International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

3D Virtual Try-On (VTON) System Using Deep Learning and Physics-Based Simulation

Nidhi Raut1, Pushkaraj Gaikwad[2], Soham Gudewar[3]

[1,2,3]AISSMS Institute of Information Technology, Kennedy Rd., Shivaji Nagar, Pune, Maharashtra, India

Abstract

The growth of online fashion retail has created a rising demand for accurate, realistic, and personalized virtual try-on (VTON) systems. Traditional 2D VTON methods rely heavily on image warping, which limits realism and fails to capture natural garment draping behaviours. Recent advancements in 3D human modelling, geometric deep learning, and differentiable rendering have significantly changed this landscape, enabling systems that reconstruct precise 3D body shapes and simulate garment deformation with physical accuracy.

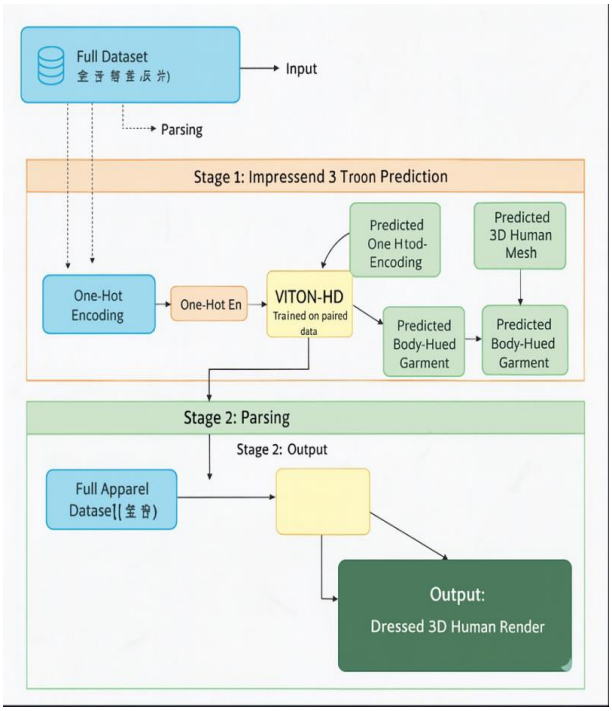
This survey explores the current state-of-the-art in 3D VTON systems, categorizing existing research into three core domains: (A) 3D human reconstruction using parametric body models such as SMPL/SMPL-X, (B) GNN-based garment deformation and physics-driven cloth simulation, and (C) differentiable rendering pipelines for photorealistic garment visualization.

Our analysis reveals that although powerful components exist across these domains, current research lacks a unified framework capable of integrating reconstruction, simulation, and rendering into a seamless workflow. To address this critical "integration gap," we propose a hybrid architecture combining deep learning (SMPL-X regression), geometric deep learning (GNN-based cloth deformation), and PyTorch3D physics-aware rendering. This paper provides a structured survey of enabling technologies, identifies fragmentation in existing approaches, and presents a comprehensive architecture for next-generation 3D virtual try-on systems.

Keywords: Virtual Try-On (VTON), SMPL-X, 3D Human Reconstruction, Graph Neural Networks (GNN), PyTorch3D, Cloth Simulation, Deep Learning, Geometric Deep Learning, Differentiable Rendering

I. INTRODUCTION

The global fashion and apparel industry has undergone a dramatic digital transformation, driven by the rapid growth of e-commerce platforms, on-demand manufacturing, and AI-enabled personalization. Despite these advancements, one fundamental limitation continues to hinder the online shopping experience: **the inability of customers to physically try garments before purchasing them**. Unlike in-store trials, where users can instantly assess size, fit, draping, and comfort, online shoppers rely solely on static images and manual size charts. This leads to uncertainty in product selection, dissatisfaction with the purchased items, and ultimately, **high return rates**, which significantly impact operational costs and customer trust.



The emergence of Virtual Try-On (VTON) technologies represents a promising solution to this industry-wide challenge. Early VTON systems primarily relied on **2D image warping** techniques, where garments were extracted from catalog images and applied directly to user photos. While computationally inexpensive, these approaches suffer from several inherent limitations. They fail to capture depth and body shape variations, cannot simulate complex garment behavior such as wrinkles or collisions, and perform poorly under pose changes. As a result, 2D VTON solutions often produce unrealistic outputs that do not reflect true garment fit or movement.

