Virtual Try-On-Enhancing Fashion Exploration for Gen-Z

Anushka Mane
Department of Artificial Intelligence and Data Science
AISSMS Institute of Information Technology, Pune, Maharashtra, India
anushkamane.10@gmail.com

Sammrudhi Navaghane
Department of Artificial Intelligence and Data Science
AISSMS Institute of Information Technology, Pune, Maharashtra, India
samnavaghane11@gmail.com

Eeshan Prabhu
Department of Artificial Intelligence and Data Science
AISSMS Institute of Information Technology, Pune, Maharashtra, India
eeshanprabhu29@gmail.com

Atharva More
Department of Artificial Intelligence and Data Science
AISSMS Institute of Information Technology, Pune, Maharashtra, India
moreatharva193@gmail.com

Prof. Snehal Bagal
Department of Artificial Intelligence and Data Science
AISSMS Institute of Information Technology, Pune, Maharashtra, India
snehal.bagal@aissmsioit.org

Abstract

Virtual Try-On (VTON) is a rapidly emerging technology designed to digitally visualize how garments might appear when worn by individuals, thus transforming traditional fashion and retail experiences. This survey paper meticulously explores various state-of-the-art techniques employed in VTON systems, primarily focusing on image-based, 3D-based, and hybrid approaches. Initially, the paper introduces fundamental concepts of virtual try-on systems, tracing their historical progression and relevance within modern e-commerce and fashion industries. It systematically categorizes methodologies into distinct groups, highlighting pioneering approaches such as warping methods, generative adversarial networks (GAN), and advanced 3D garment simulation techniques. The paper further emphasizes crucial technologies and datasets pivotal to VTON advancements, including GANs, transformers, diffusion models, and benchmarking datasets like DeepFashion and VITON. In addressing existing limitations, this survey underscores critical challenges such as achieving photorealistic rendering, effectively handling occlusions and diverse human poses, ensuring real-time processing, and generalizing across various fabric textures and garment styles. Moreover, recent innovations and their implications on commercial and real-time applications are thoroughly discussed. Finally, the paper delineates future research directions aimed at enhancing system realism, scalability, personalization, and integration of emerging generative AI methodologies, highlighting the significant potential for continued innovation and application in the digital retail landscape.

Keywords — *Virtual Try-On, VTON, Image-based VTON, 3D-based VTON, Generative Adversarial Networks (GAN), Deep Learning, Garment Simulation*

I. INTRODUCTION

The rapid growth of e-commerce and the fashion industry has accelerated the need for innovative technologies that provide consumers with interactive and personalized shopping experiences. Among these innovations, Virtual Try-On (VTON) systems have emerged as highly influential solutions capable of significantly enhancing consumer engagement and reducing the uncertainty associated with online garment purchases (Han et al., 2018; Wang et al., 2018). VTON systems allow users to virtually visualize how clothing items would look on them without physically trying the garments, effectively bridging the gap between online and offline shopping experiences.

Historically, virtual try-on began as a basic technology utilizing simple 2D image processing methods. These early systems primarily used manual landmarks and relied on rigid transformations and image warping to fit garments onto users, resulting in limited realism and scalability (Jetchev & Bergmann, 2017). However, the landscape drastically changed with advancements in artificial intelligence and deep learning, particularly with the introduction of Generative Adversarial Networks (GANs). Modern VTON solutions such as VITON (Han et al., 2018) and CP-VTON (Wang et al., 2018) now utilize GAN-based architectures to achieve higher fidelity, producing realistic textures and maintaining garment characteristics and details.

Further advancements have seen the integration of 3D modeling and simulation technologies, providing more accurate garment fitting and realistic draping dynamics. These 3D-based approaches offer substantial improvements by considering physical properties of fabrics, user-specific body measurements, and real-time interactions, significantly enhancing the overall user experience and accuracy of virtual trials (Patel & Shaikh, 2021).

Despite substantial progress, several significant challenges persist, including the generation of photorealistic results, robustness to pose variations and occlusions, achieving real-time performance, and generalization across diverse clothing types and textures. Addressing these challenges remains critical for the widespread adoption of VTON technologies in both industry and consumer applications.

This survey aims to comprehensively review recent methodologies, outline key challenges, highlight important advancements, and suggest promising future research directions in the domain of Virtual Try-On systems, thereby offering a clear perspective for researchers, developers, and industry professionals alike.

II. LITERATURE SURVEY

VITON (Han et al., 2018) introduced a pioneering coarse-to-fine warping mechanism to synthesize photorealistic try-on images. The approach consists of a geometric matching module that warps the target garment to align with the human body, followed by a refinement network that synthesizes realistic try-on results. The study demonstrated significant improvements in generating natural-looking images, with enhanced preservation of garment details. However, its reliance on predefined keypoints limited its ability to handle complex poses and occlusions.

