try-on models. The first [41] and third [182] rows are multi-pose models that did not preserve the facial identity, while the second row [42] is a video model that also failed to preserve the facial identity.

responsible for predicting the semantic layout is not always accurate [82], [85]. 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 [69]. 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. [69] argue that datasets like MPV [41] 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 [181], [195] 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 [53], [54], [111].

In FIGURE 10b, it is evident that some models have generated low-quality virtual try-on images. The first row [42] shows that the model failed to incorporate the bike that appears in the target garment in the synthesised image. The second row [181] displays poor occlusion handling, as the exposed arms have unnaturally blended behind the garment. The last row [133] 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 [181]. 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], [42], [182]. 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 [101], [204] and that minors (under 18) are more likely to engage with and utilise virtual try-on systems compared to adults [204]. 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 [111]. 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], [2], [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 [102]. 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 [101], [204]. 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], [63], [94], [118], [211].

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 [69] 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 [179] and transformer model [179], 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 [60], [73], [167], [186].

Virtual try-ons have been introduced as an alternative to physical fitting rooms, but not many customers have embraced this technology yet [86], [114]. 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. CONCLUSION

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] *Official Hugo Boss Website*. Accessed: Nov. 12, 2023. [Online]. Available: <https://www.ray-ban.com>
- [2] *Official Ray-Ban Website*. Accessed: Nov. 12, 2023. [Online]. Available: <https://www.hugoboss.com/uk/all-brands/men/new-in/virtual-try-on/>
- [3] *Virtual Try On for Hair Colour & Makeup | L'Oréal Paris UK*. Accessed: Nov. 12, 2023. [Online]. Available: <https://www.loreal-paris.co.uk/virtual-try-on>
- [4] (Jun. 2023). *Fashion E-Commerce Worldwide—Statistics & Facts*. [Online]. Available: <https://www.statista.com/topics/9288/fashion-e-commerce-worldwide/#topicOverview>
- [5] S. Adhikari, B. Bhushal, P. Ghimire, and A. Shrestha, "VTON-IT: Virtual try-on using image translation," 2023, *arXiv:2310.04558*.
- [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 Trans. Graph.*, vol. 40, no. 6, pp. 1–11, Dec. 2021.
- [7] A. Ayanso and R. Yoogalingam, "Profiling retail web site functionalities and conversion rates: A cluster analysis," *Int. J. Electron. Commerce*, vol. 14, no. 1, pp. 79–114, Sep. 2009.
- [8] B. J. Babin, W. R. Darden, and M. Griffin, "Work and/or fun: Measuring hedonic and utilitarian shopping value," *J. Consum. Res.*, vol. 20, no. 4, pp. 644–656, Mar. 1994.
- [9] S. Bai, H. Zhou, Z. Li, C. Zhou, and H. Yang, "Single stage virtual try-on via deformable attention flows," in *Proc. Eur. Conf. Comput. Vis. Tel Aviv*, Israel: Springer, 2022, pp. 409–425.
- [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 *Proc. Int. Textile Apparel Assoc. Annu. Conf.*, vol. 73. Iowa, USA: Iowa State Univ. Digital Press, 2016, pp. 1–2.
- [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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5968–5976.
- [12] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," 2018, *arXiv:1801.01401*.
- [13] M. Blázquez, "Fashion shopping in multichannel retail: The role of technology in enhancing the customer experience," *Int. J. Electron. Commerce*, vol. 18, no. 4, pp. 97–116, Jul. 2014.
- [14] P. H. Bloch, D. L. Sherrell, and N. M. Ridgway, "Consumer search: An extended framework," *J. Consum. Res.*, vol. 13, no. 1, pp. 119–126, Jun. 1986.
- [15] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.
- [16] A. Borji, "Pros and cons of GAN evaluation measures: New developments," *Comput. Vis. Image Understand.*, vol. 215, Jan. 2022, Art. no. 10329.
- [17] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," Sep. 2018, *arXiv:1809.11096*.
- [18] F. Buttle and S. Maklan, *Customer Relationship Management: Concepts and Technologies*. Evanston, IL, USA: 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," 2023, *arXiv:2302.06826*.
- [20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
- [21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.
- [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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2295–2299.
- [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 *Proc. Int. Conf. Multimedia Model.* Phu Quoc, Vietnam: Springer, 2022, pp. 28–40.
- [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 *Proc. Int. Conf. Multimedia Model.* Quoc, Vietnam: Springer, 2022, pp. 82–94.
- [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 *Proc. Int. Conf. Multimedia Retr.*, 2022, pp. 313–321.
- [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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13809–13818.