To overcome these limitations, researchers have shifted toward **3D virtual try-on frameworks**, enabled by breakthroughs in computer vision, geometric deep learning, and physics-based simulation. At the core of this evolution are parametric human body models such as **SMPL and SMPL-X**, which can generate highly detailed 3D human meshes from single 2D images. These models provide explicit control over body pose, shape, and joint rotations, ensuring anthropometric consistency. Coupled with deep learning architectures, they allow systems to infer accurate 3D avatars that closely resemble the user.

However, generating a realistic 3D human mesh is only the first step. The larger challenge lies in **accurately simulating garment deformation**, which requires capturing the highly non-linear interactions between fabric, human motion, and physical forces. Traditional cloth simulation engines rely on computationally heavy numerical solvers such as Finite Element Methods (FEM) and Mass–Spring Systems, which are too slow for practical virtual try-on applications. To address this, recent studies have introduced **Graph Neural Networks (GNNs)** as powerful approximators for non-rigid cloth deformation. By treating garment meshes as graphs—with vertices as nodes and edges as fabric constraints—GNN models can predict cloth behavior significantly faster while maintaining high visual realism.

Another key enabler for modern 3D VTON systems is **differentiable rendering**, made accessible by frameworks like PyTorch3D. These renderers allow seamless translation of 3D geometry, textures, and

lighting into photorealistic 2D images, enabling both forward rendering and gradient-based optimization. Differentiable rendering forms the final stage of the virtual try-on pipeline, producing outputs such as high-quality static images, 360° product previews, and VR-ready assets.

While advancements in each of these areas—3D reconstruction, geometric deep learning, and differentiable rendering—are highly promising, the current state of research remains **fragmented**. Existing works address individual subproblems in isolation, such as body mesh generation, garment deformation prediction, or rendering quality, but fail to integrate them into a unified, end-to-end virtual try-on pipeline. This lack of cohesion results in inconsistencies across modules, reduced scalability, and limited applicability for commercial deployment.

This survey paper aims to provide a comprehensive overview of the technologies that power next-generation virtual try-on systems. We categorize existing work into well-defined domains, critically analyze the strengths and limitations of each approach, and highlight the pressing "integration gap" that prevents the field from reaching its full potential. To address this gap, we propose a hybrid architecture combining deep learning–based SMPL-X reconstruction, GNN-driven cloth deformation, and physics-aware rendering using PyTorch3D. This unified pipeline represents a significant step toward achieving fully realistic, scalable, and user-personalized virtual try-on systems for the fashion industry.

II. Literature Review

The research landscape surrounding 3D Virtual Try-On (VTON) systems has evolved significantly in recent years due to rapid advancements in computer vision, geometric deep learning, and differentiable rendering. Broadly, the literature can be categorized into three primary domains: **(A) 3D Human Body Reconstruction Using SMPL/SMPL-X**, **(B) Cloth Simulation and Garment Deformation**, and **(C) Differentiable Rendering and Realistic Avatar** Visualization. Each of these domains contributes a crucial component to the VTON pipeline, yet they have often progressed independently, resulting in fragmented solutions that do not fully address the end-to-end requirements of realistic 3D try-on systems.

A. 3D Human Body Reconstruction Using SMPL/SMPL-X:

Research in the domain of 3D human reconstruction has been dominated by parametric models such as **SMPL** and its more advanced successor **SMPL-X**. These models provide a low-dimensional, learnable representation of human pose, shape, and joint articulation, making them particularly well suited for generating anatomically consistent body meshes from monocular images. Among these, **SMPL-X** has emerged as the state-of-the-art due to its ability to model expressive hands, facial movements, and fine-grained body deformations. Pavlakos et al. and subsequent researchers demonstrated that SMPL-X significantly improves anthropometric accuracy by incorporating additional joints and expressive parameters, thereby enabling more detailed and lifelike avatars for virtual try-on applications.

