

Advances in Virtual Try-On Systems: A Consolidated Research Review

Shalu Verma

Guest Faculty, Department of Computer Science, Specialization in Artificial Intelligence, University of Delhi

Email: shallverma18@gmail.com

ORCID ID: 0009-0002-7248-2654

Abstract- Despite the explosive growth of fashion e-commerce, the sector has also seen a significant shift in consumer behavior, with product returns rates typically over 30%. This is mainly hindered by the inability to accurately visualize products and the lack of personalization in online shopping. The rise of Virtual Try-On (VTO) systems has offered a game-changing solution, combining AI, computer vision, and AR to seamlessly connect the physical and virtual retail environments. This research survey consolidates and systematically reviews the progress of VTO methods from high fidelity diffusion-based generative models and Vision Transformer alignment to real-time pose-guided methods and 'phygital' IoT integration.

There's a constant underlying conflict between visual fidelity and computational efficiency, including problems with garment physics simulation, maintaining identity preservation, and mobile scalability. Moreover, this review highlights the future direction of the field, focusing on the transition to physics-driven hybrid architectures, culture-inclusive datasets, and privacy-focused pipelines. This paper combines the results of seventeen key studies to give a shared view on the potential impact of VTO technology on the future of digital identity and sustainable consumer experiences in the emerging Metaverse.

Keywords: Virtual Try-On (VTO), GANs, Generative AI, Encoder Decoder Transformers, Diffusion Models, Augmented Reality, Fashion E-commerce, Physics-Based Rendering.

1. INTRODUCTION

With the explosion of fashion e-commerce in today's fast fashion world, it has revolutionized the way consumers shop, but there are limitations – such as high return rates, lack of ability to “try it on” and limited ability to personalize choices – that still stand in the way. But, the commonly used 2D online shopping platforms are based on a single size chart and static product photos, which fail to represent the many different body shapes and poses. Consumer dissatisfaction and inefficient supply chain operations are among the consequences of these limitations, and return rates in fashion ecommerce are average over 30%.

Virtual try-on (VTO) systems are a powerful technology that uses artificial intelligence (AI), augmented reality (AR), computer vision and generative modelling to simulate the virtual dressing of users in a realistic and interactive way [1]-[3]. Such systems attempt to close the gap between physical and online shopping experience by providing personalized dynamic virtual try-on of clothes.

Recent work has shown that VTO technologies are not only giving a boost to garment visualization, but also re-defining the shopping experience by building consumer trust, decreasing decision time and decreasing returns rate [9], [12], [14]. Along with, VTO systems are also being integrated with the recommendation