- [27] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 4, pp. 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," *J. Consum. Behaviour*, vol. 23, no. 1, pp. 107–125, Jan. 2024.
- [29] T. L. Childers, C. L. Carr, J. Peck, and S. Carson, "Hedonic and utilitarian motivations for online retail shopping behavior," *J. Retailing*, vol. 77, no. 4, pp. 511–535, Dec. 2001.
- [30] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14131–14140.
- [31] M. J. Chong and D. Forsyth, "Effectively unbiased FID and inception score and where to find them," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6070–6079.
- [32] Z. Chong and L. Mo, "ST-VTON: Self-supervised vision transformer for image-based virtual try-on," *Image Vis. Comput.*, vol. 127, Nov. 2022, Art. no. 104568.
- [33] A. Chopra, R. Jain, M. Hemani, and B. Krishnamurthy, "ZFlow: Gated appearance flow-based virtual try-on with 3D priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5413–5422.
- [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*. Perth, Australia: Springer, 2019, pp. 654–668.
- [35] A. V. Citrin, D. E. Stem Jr., E. R. Spangenberg, and M. J. Clark, "Consumer need for tactile input: An internet retailing challenge," *J. Bus. Res.*, vol. 56, no. 11, pp. 915–922, 2003.
- [36] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," 2019, *arXiv:1911.03584*.
- [37] F.-A. Croitoru, V. Hundru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.
- [38] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and intrinsic motivation to use computers in the workplace 1," *J. Appl. Social Psychol.*, vol. 22, no. 14, pp. 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," *J. Customer Behaviour*, vol. 9, no. 2, pp. 151–174, Jun. 2010.
- [40] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 8780–8794.
- [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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9026–9035.
- [42] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1161–1170.
- [43] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8120–8128.
- [44] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3480–3489.
- [45] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
- [46] C. Du and S. Xiong, "CF-VTON: Multi-pose virtual try-on with cross-domain fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
- [47] 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 *Proc. IEEE 33rd Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2021, pp. 224–231.
- [48] 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 Trans. Multimedia*, vol. 25, pp. 777–791, 2023.
- [49] C. Du, F. Yu, M. Jiang, X. Wei, T. Peng, and X. Hu, "Multi-pose virtual try-on via self-adaptive feature filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2544–2548.
- [50] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.
- [51] T. Y. Evelina, A. Kusumawati, and U. Nimran, "The influence of utilitarian value, hedonic value, social value, and perceived risk on customer satisfaction: Survey of e-commerce customers in Indonesia," *Bus., Theory Pract.*, vol. 21, no. 2, pp. 613–622, Sep. 2020.
- [52] 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," *Inf. Technol. People*, vol. 34, no. 3, pp. 1152–1174, May 2021.
- [53] 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," 2023, *arXiv:2304.08956*.
- [54] B. Fele, A. Lampe, P. Peer, and V. Struc, "C-VTON: Context-driven image-based virtual try-on network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2203–2212.
- [55] S. Gallino and A. Moreno, "The value of fit information in online retail: Evidence from a randomized field experiment," *Manuf. Service Oper. Manag.*, vol. 20, no. 4, pp. 767–787, Sep. 2018.
- [56] C. Gan and W. Wang, "The influence of perceived value on purchase intention in social commerce context," *Internet Res.*, vol. 27, no. 4, pp. 772–785, Aug. 2017.
- [57] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu, and P. Luo, "Disentangled cycle consistency for highly-realistic virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16923–16932.
- [58] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8481–8489.
- [59] 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 Appl.*, vol. 81, no. 14, pp. 19967–19998, Jun. 2022.
- [60] 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," 2023, *arXiv:2311.10709*.
- [61] P. Goel, K. Mahadevan, and K. K. Punjani, "Augmented and virtual reality in apparel industry: A bibliometric review and future research agenda," *foresight*, vol. 25, no. 2, pp. 167–184, Apr. 2023.
- [62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–14.
- [63] 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 *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7599–7607.
- [64] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "DRAPE: Dressing any person," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Jul. 2012.
- [65] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.
- [66] X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10470–10479.
- [67] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.
- [68] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
- [69] S. He, Y.-Z. Song, and T. Xiang, "Single stage multi-pose virtual try-on," 2022, *arXiv:2211.10715*.
- [70] S. He, Y.-Z. Song, and T. Xiang, "Style-based global appearance flow for virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3460–3469.
- [71] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
- [72] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [73] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.