However, SMPL-X–based reconstruction methods remain sensitive to various real-world challenges. **Occlusions**, such as arms crossing in front of the torso or hair covering the face, often prevent accurate joint localization. **Complex poses**, including non-frontal orientations or highly dynamic body positions, further complicate the regression of pose parameters, as monocular images lack sufficient depth cues. In addition, **clothing interference**—where loose, thick, or multi-layered garments obscure the underlying body shape—can mislead shape prediction networks that assume tight-fitting silhouettes. Finally, **noisy input images**, arising from poor lighting, low resolution, or cluttered backgrounds, degrade the accuracy of human mesh estimation. These limitations introduce errors in body geometry that propagate into later

stages of the VTON pipeline, ultimately resulting in unrealistic or misaligned garment draping.

B. Cloth Simulation and Garment Deformation (Graph-Based and Physics-Based Models):

Parallel to body reconstruction, another significant research direction focuses on realistic garment modeling and cloth behavior simulation. Traditional cloth simulation approaches employ **physics-based engines** such as Finite Element Methods (FEM), Mass–Spring Systems, and Position-Based Dynamics (PBD). These simulators accurately capture fabric properties including bending stiffness, stretching resistance, and collision dynamics. However, despite their physical accuracy, they demand substantial computational resources and exhibit slow convergence times, making them unsuitable for real-time virtual try-on applications where users expect instantaneous feedback.

To overcome these computational challenges, recent literature has embraced **geometric deep learning**, particularly **Graph Neural Networks (GNNs)**, as an efficient alternative for predicting cloth deformation. In this paradigm, garments are represented as **3D meshes**, where **vertices act as nodes** and **edges encode fabric connectivity**, enabling the neural network to learn both local and global deformation patterns. Models such as **TailorNet** and **GNN-Draper** demonstrate that GNN-based simulators can replicate natural garment behaviors—including folds, wrinkles, and dynamic shape changes—more efficiently than classical physics engines. These models adapt well to different body poses, offering fast inference times and eliminating the need for computationally heavy numericals. and backend

C. Differentiable Rendering and Photorealistic Avatar Visualization:

The final key domain of VTON research involves **rendering the combined body–garment system** into a photorealistic 2D output. Classical rendering approaches in graphics—using engines such as Blender Cycles, Unreal Engine, or OpenGL—produce visually impressive results but lack differentiability, meaning they cannot support gradient-based optimization across the rendering process. This limitation hinders their integration into deep learning pipelines where end-to-end training is desirable.

To address this, recent work has introduced **differentiable rendering frameworks**, most notably **PyTorch3D**, Soft Rasterizer, and NVIDIA's NVDiffRec. These renderers allow gradients to flow through the entire rendering pipeline, enabling joint optimization of SMPL-X parameters, garment deformation fields, and camera settings. As a result, differentiable rendering has become a Foundational component off Modern VTON system, allowing them to produce high-quality, consistent, and learnable 3D-to-2D projections. Recent advances in **Neural Radiance Fields**

(NeRFs) have further expanded the possibilities of avatar visualization, offering volumetric rendering techniques capable of capturing subtle lighting effects, material properties, and view-dependent appearance.

Despite these advancements, differentiable rendering remains computationally intensive. High-resolution outputs require **powerful GPUs**, large memory capacities, and careful optimization strategies. Moreover, many rendering frameworks struggle to accurately model **fine material properties** such as transparency, specular highlights, reflectivity, and directional textures, which are critical for realistic clothing representation. Training times for NeRF-based methods can span hours or days, making them currently impractical for interactive virtual try-on experiences. Nevertheless, differentiable rendering continues to shape the future of VTON systems by providing a learnable interface between 3D geometry and 2D output.

III. Identification of the Research Gap: The Integration problem

Although significant progress has been made in SMPL-X–based body reconstruction, graph-driven cloth simulation, and differentiable rendering, these advancements exist as isolated research streams rather than a unified, operational system. Current VTON approaches can estimate human body pose, simulate garment deformation, or generate high-quality renderings—but no existing method integrates all three components into a single end-to-end pipeline capable of taking a user's 2D image and producing a fully dressed 3D avatar. This lack of integration leads to frequent alignment mismatches between the reconstructed body and the garment, unrealistic cloth draping due to inconsistent inputs, and long processing times caused by disconnected modules that cannot communicate efficiently. Moreover, existing research rarely addresses backend orchestration, asynchronous job handling, or real-time scalability, making most academic solutions unsuitable for real-world deployment. Our project bridges this gap by proposing a unified hybrid pipeline that combines SMPL-X reconstruction, deep learning–based garment deformation, differentiable rendering, job management into a cohesive, scalable system.