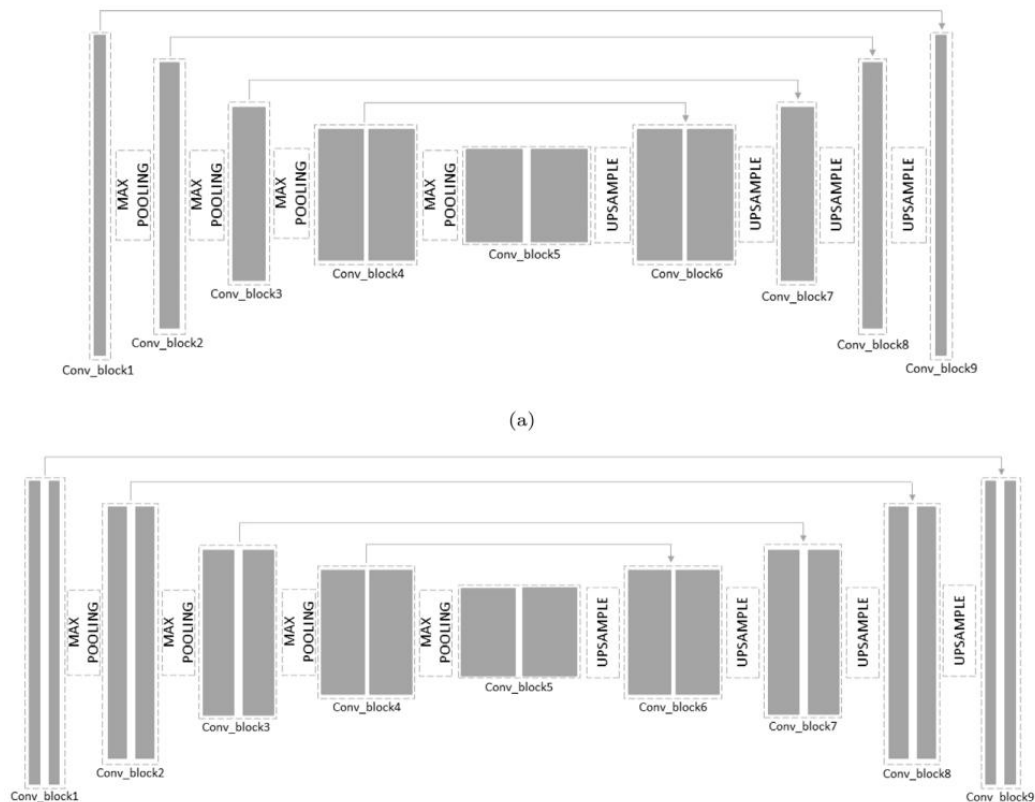
Furthermore, VTO systems are more and more becoming built-in with recommendation engines, social media, and mobile apps, and can be used by a larger target market. This is a consolidated review that combines the results of various studies into a single review and provide a consolidated view on the evolution of VTO systems, methods, limitations and directions for future research.

2. RELATED WORK

Within the Researches there are a number of different methodological categories, each with their own strengths to the ecosystem:

- **Diffusion-Based Generative Models:**

The latent diffusion architecture, such as ARStyle [12] and StyleTailor [8] and SiCo [15] etc., can capture the texture of garments to produce highly photo-realistic images, while supporting the size-adjustable overlay. These models employ the techniques of skip connections, IP-Adapters, and semantic segmentation to preserve garment fidelity. User features are maintained via identity preserving mechanisms like facial embedding constraints and body shape regularization. But diffusion models are also very computation-heavy, on high-end GPUs and with long inference times, making them less scalable in real time and less mobile.

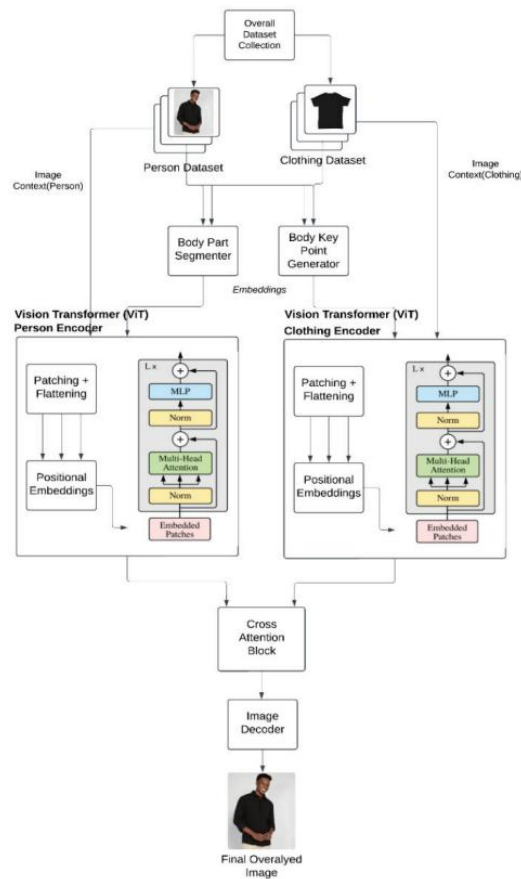


Ref : [12] Comparison of baseline (a) and enhanced (b) encoder-decoder architectures featuring hierarchical convolutional blocks and skip connections for high-resolution garment synthesis.