CP-VTON (Wang et al., 2018) improved upon VITON by introducing a geometric matching module (GMM) that learned to align garments more effectively without requiring explicit keypoint annotations. This resulted in better shape consistency and texture preservation. The method employed a conditional generative adversarial network (cGAN) for fine-grained garment warping and an additional refinement network to improve image realism. The approach was evaluated on the VITON dataset and showed superior garment alignment compared to previous methods. However, it struggled with loose-fitting garments and complex drapes.

ACGPN (Yang et al., 2020) introduced an adaptive content generation and preservation network to enhance realism in virtual try-on systems. The key contribution was an attention-driven mechanism that dynamically adjusted garment alignment while preserving texture details. ACGPN outperformed previous models in handling intricate garment structures and occlusions, making it a more robust solution for real-world applications. However, the model required extensive computational resources for training and inference.

FashionGAN (Zhu et al., 2017) was among the earliest works to leverage GANs for fashion image synthesis. It utilized a multi-stage architecture where clothing items were transformed and synthesized onto different body shapes. The method introduced a novel perceptual loss function that helped improve texture representation. While effective for fashion design applications, it lacked precise garment fitting capabilities required for realistic try-on experiences.

ClothFlow (Han et al., 2019) introduced a flow-based warping technique that improved garment fitting through dense correspondences. Unlike conventional GAN-based warping, ClothFlow utilized optical flow estimation to predict detailed garment deformations. The method showed remarkable accuracy in preserving fabric details but required a significant amount of labeled training data.

C-VTON (Dong et al., 2019) built upon prior VTON architectures by improving coarse-to-fine garment alignment. It incorporated an iterative refinement strategy that progressively adjusted clothing alignment, leading to more seamless try-on results. However, the iterative approach increased computational complexity and inference time.

3D-VTON (Patel & Shaikh, 2021) pioneered the integration of 3D human models with virtual try-on systems. By leveraging physics-based cloth simulation techniques, the model was capable of accurately representing garment draping and fabric movement. This approach significantly improved realism but introduced high computational demands that made real-time processing challenging.

VITON-HD (Choi et al., 2021) extended the original VITON model by incorporating high-resolution generative networks. It leveraged advanced GAN architectures to generate high-definition try-on images, allowing for greater texture fidelity. The approach was particularly useful for e-commerce applications requiring detailed visualization. However, high-resolution processing made it difficult to scale the model for real-time applications.

III. METHDOLOGY

Virtual Try-On (VTON) systems employ a variety of methodologies that can be broadly categorized into three main approaches: **2D Image-Based Try-On, 3D Model-Based Try-On, and Hybrid Approaches**. Each of these methodologies incorporates specific techniques tailored to enhance realism, garment alignment, and computational efficiency.

3.1 2D Image-Based Try-On

2D-based VTON systems focus on modifying existing images to seamlessly overlay new garments onto a target individual. These methods rely on deep learning techniques such as generative adversarial networks (GANs) and spatial transformation networks to generate photorealistic outputs. One of the most commonly used techniques in 2D-based VTON is warping-based garment transfer, where the clothing item is spatially deformed to fit the user's body shape while preserving texture details. The VITON (Han et al., 2018) model introduced a coarse-to-fine warping mechanism that first aligns the garment using a geometric matching module and then refines the output using a GAN-based synthesis model. CP-VTON (Wang et al., 2018) further improved this approach by introducing a geometric matching module GMM) that allowed for more accurate garment alignment while maintaining realistic fabric structures.

Another significant 2D approach involves GAN-based synthesis, where generative models are trained to blend the garment naturally into the person’s body while adjusting texture, shading, and occlusions. ACGPN (Yang et al., 2020) introduced an attention-based generative framework that adaptively refines clothing placement while ensuring photorealism. FashionGAN (Zhu et al., 2017) pioneered the use of GANs for fashion image synthesis by incorporating a perceptual loss function to retain garment details effectively.

3.2 3D Model-Based Try-On

Unlike 2D approaches, 3D-based VTON systems aim to simulate realistic garment draping and fitting using physics-based modeling techniques. These systems generate a 3D representation of the user’s body and overlay clothing items with fabric behavior simulations to account for folds, wrinkles, and body interactions. 3D-VTON (Patel & Shaikh, 2021) was one of the earliest models to integrate detailed 3D human modeling with virtual try-on systems, allowing for better fit prediction and visualization. Another approach, ClothFlow (Han et al., 2019), used optical flow-based warping to align garments more naturally with 3D human models, ensuring realistic draping effects. These models, while more accurate in representing clothing physics, often require high computational power and extensive data collection, making real-time implementation challenging.