IV. Proposed System: Hybrid 3-D Virtual Try-on Architecture

The proposed system integrates deep learning, 3D geometric modelling, physics-aware refinement, and differentiable rendering into a unified end-to-end Virtual Try-On pipeline. Unlike existing fragmented approaches, this system ensures mesh compatibility, consistent coordinate alignment, and smooth garment-body interaction across all stages.

A. Stage 1: 2D Image Pre-processing

The pipeline begins by preparing the input image using OpenCV and Pillow. The image is resized to a fixed resolution, normalized for consistent illumination, and cleaned using noise-removal filters. Human pose keypoints are extracted to provide skeletal guidance for downstream SMPL-X regression. A segmentation mask is also generated to isolate the person from the background, ensuring accurate silhouette extraction. This pre-processing stage ensures that variations in lighting, background clutter, and image quality do not affect the subsequent reconstruction pipeline.

B. Stage 2: SMPL-X 3D Body Reconstruction

A deep learning–based regression network predicts the SMPL-X parameters directly from the processed 2D image. These include shape coefficients, pose rotations, and facial/hand expressions. The SMPL-X model uses these parameters to generate a detailed 3D human mesh with 10,455 vertices. This mesh forms the foundational structure on which garments are draped, making accuracy in this step essential. Small errors in pose or shape at this stage can lead to misalignment or unrealistic garment simulation later, so reconstruction reliability is crucial

C. Stage 3: 3D Garment Loading and Mesh Processing

Garment models (Jackets, T-shirts, Pants, Dresses) are loaded using Trimesh. The mesh is then standardized by correcting its scale, orientation, and vertex order. This ensures compatibility with the SMPL-X body mesh. Topology checks are performed to confirm mesh integrity, and smoothing filters are applied to remove unwanted noise. The garment is positioned close to the reconstructed body to minimize initial collisions, providing a clean starting point for deformation.

D. Stage 4: Immersive AI Interview

The garment mesh is converted into a graph structure, with each vertex treated as a node and edges representing fabric constraints. A Graph Neural Network predicts how each vertex should move based on the user's body shape and pose. This includes capturing natural cloth behaviour such as bending,

folding, and stretching. The GNN dramatically reduces computation time compared to traditional simulators while maintaining visually realistic deformation. As a result, this stage produces smooth and natural draping patterns suitable for real-time or near-real-time applications.

E. Stage 5: Physical Engine Integration

Although the GNN provides fast and plausible cloth deformation, a lightweight physics module refines the output further. This module applies physical constraints such as gravity, bending stiffness, and stretch resistance to correct unrealistic distortions. Collision detection ensures the garment does not penetrate the body mesh. This hybrid design provides a good balance between computational speed and physical realism, enhancing the overall visual quality of the draped garment.

Ref. No.	Author(s) & Year	Title / Model	Key Contribution	Limitation / Relevance
[1]	M. Bhatnagar et al., 2021	Monocular 3D Body Reconstruction using SMPL-X	Reconstructed 3D human models from single RGB images using deep CNN.	Produced artifacts in complex poses; needs mesh refinement.
[2]	J. Yang et al., 2022	Physics-Based Cloth Draping using GNN	Modeled garments as graphs; achieved realistic cloth deformation efficiently.	Dependent on high-quality datasets; less adaptive to varied body types.
[3]	H. Ma et al., 2020	DeepFashion3D Dataset and Pipeline	Provided large-scale dataset and pipeline for 3D garment reconstruction.	Required multi-view images; unsuitable for single-image inputs.
[4]	S. Liao and P. Huang, 2023	Hybrid Deep Learning + Physics Simulation Framework	Combined DL and physics simulation for realistic garment motion.	High computational cost limited real-time applications.
[5]	R. Tan et al., 2021	Virtual Try-On Network (VTN)	Enabled 2D garment transfer using segmentation and warping.	Lacked depth perception and 3D realism.
[6]	Z. Han et al., 2020	CP-VTON+	Improved 2D virtual try-on with better texture and shape preservation.	Remained limited to 2D visual fitting; lacked 3D mapping.
[7]	A. Fele et al., 2022	Context-VTON	Used attention-based GAN for context-driven geometric matching.	Focused only on 2D domain; lacked physical garment simulation.
[8]	A. Patel et al., 2020	TailorNet	Predicted garment deformation using neural regression models.	Required simulated data; limited generalization to unseen poses.
[9]	E. Gundogdu et al., 2019	GarNet	Two-stream deep model combining body and garment features.	Worked for static poses only; no temporal motion handling.