Source: Islam, M., et al. (2024). *Image-based virtual try-on*.

- **Vision Transformer-Based Alignment:**

Vision Transformers (ViTs) add global spatial reasoning to the model, which is better suited for complex poses and non-rigid garment deformations than CNNs [14], [16]. The dual-encoder architectures combine the human embedding and garment embedding with cross-attention, which enhances structural consistency and multi-region alignment. However, they require significant amounts of annotated data at large scale and the optimization of the transformer, making them resource-intensive.



Ref: [14] Multimodal ViT-based pipeline showing parallel person-clothing encoding via multi-head attention and feature integration through a cross-attention block for accurate image synthesis.

Source; Devi, R., et al. (2024). *Enhancing virtual try-on with vision transformers and multi-part segmentation*.

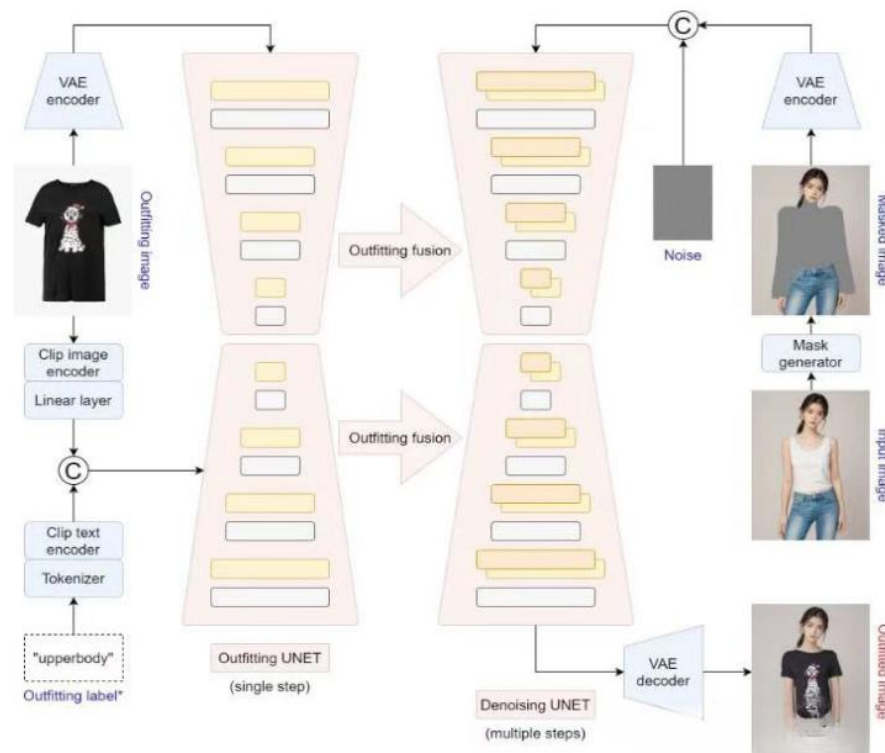
- **Pose-Guided Real-Time Systems:**

MediaPipe, OpenPose and skeleton-based feature extraction are frameworks that provide responsive try-ons [11] and [13]. These systems capture keypoints and skeletal maps to be used to inform garment warping and alignment. They interact with garments dynamically to user motion, together with 3D modeling (such as PIFuHD, SMPL), which allows for interactive experience on consumer devices [2] and [3]. They are quick and compact, but can also be sensitive to changes in lighting and background complexity, which can impact segmentation precision.

- **Phygital Integrations:**

Systems using IoT, AR, and VR combine the physical and digital world. Physics-based rendering provides a simulation of the appearance of fabrics, wrinkles and behavior of fabric in various lighting conditions, to provide immersive environments, which have been validated by users in user testing [9]. Often these systems will include haptic feedback, spatial audio and environmental mapping for increased realism. They do, however, present engineering challenges, as they will need to be integrated across platforms and have sophisticated set-ups.