3.3 Hybrid Approaches

Hybrid VTON methods aim to merge the benefits of both 2D and 3D approaches by using a combination of GAN-based synthesis and 3D modeling. These methods provide enhanced realism while maintaining computational efficiency. MG-VTON (Dong et al., 2019) introduced a multi-pose guidance framework that allowed garments to be adjusted dynamically based on pose variations, significantly improving garment alignment accuracy. C-VTON (Dong et al., 2019) employed a coarse-to-fine iterative refinement strategy, where a GAN-based synthesis network was combined with physics-driven modeling to ensure better alignment and fabric consistency.

Methodology	Strengths	Limitations
2D Warping-Based	Fast, requires fewer computational resources	Limited in handling complex poses and occlusions
GAN-Based Synthesis	High photorealism, texture retention	May struggle with generalization to unseen clothing
3D Cloth Simulation	Accurate fabric behavior, handles occlusions	Computationally expensive, requires detailed models
Hybrid Approaches	Combines best of both 2D & 3D	Still evolving, needs optimization for real-time performance

Table 1. Comparative Analysis of VTON Methodologies

IV.CHALLENGES AND LIMITATIONS

One of the primary challenges in virtual try-on systems is achieving **photorealism**. Generating high-fidelity images that accurately capture fabric textures, lighting, and natural garment folds remains difficult. Many existing models suffer from blurry results or unrealistic blending of garments onto the body. The ability to simulate fine-grained details such as wrinkles, shadows, and reflections continues to be a major research challenge. Additionally, realistic garment physics, including how fabric stretches, folds, and interacts with body motion, is still an open problem. The use of adversarial training and perceptual loss functions has helped improve realism, but results are often limited by training data quality and generalization issues.

Another critical limitation is **pose and occlusion handling**. Many VTON systems fail to generalize well to diverse human poses, leading to misalignment and distortions. This issue is exacerbated when dealing with occluded body parts, as models often struggle to infer missing information. Current systems primarily rely on pose estimation networks, but these networks may introduce errors, especially in cases of extreme occlusion or unusual body positioning. Some recent models attempt to use multi-view synthesis to better predict occluded regions, but these methods require additional data sources and significant computational overhead. Ensuring robust, pose-invariant, and occlusion-resilient VTON models remains an ongoing challenge.

Real-time processing challenges also hinder widespread adoption. Most high-fidelity virtual try-on methods rely on complex neural networks that demand substantial computational resources, making real-time deployment difficult. Processing an image through a generative network often requires powerful GPUs, limiting accessibility on consumer devices. Some lightweight models use neural network compression techniques, such as knowledge distillation or pruning, but often at the expense of quality. Furthermore, balancing efficiency with accuracy remains difficult, as reducing computational cost often results in artifacts or lower-resolution outputs. Addressing these challenges will require the development of more efficient architectures, optimized rendering pipelines, and specialized hardware accelerations tailored for VTON applications.

V.RECENT INNOVATIONS

Several recent advancements have attempted to address these challenges. **Transformer models** have been increasingly explored for virtual try-on applications, as they provide better feature representations and global context understanding. Unlike convolutional neural networks (CNNs), transformers can capture long-range dependencies, allowing for improved garment alignment and texture preservation. Some recent studies have combined transformers with GAN architectures to enhance realism while ensuring better spatial coherence. These models have shown promising results in generating high-resolution, high-fidelity try-on images, making them an attractive alternative to traditional CNN-based architectures.

Diffusion models have emerged as a groundbreaking innovation for VTON. Unlike traditional GANs, diffusion models progressively refine image generation through iterative noise reduction. This process allows for superior detail retention,

smoother texture synthesis, and improved garment blending. By iteratively denoising a sample from a noise distribution, diffusion models can generate high-quality outputs that capture intricate fabric patterns and realistic lighting effects. Recent implementations have demonstrated the potential of diffusion models in generating photorealistic virtual try-on images with minimal artifacts and improved adaptability across diverse clothing types.

Real-time implementations are becoming increasingly important as VTON applications move toward broader consumer adoption. Optimizing neural network architectures to reduce computational overhead while maintaining image quality remains a key focus area. Techniques such as model quantization, pruning, and knowledge distillation have been leveraged to develop lightweight VTON models capable of running on edge devices and mobile platforms. Additionally, advances in parallelized rendering pipelines and hardware acceleration (e.g., TensorRT and TPU-based optimization) have enabled faster processing speeds, making real-time virtual try-on experiences feasible. Some commercial platforms have already started integrating these optimized models, allowing users to try on garments virtually with near-instant feedback.