- [74] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
- [75] D. L. Hoffman and T. P. Novak, "Marketing in hypermedia computer-mediated environments: Conceptual foundations," *J. Marketing*, vol. 60, no. 3, pp. 50–68, Jul. 1996.
- [76] S. Honda, "VITON-GAN: Virtual try-on image generator trained with adversarial loss," 2019, *arXiv:1911.07926*.
- [77] 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 *Proc. ACM MM*, 2019, pp. 275–283.
- [78] B. Hu, P. Liu, Z. Zheng, and M. Ren, "SPG-VTON: Semantic prediction guidance for multi-pose virtual try-on," *IEEE Trans. Multimedia*, vol. 24, pp. 1233–1246, 2022.
- [79] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [80] 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 *Proc. 18th ACM SIGGRAPH Int. Conf. Virtual-Reality Continuum Appl. Ind.*, 2022, pp. 1–8.
- [81] 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*, vol. 8, pp. 189479–189489, 2020.
- [82] T. Islam, A. Miron, X. Liu, and Y. Li, "SVTON: Simplified virtual try-on," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 369–374.
- [83] T. Islam, A. Miron, X. Liu, and Y. Li, "Image-based virtual try-on: Fidelity and simplification," *Preprint*, 2023.
- [84] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 1125–1134.
- [85] 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*. Glasgow, U.K.: Springer, 2020, pp. 619–635.
- [86] A. Ivanov, Y. Mou, and L. Tawira, "Avatar personalisation vs. privacy in a virtual try-on app for apparel shopping," *Int. J. Fashion Des., Technol. Edu.*, vol. 16, no. 1, pp. 100–109, Jan. 2023.
- [87] S. Jandial, A. Chopra, K. Ayush, M. Hemani, A. Kumar, and B. Krishnamurthy, "SieveNet: A unified framework for robust image-based virtual try-on," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2171–2179.
- [88] N. Jetchev and U. Bergmann, "The conditional analogy GAN: Swapping fashion articles on people images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2287–2292.
- [89] J. Jiang, T. Wang, H. Yan, and J. Liu, "ClothFormer: Taming video virtual try-on in all module," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10799–10808.
- [90] H.-W. Jin and D.-O. Kang, "Versatile-VTON: A versatile virtual try-on network," in *Proc. IEEE Int. Conf. Consum. Electronics-Asia (ICCE-Asia)*, Oct. 2023, pp. 1–4.
- [91] Y. Jo and J. Park, "SC-FEGAN: Face editing generative adversarial network with user's sketch and color," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1745–1753.
- [92] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
- [93] J.-Y.-M. Kang and K. K. P. Johnson, "How does social commerce work for apparel shopping? Apparel social e-shopping with social network storefronts," *J. Customer Behaviour*, vol. 12, no. 1, pp. 53–72, Apr. 2013.
- [94] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "DreamPose: Fashion image-to-video synthesis via stable diffusion," 2023, *arXiv:2304.06025*.
- [95] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. 6th Int. Conf. Learn. Represent.*, Oct. 2018, pp. 1–26.
- [96] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
- [97] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, Jun. 2020, pp. 8110–8119.
- [98] K. Kikuchi, K. Yamaguchi, E. Simo-Serra, and T. Kobayashi, "Regularized adversarial training for single-shot virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3149–3152.
- [99] D.-E. Kim, "Psychophysical testing of garment size variation using three-dimensional virtual try-on technology," *Textile Res. J.*, vol. 86, no. 4, pp. 365–379, Mar. 2016.