For realistic outfitting fusion, a VAE-based diffusion network framework is used, with a two-stage U-Net process (outfitting and denoising) as can be seen in [9].



Ref: [9] Diffusion-based VTO framework utilizing VAE encoders and a two-stage U-Net process (outfitting and denoising) for realistic outfitting fusion.

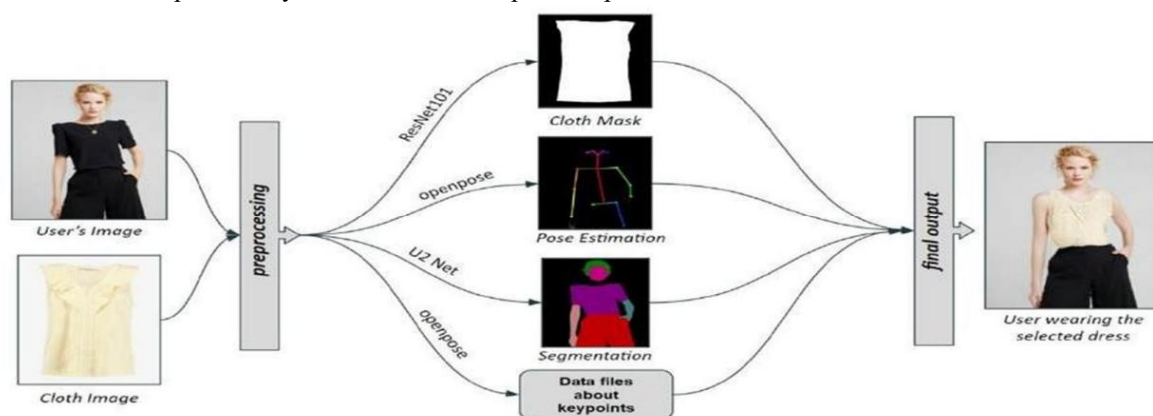
Source: Lin, X. (2024). *Exploration of artificial intelligence and phygital technology in virtual try-ons*.

- **Multi-Agent and Vision-Language Systems:**

Agent-driven frameworks combine reasoning and generative synthesis, allowing for coherent styling and minimizing the occurrence of generative hallucinations [8, 10]. Designer and consultant agents talk to and interpret the user requirements while they are searching for garment options and making selections. Typically these systems are reinforced by reinforcement learning and feedback loops to enhance recommendation accuracy over the time. Methods for recommending cultural fashion items based on culture.

- **Cultural Fashion Recommendation Models:**

CNN-based pipelines with Color Detection and Similarity Retrieval give culturally relevant recommendations, where suggestions are made according to ethnic/regional contexts [4, 17]. Cultural database of garbages and regional classifiers are included in these models. The realism for visualization is still relatively limited, because the garments are still modeled in a simplified way and there is still no pose adaptation.