VI. FUTURE DIRECTIONS

Future research in virtual try-on systems should focus on **enhancing photorealism and scalability**. Current approaches struggle with generalizing across different fabric types, body shapes, and lighting conditions. Developing models capable of adapting to diverse garment textures while maintaining realistic physics-based simulations will be crucial. One promising avenue is the use of hybrid models that integrate deep learning-based synthesis with physically based rendering techniques to improve realism. Additionally, enhancing datasets with more diverse human and clothing samples will improve model robustness.

Another important research avenue is **self-learning VTON systems** that can adapt to new garments and poses without requiring extensive retraining. Traditional VTON models rely on large-scale supervised datasets, which can be expensive and time-consuming to curate. Leveraging **self-supervised learning and few-shot learning techniques** will enable models to generalize to new clothing items with minimal labeled data. Additionally, reinforcement learning-based optimization strategies could be used to fine-tune models dynamically based on real-world feedback, further improving accuracy and adaptability.

Integration of **multi-modal AI techniques**, including natural language processing (NLP), could allow users to describe their desired outfits and generate personalized recommendations in real-time. For example, combining VTON with conversational AI models would enable users to input natural language queries such as "Show me a formal blazer with a slim fit" and receive interactive virtual try-on results. Additionally, integrating **3D body scanning technology** will enhance personalization by ensuring better garment fit and reducing discrepancies between virtual and real-world try-on experiences.

Another potential direction is the development of **cross-platform VTON solutions** that seamlessly integrate with e-commerce platforms, AR applications, and digital fashion environments in the metaverse. As the fashion industry increasingly embraces virtual experiences, VTON models will need to be optimized for interoperability across various digital ecosystems. Future advancements in **blockchain-based digital clothing ownership** could also introduce new opportunities for virtual fashion markets, allowing users to purchase, trade, and showcase digital garments within immersive environments.

Lastly, **ethical considerations and fairness in VTON systems** should be a priority. Many current models struggle with bias in clothing visualization across different skin tones, body types, and cultural fashion preferences. Future research should focus on making VTON models more inclusive by ensuring diverse training datasets and fairness-aware AI techniques that eliminate potential biases in generated outputs.

V. REFERENCE

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An Image-Based Virtual Try-On Network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7543-7552.

[2] B. Wang, H. Zheng, X. Liang, Y. Chen, and L. Lin, "Toward Characteristic-Preserving Image-Based Virtual Try-On Network," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 589-604

[3] H. Yang, Y. Yu, S. Zhang, W. Liu, Y. Yang, and W. Wang, "Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Network (ACGPN)," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 10487-10496.

[4] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "FashionGAN: Virtual Try-On with Adversarial Generative Networks," IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), vol. 40, no. 8, pp. 2004-2016, 2017.

[5] H. Dong, X. Liang, B. Gong, H. Lai, J. Zhu, X. Shen, and L. Lin,
"MG-VTON: Multi-Pose Guided Virtual Try-On," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 9026-9035.
H. Dong, X. Liang, B. Gong, H. Lai, J. Zhu, X. Shen, and L. Lin,
"C-VTON: Clothing-Conditioned Image-Based Virtual Try-On," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 0-7.

[6] Y. Ge, R. Zhang, H. Zhang, Z. Li, P. Luo, and X. Li, "PF-AFN: Progressive Feature Alignment for Unsupervised Domain Adaptation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 10302-10311.

[7] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "ClothFlow: A Flow-Based Model for Clothed Human Reconstruction," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 10443-10452.

[8] M. V. Patel and S. H. Shaikh, "Advances in Virtual Try-On Systems: A Review," Multimedia Tools and Appl., vol. 80, no. 15, pp. 1-28, 2021.

[9] K. Choi, J. Kim, S. Hwang, and K. Lee, "VITON-HD: High-Resolution Virtual Try-On via Adversarial Learning," IEEE Trans. Multimedia, vol. 23, pp. 2381-2392, 2021.

[10] D. Kim, J. Lee, S. Kim, and J. Park, "Real-Time Virtual Try-On using MobileNet-based GANs," in Proc. ACM Int. Conf. Multimedia, 2021, pp. 456-465.

[11] T. Jetchev and U. Bergmann, "The Conditional Analogy GAN: Swapping Fashion Articles on People Images," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), 2017, pp. 0-7.

[12] 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. (NeurIPS), 2017, pp. 6000-6010.

[13] Ramesh, M. Pavlov, G. Goh, S. Gray, and V. Misra, "DALL-E 2: A Generative Model for Image Synthesis," OpenAI, 2022. Available: https://openai.com/research.

[14] M. Mirza and S. Osindero, "Conditional Generative Adversarial Networks (cGANs)," arXiv preprint arXiv:1411.1784, 2014.