- [100] J. Kim and S. Forsythe, "Hedonic usage of product virtualization technologies in online apparel shopping," *Int. J. Retail Distribution Manag.*, vol. 35, no. 6, pp. 502–514, May 2007.
- [101] J. Kim and S. Forsythe, "Adoption of virtual try-on technology for online apparel shopping," *J. Interact. Marketing*, vol. 22, no. 2, pp. 45–59, May 2008.
- [102] N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli, "Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda," *Int. J. Inf. Manag.*, Oct. 2023, Art. no. 102716.
- [103] W. H. Kunz and J. Wirtz, "Corporate digital responsibility (CDR) in the age of AI: Implications for interactive marketing," *J. Res. Interact. Marketing*, vol. 18, no. 1, pp. 31–37, Jan. 2024.
- [104] 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 *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2021, pp. 191–200.
- [105] M. Law and M. Ng, "Age and gender differences: Understanding mature online users with the online purchase intention model," *J. Global Scholars Marketing Sci.*, vol. 26, no. 3, pp. 248–269, Jul. 2016.
- [106] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.
- [107] H. Lee and Y. Xu, "Classification of virtual fitting room technologies in the fashion industry: From the perspective of consumer experience," *Int. J. Fashion Des., Technol. Edu.*, vol. 13, no. 1, pp. 1–10, Jan. 2020.
- [108] 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," *Int. J. Fashion Des., Technol. Educ.*, vol. 15, no. 3, pp. 297–307, Sep. 2022.
- [109] H. J. Lee, R. Lee, M. Kang, M. Cho, and G. Park, "LA-VTON: A network for looking-attractive virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3129–3132.
- [110] J. Lee, M. Lee, and Y. Kim, "MT-VTON: Multilevel transformation-based virtual try-on for enhancing realism of clothing," *Appl. Sci.*, vol. 13, no. 21, p. 11724, Oct. 2023.
- [111] 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*. Tel Aviv, Israel: Springer, 2022, pp. 204–219.
- [112] Y. Lei, W. Du, and Q. Hu, "Face sketch-to-photo transformation with multi-scale self-attention GAN," *Neurocomputing*, vol. 396, pp. 13–23, Jul. 2020.
- [113] K. M. Lewis, S. Varadharajan, and I. Kemelmacher-Shlizerman, "TryOn-GAN: Body-aware try-on via layered interpolation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–10, Aug. 2021.
- [114] A. Li and Y. Xu, "A study of Chinese consumers' adoption behaviour toward virtual fitting rooms," *Int. J. Fashion Des., Technol. Educ.*, vol. 13, no. 2, pp. 140–149, 2020.
- [115] K. Li, M. J. Chong, J. Zhang, and J. Liu, "Toward accurate and realistic outfit visualization with attention to details," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15546–15555.
- [116] K. Li, J. Zhang, and D. Forsyth, "POVNet: Image-based virtual try-on through accurate warping and residual," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12222–12235, Oct. 2023.
- [117] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3688–3697.
- [118] Z. Li, P. Wei, X. Yin, Z. Ma, and A. C. Kot, "Virtual try-on with pose-garment keypoints guided inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22788–22797.
- [119] J.-W. Lian and D. C. Yen, "Online shopping drivers and barriers for older adults: Age and gender differences," *Comput. Hum. Behav.*, vol. 37, pp. 133–143, Aug. 2014.
- [120] F. Liu, J. Li, D. Mizerski, and H. Soh, "Self-congruity, brand attitude, and brand loyalty: A study on luxury brands," *Eur. J. Marketing*, vol. 46, no. 7/8, pp. 922–937, Jul. 2012.