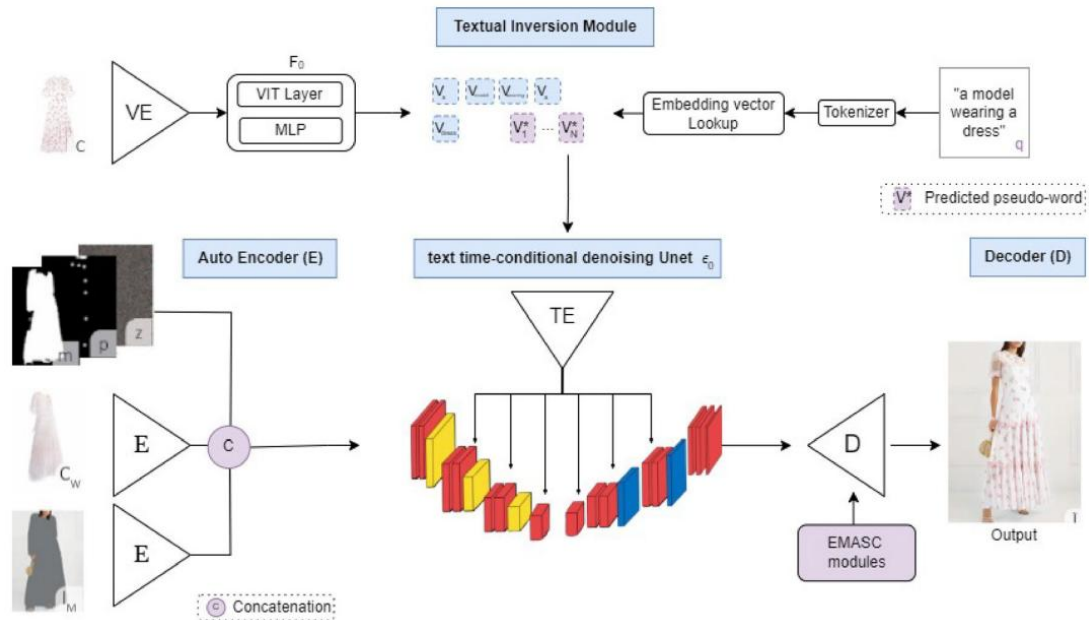
Ref: [17] . Systematic VTO preprocessing workflow incorporating ResNet101 for cloth masking, OpenPose for human pose estimation, and U2-Net for semantic segmentation to guide the final image generation

Source: A, M. (2023). *Online clothing recommendations and virtual trials using AI*.

- **Lightweight & Mobile AR Systems:**

In conjunction with the mobile-first AR solutions and pipelines such as MediaPipe + OpenCV, we can reach high FPS and accessibility, and enable VTO to be deployed in broader consumer applications [2], [3], [5]. These systems focus on speed and usability, and may involve heuristic garment fitting and simplified physics. While highly efficient, they are less effective with dynamics of cloth, multiple layers, and adapting to real-time light conditions. To tackle both challenges, we propose FITMI, which integrates a Textual Inversion Module for pseudo-word prediction and an

EMASC-integrated decoder for realistic garment outfitting. To address both challenges, we introduce FITMI, which incorporates a Textual Inversion Module for pseudo-word prediction along with an EMASC-integrated decoder for realistic garment outfitting.



Ref: [3] FITMI architecture featuring a Textual Inversion Module for pseudo-word prediction and an EMASC-integrated decoder for realistic garment outfitting.

Source: A, M. (2023). *Online clothing recommendations and virtual trials using AI*.

- **Visual Search Integration:**

Using the visual search system based on AI, users can upload images and find similar garments by using the ResNet-50 feature extraction and similarity retrieval algorithm KNN [6], [7]. Integrated with VTO, this enhances product discovery and personalization. These systems depend heavily on the diversity of the datasets and have embedding quality that may result in low performance within niche fashion categories.

3. COMPARATIVE OVERVIEW OF RESEARCH APPROACHES

Approach	Core Model	Techniques Used	Strengths	Limitations	Year
Diffusion-Based	Latent Diffusion, SDXL	Segmentation, skip connections, IP-Adapters	High realism, size-aware overlays	High computational cost	2024–2025
Transformer-Based	Vision Transformers	Cross-attention, multi-part parsing	Strong alignment, pose handling	GPU intensive	2024
Pose-Guided Real-Time	MediaPipe, OpenPose	Skeleton keypoints, garment warping	Fast, responsive, interactive	Moderate realism, lighting sensitivity	2025
Phygital Systems	IoT + AR/VR + PBR	Physics-based rendering	Immersive, reduced return rates	Complex integration	2024
Multi-Agent Systems	Agent Collaboration	Reasoning modules, feedback loops	Coherent styling, reduced errors	Complex orchestration	2024–2025
Cultural Fashion Models	CNN + ML	Color detection, classification	Culturally relevant recommendations	Limited synthesis realism	2023
Mobile AR Systems	MediaPipe + SMPL	3D modeling, AR overlays	High FPS, interactive experiences	Limited cloth dynamics	2024–2025

Visual Search Integration	ResNet-50 + KNN	Feature extraction, similarity retrieval	Improved product discovery	Dataset diversity limitations	2025
----------------------------------	-----------------	--	----------------------------	-------------------------------	------

4. Challenges and Limitations

While significant strides are being made, there are still a number of hurdles:

Virtual Try-On (VTO) technology is available in a range of varying degrees of realism and utility. The current state of Virtual Try-On (VTO) technology is characterized by a persistent trade-off between visual realism and computational practicality. One of the main research challenges is the accurate simulation of garment physics, for which two-dimensional image warping is commonly used but is not capable of simulating multi-layered draping or the elasticity of the cloth itself, causing "interpenetration" artifacts that occur when the virtual garment passes through the body of the user. Such challenges are exacerbated by high level of pose variation and environmental sensitivity (e.g. illumination variations, noise, clutter), which often affect segmentation quality and identity preservation.

Further, switching to Diffusion and Transformer architectures has created a significant computational load, requiring significant VRAM and GPU power which can lead to thermal throttling or loss of fine texture detail when optimizing for mobile.

To overcome these constraints, the next generation of VTO systems is coming into the hybrid, physics-aware era. Future work involves integrating a photorealistic approach with the structural integrity of 3D mesh representations to preserve spatial consistency of garments under multi-angle viewing and dynamic movement.

Researchers are testing real-time edge inference, model distortion and cloud-assisted rendering in order to provide mobile access without compromising fidelity. Additionally, with the sensitive nature of biometric data, there is increasing focus on privacy-preserving pipelines, which leverage federated learning and differential privacy to ensure that sensitive user information remains protected while providing a personalized experience for users.

VTO has extended beyond optimization and is being developed within inclusive and social aspects. One of the most important challenges is the management of a variety of cultural and demographic information to reduce algorithmic bias and make sure that the technology works well for different kinds of body and ethnic outfits. This transformation is moving towards the concept of multimodal personalization, which combines body measurements with context information such as weather conditions, time of day, and occasion to give personalized fashion recommendations. Finally, VTO is projected to become a social experience that will be worked on together, in line with the evolution of the Metaverse and multi-users AR environments. This revolution will enable customers to style their avatars and shop alongside others, creating an interactive and social digital fitting room.

5. Future Scope

Future VTO systems will include the following:

Future development of VTO systems is increasingly focused on the issue of seamlessly connecting aesthetics realism to physical accuracy in physics-informed cloth simulation and hybrid generative modelling. Recent advances include replacing traditional image warping with differentiable physics engines and hybrid mesh/diffusion models to simulate complex fabric behaviors, such as elasticity, dynamic draping, and environmental interactions, like wind effects.

This structural integrity is reinforced by the appearance of hybrid 2D-3D generative models. These architectures combine the high fidelity that 2D diffusion models can offer with the spatial coherence of 3D mesh representations so that a virtual garment remains realistic and coherent across a range of camera angles and multi-angle display setups.

At the same time, scientists are increasingly focused on accessibility and inclusiveness by increasing cultural and demographic datasets. Today, efforts are underway to address the issue of algorithmic bias by incorporating a more diverse range of body shapes, clothing styles, and ethnic outfits to make the technology more widely applicable for users around the world. In order to be able to apply these complex models on consumers' devices, there is much emphasis on real-time optimization. Model distillation, quantization, and edge inference on GPUs are among the techniques used to speed up the Transformer and diffusion architectures for mobiles. Cloud-accelerated rendering pipelines are also being adopted to reduce render times for more complex applications, delegating 3D modelling to a remote server to provide low latency, high quality performance at the user's fingertips despite their local hardware limitations.