- [121] J. Liu, X. Song, Z. Chen, and J. Ma, "MGCM: Multi-modal generative compatibility modeling for clothing matching," *Neurocomputing*, vol. 414, pp. 215–224, Nov. 2020.
- [122] K.-H. Liu, T.-Y. Chen, and C.-S. Chen, "MVC: A dataset for view-invariant clothing retrieval and attribute prediction," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 313–316.
- [123] L. Liu, H. Zhang, Y. Ji, and Q. M. Jonathan Wu, "Toward AI fashion design: An attribute-GAN model for clothing match," *Neurocomputing*, vol. 341, pp. 156–167, May 2019.
- [124] Y. Liu, W. Chen, L. Liu, and M. S. Lew, "SwapGAN: A multistage generative approach for person-to-person fashion style transfer," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2209–2222, Sep. 2019.
- [125] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
- [126] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
- [127] D. Morelli, S. Bianco, and G. Ciocca, "Designing an AI-based virtual try-on web application," *Sensors*, vol. 22, no. 10, p. 3832, May 2022.
- [128] A. Merle, S. Senecal, and A. St-Onge, "Whether and how virtual try-on influences consumer responses to an apparel web site," *Int. J. Electron. Commerce*, vol. 16, no. 3, pp. 41–64, Apr. 2012.
- [129] M. R. Minar and H. Ahn, "CloTH-VTON: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 1–19.
- [130] 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 *Proc. CVPR Workshops*, 2020, pp. 10–14.
- [131] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
- [132] S. O. Mohammadi and A. Kalhor, "Smart fashion: A review of AI applications in virtual try-on & fashion synthesis," *December*, vol. 3, no. 4, pp. 284–304, Nov. 2021.
- [133] D. Morelli, A. Baldari, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "LaDI-VTON: Latent diffusion textual-inversion enhanced virtual try-on," 2023, *arXiv:2305.13501*.
- [134] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2230–2234.
- [135] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert, "Image based virtual try-on network from unpaired data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5183–5192.
- [136] 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 *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2023, pp. 695–700.
- [137] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," 2020, *arXiv:2006.13846*.
- [138] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.
- [139] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
- [140] I. Pachoulakis and K. Kapetanakiss, "Augmented reality platforms for virtual fitting rooms," *Int. J. Multimedia Appl.*, vol. 4, no. 4, pp. 35–46, Aug. 2012.
- [141] 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*, vol. 463, pp. 1–16, Nov. 2021.
- [142] N. Pandey and A. Savakis, "Poly-GAN: Multi-conditioned GAN for fashion synthesis," *Neurocomputing*, vol. 414, pp. 356–364, Nov. 2020.
- [143] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2337–2346.
- [144] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7363–7373.
- [145] D. L. Pham, N. T. Ngyuen, and S.-T. Chung, "Keypoints-based 2D virtual try-on network system," *J. Korea Multimedia Soc.*, vol. 23, no. 2, pp. 186–203, 2020.
- [146] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, Aug. 2017.
- [147] A. Pumarola, V. Goswami, F. Vicente, F. De la Torre, and F. Moreno-Noguer, "Unsupervised image-to-video clothing transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3181–3184.
- [148] Z. Qi, J. Sun, J. Qian, J. Xu, and S. Zhan, "PCCM-GAN: Photographic text-to-image generation with pyramid contrastive consistency model," *Neurocomputing*, vol. 449, pp. 330–341, Aug. 2021.
- [149] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays, "SwapNet: Garment transfer in single view images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 666–682.
- [150] H. Rawal, M. Junaid Ahmad, and F. Zaman, "GC-VTON: Predicting globally consistent and occlusion aware local flows with neighborhood integrity preservation for virtual try-on," 2023, *arXiv:2311.04932*.
- [151] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, May 2016, pp. 1060–1069.
- [152] B. Ren, H. Tang, F. Meng, D. Runwei, P. H. S. Torr, and N. Sebe, "Cloth interactive transformer for virtual try-on," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 4, pp. 1–20, Apr. 2024.
- [153] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor, "Garment replacement in monocular video sequences," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 1–10, Dec. 2014.
- [154] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
- [155] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.
- [156] D. Roy, S. Santra, and B. Chanda, "LGVTON: A landmark guided approach to virtual try-on," 2020, *arXiv:2004.00562*.
- [157] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH*, 2022, pp. 1–10.
- [158] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
- [159] K. Sarkar, V. Golyanik, L. Liu, and C. Theobalt, "Style and pose control for image synthesis of humans from a single monocular view," 2021, *arXiv:2102.11263*.
- [160] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 533–540, Jul. 2006.
- [161] M. Seitzer. (Aug. 2020). *PyTorch-FID: FID Score for PyTorch Version 0.3.0*. [Online]. Available: <https://github.com/mseitzer/pytorch-fid>
- [162] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in *Proc. 5th Int. Conf. 3D Body Scanning Technol.*, Oct. 2014, pp. 406–413.
- [163] 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 *Proc. IEEE Int. Conf. Consum. Electronics-Taiwan (ICCE-TW)*, Sep. 2021, pp. 1–2.
- [164] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 4030–4038.
- [165] S. In Shim and Y. Lee, "Consumer's perceived risk reduction by 3D virtual model," *Int. J. Retail Distrib. Manag.*, vol. 39, no. 12, pp. 945–959, Nov. 2011.
- [166] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4645–4653.

- [167] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," 2022, *arXiv:2209.14792*.
- [168] S. K. Raghav, A. B. Jahnnavi, S. D. Vivek, T. S. Kirtan, and P. Agarwal, "Detail-preserving video-based virtual try-on (DPV-VTON)," in *Proc. 8th Int. Conf. Multimedia Image Process.*, 2023, pp. 57–66.
- [169] J.-M. Sohn, S. 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 Res. J.*, vol. 90, nos. 17–18, pp. 1906–1930, Sep. 2020.
- [170] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.
- [171] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456*.
- [172] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
- [173] J. C. Sweeney and G. N. Soutar, "Consumer perceived value: The development of a multiple item scale," *J. Retailing*, vol. 77, no. 2, pp. 203–220, Jun. 2001.
- [174] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.
- [175] 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," *Sci. Rep.*, vol. 11, no. 1, p. 19950, Oct. 2021.
- [176] W.-J. Tsai and Y.-C. Tien, "Attention-based video virtual try-on," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2023, pp. 209–216.
- [177] 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 *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, p. 6558.
- [178] T. T. Tuan, M. R. Minar, H. Ahn, and J. Wainwright, "Multiple pose virtual try-on based on 3D clothing reconstruction," *IEEE Access*, vol. 9, pp. 114367–114380, 2021.
- [179] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [180] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quart.*, vol. 27, no. 3, pp. 425–478, 2003.
- [181] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.
- [182] J. Wang, T. Sha, W. Zhang, Z. Li, and T. Mei, "Down to the last detail: Virtual try-on with fine-grained details," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 466–474.
- [183] R. Wang, M. Dresner, and X. Pan, "Strategic responses to the pandemic: A case study of the US department store industry," *Int. J. Phys. Distribution Logistics Manag.*, vol. 53, no. 2, pp. 273–294, Mar. 2023.
- [184] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [185] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.
- [186] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 7623–7633.
- [187] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2E-try on net: Fashion from model to everyone," in *Proc. ACM MM*, 2019, pp. 293–301.
- [188] Z. Xiel, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23550–23559.
- [189] 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," 2022, *arXiv:2207.13475*.
- [190] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2598–2610.
- [191] Z. Xie, J. Lai, and X. Xie, "LG-VTON: Fashion landmark meets image-based virtual try-on," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Nanjing, China: Springer, 2020, pp. 286–297.
- [192] 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 *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3350–3359.
- [193] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
- [194] H. Yang, X. Yu, and Z. Liu, "Full-range virtual try-on with recurrent tri-level transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3450–3459.
- [195] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7847–7856.
- [196] 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," *J. Manuf. Syst.*, vol. 33, no. 4, pp. 690–698, Oct. 2014.
- [197] J. Yao and H. Zheng, "LC-VTON: Length controllable virtual try-on network," *IEEE Access*, vol. 11, pp. 88451–88461, 2023.
- [198] 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 Trans. Consum. Electron.*, vol. 69, no. 4, pp. 1101–1113, Nov. 2023.
- [199] L. Yu, Y. Zhong, and X. Wang, "Inpainting-based virtual try-on network for selective garment transfer," *IEEE Access*, vol. 7, pp. 134125–134136, 2019.
- [200] R. Yu, X. Wang, and X. Xie, "VTNFP: An image-based virtual try-on network with body and clothing feature preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10510–10519.
- [201] Y. Y. Yuen, K. Narayanasamy, Y. L. Cheng, D. Rasiah, and S. Ramasamy, "Consumer's perception towards real-time virtual fitting system," in *Proc. 6th Int. Conf. Comput. Inform.*, 2017, pp. 311–316.
- [202] P. Zablotckaia, A. Siarohin, B. Zhao, and L. Sigal, "DwNet: Dense warp-based network for pose-guided human video generation," 2019, *arXiv:1910.09139*.
- [203] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
- [204] 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 Res.*, vol. 29, no. 3, pp. 529–551, Jun. 2019.
- [205] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13219–13229.
- [206] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie, "Virtually trying on new clothing with arbitrary poses," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 266–274.
- [207] X. Zhong, Z. Wu, T. Tan, G. Lin, and Q. Wu, "MV-TON: Memory-based video virtual try-on network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 908–916.
- [208] H. Zhou, T. Lan, and G. Venkataramani, "PT-VTON: An image-based virtual try-on network with progressive pose attention transfer," 2021, *arXiv:2111.12167*.
- [209] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 286–301.
- [210] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2223–2232.