The last frontier of VTO is the combination of multimodal personalisation and collaborative social environments. A modern pipeline is becoming more and more like a fashion adviser, and is now gathering more information from a variety of sources such as biometrics, individual preferences, and contextual information like weather and certain social events. With the processing of more sensitive personal information these systems are increasingly facing, the adoption of privacy-preserving techniques like federated learning, differential privacy and secure multi-party computation is becoming essential. Finally, VTO is moving from the personal, to the social. The technology's strides in the digital fitting room are also reshaping it into a collaborative experience

via Metaverse integration and multi-user AR environments, allowing avatars to style and experience shopping together in VR spaces.

6. Conclusion

Virtual try-on (VTO) systems are quickly reshaping the fashion retail scene by combining artificial intelligence, augmented reality, and computer vision into a unified, user-friendly platform. From simple image projections to complex generative pipelines, these systems have advanced to produce photorealistic, custom, and dynamic garment images. The combination of diffusion models, Vision Transformers, pose-guided approaches, and mobile AR systems have greatly enhanced garment alignment, texture fidelity, and responsiveness.

This review has brought together a variety of research lines, and it has shown the strengths and limitations of each of them in addressing some of the key challenges in VTO, ranging from the preservation of the identity of the user to the flexibility of the pose, to the simulation of realistic cloth dynamics and to the feasibility of cultural relevance. The comparative analysis highlights the compromise between realism, computational efficiency, and accessibility and contributes to a nuanced understanding of the present technological landscape.

Despite these strides, there are still critical challenges with physics simulation, dataset diversity and mobile scalability for VTO systems. This will take science from different fields—AI, fashion design, human-computer interaction and ethics—to fix these gaps. The vision for the future described in this paper is focused on the creation of new generation of VTO platforms that will be technically robust, inclusive, privacy-friendly and socially active.

In conclusion, virtual try-on technologies will transform the realm of digital fashion shopping. VTO systems can help minimise return rates, boost customer satisfaction, and encourage sustainable shopping habits by allowing consumers to visualize garments in a context-rich, culturally sensitive, and physically accurate manner. These systems will continue to grow in maturity, and will be instrumental to the future of retail, digital identity, and immersive consumer experience.

References

- [1] A, M. (2023). *Online clothing recommendations and virtual trials using AI*.
- [2] Agrawal, K. (2025). *AI-enhanced intelligent fashion eCommerce*.
- [3] Bhagat, S., et al. (2025). *AI-powered virtual try-on*.
- [4] Chen, H., et al. (2025). *SiCo*.
- [5] Devi, R., et al. (2024). *Enhancing virtual try-on with vision transformers and multi-part segmentation*.
- [6] Gogate, U., et al. (2024). *ARStyle*.
- [7] Islam, M., et al. (2024). *Image-based virtual try-on*.
- [8] J, S., et al. (2025). *Applying pose-guided deep learning for real-time virtual try-on*.
- [9] Krishnapriya, N., Panchal, D., Khammar, J., & Bhatt, D. (2025). Virtual try-on system using MediaPipe and OpenCV for AI-based clothing overlay. *International Journal on Science and Technology (IJSAT)*, 16(2), 1–10.
- [10] Lin, X. (2024). *Exploration of artificial intelligence and phygital technology in virtual try-ons*.
- [11] Ma, L., et al. (2024). *StyleTailor*.
- [12] Moharir, P., et al. (2025). *VTON-IT*.
- [13] Muppalla, S. (2025). *AI-powered visual search and virtual try-on for e-commerce*.
- [14] S, A., et al. (2023). *Culture craze couture*.
- [15] Samy, A., et al. (2024). *Revolutionizing online shopping with FITMI: A realistic virtual try-on solution*.
- [16] Shakeel, M., et al. (2024). *SyncFit*.
- [17] Shete, S., et al. (2023). *AI based fashion stylist recommendation system*.