- [211] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4606–4615.
- [212] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, “Progressive pose attention transfer for person image generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2342–2351.



TASIN ISLAM received the B.Sc. degree in computer science from Brunel University London, U.K., in 2020, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include the fields of machine learning, artificial intelligence, image processing, and computer vision, with a particular focus on their applications.



ALINA MIRON received the B.Eng. and M.Eng. degrees from Babes-Bolyai University, Romania, and the Ph.D. degree in machine learning in the field of autonomous vehicles from INSA de Rouen. She is currently a Lecturer with Brunel University London. She is also an artificial intelligence Researcher, a Developer, and an Educator. Her current research interests include computer vision and machine learning applied to medical imagining, the analysis of human behavior from videos and other sensors, action recognition, and object detection.



XIAOHUI LIU is currently a Professor in computing with Brunel University London, where he conducts research in artificial intelligence, data science and optimization, with applications in diverse areas, including biomedicine and engineering. He has held senior visiting positions in Leiden, Harvard, and Chinese Academy of Sciences, advised U.K. Research Councils on data analytics, genomics, and security and the Royal Statistical Society/the Institute and Faculty of Actuaries on Statistical Education, U.K. schools in light of big data. He founded the International Symposium Series on Intelligent Data Analysis, from 1995 to 1997, served on the international panel to assess the quality of computer science research in the Netherlands, from 2015 to 2016, and since 2014, he has been named as a Highly Cited Researcher for ten consecutive years in computer science, engineering, or cross-field (Clarivate/Web of Science).



YONGMIN LI (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from Tsinghua University, China, and the Ph.D. degree from the Queen Mary University of London. Before joining Brunel University London, he was a Research Scientist with the British Telecom Laboratories. His research interests include 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 First Prize in MICCAI RETOUCH Challenge (Online) 2023, the Second Prize in MICCAI FeTA Challenge 2022, the Most Influential Paper over the Decade Award at MVA 2019, and the Best Paper Awards at Bioimaging 2018, HIS 2012, BMVC 2007, BMVC 2001, and RATFG 2001. He is a Senior Fellow of the Higher Education Academy.