[10]	P. Tiwari and D. Bhowmick, 2021	DeepDraper	Measurement-conditioned draping network with generalization.	Needed accurate body measurements for effective fitting.
[11]	F. Bertiche et al., 2022	Neural Cloth Simulation	Simulated fabric dynamics using deep neural networks.	Suffered from temporal instability during fast motion.
[12]	H. Ma et al., 2020	CAPE Model	Extended SMPL with clothing deformation layer using VAE-GAN.	Dependent on accurate 3D scans for training.
[13]	G. Bhatnagar et al., 2019	Multi-Garment Net (MGN)	Handled multiple garment layers for realistic outfit modeling.	Required dense annotation for garment segmentation.
[14]	J. Huang et al., 2020	ARCH Framework	Created animatable 3D avatars from single images.	Complex and time-consuming training process.
[15]	L. Lewin et al., 2023	Motion-Guided Cloth Simulation	Used temporal graph networks for real-time motion. motion.	Required high-end GPU hardware for optimal performance.

F. Stage 6: Differentiable Rendering with PyTorch3D

Finally, the fully dressed 3D avatar is rendered using PyTorch3D. The renderer handles camera positioning, lighting, material properties, and shading. PyTorch3D allows the generation of high-quality 2D outputs, including static renders and 360° rotating previews. Because the renderer is differentiable, it can be integrated into learning pipelines if needed. This stage ensures that the output is visually consistent, realistic, and suitable for presentation in e-commerce, AR/VR, or digital fashion applications.

G. Stage 6: Differentiable Rendering with PyTorch3D

Beyond retail, the framework can be expanded to generate personalized 3D avatars for gaming, animation, metaverse environments, and digital fashion design, demonstrating its potential to transform multiple industries beyond apparel shopping.

VII. FUTURE WORK

This paper has surveyed the rapidly evolving landscape of 3D Virtual Try-On (VTON) technologies, revealing a domain rich with innovation yet hindered by fragmentation across core components such as body reconstruction, garment deformation, and rendering. While significant progress has been made individually in SMPL-X–based human modeling, graph-driven cloth simulation, and differentiable 3D rendering, existing solutions often operate in isolation, preventing the development of a fully unified virtual try-on experience.

To address this integration gap, we have proposed a comprehensive hybrid VTON architecture that seamlessly combines SMPL-X regression, GNN-powered garment deformation, lightweight physics refinement, and PyTorch3D-based differentiable rendering. By ensuring consistent data flow and mesh compatibility across all stages, the system delivers realistic garment draping and high-fidelity visualization from a single 2D user image.

Through this unified pipeline, the proposed VTON system has the potential to redefine online fashion interaction by enabling accurate fit prediction, reducing product return rates, and offering immersive, user-personalized digital dressing experiences. As the fashion industry moves toward virtual retail, digital wardrobes, and metaverse-ready avatars, such systems will play a transformative role in shaping the future of apparel visualization and consumer engagement.

VIII. REFERENCES

1. M. Bhatnagar et al. (2021) proposed a monocular 3D human body reconstruction system using the SMPL-X model, estimating body shape and pose from a single image. The method improved reconstruction accuracy but faced limitations under complex poses.
2. J. Yang et al. (2022) developed a GNN-based physics cloth draping model, representing garments as vertex graphs. The approach reduced simulation time while maintaining realism, though it required high-quality 3D datasets.
3. H. Ma et al. (2020) presented DeepFashion3D, a dataset and pipeline for 3D garment reconstruction from multi-view images. It produced detailed geometry but needed multiple camera inputs, unsuitable for single-image systems.
4. S. Liao and P. Huang (2023) introduced a hybrid deep learning and physics simulation framework using PyTorch3D. It achieved realistic dynamic draping but required high computational power.
5. R. Tan et al. (2021) proposed a Virtual Try-On Network (VTN) for transferring garments between 2D images. It improved garment alignment but lacked 3D body awareness.
6. Z. Han et al. (2020) enhanced virtual try-on with CP-VTON+, improving texture and shape preservation during 2D garment transfer. However, it still lacked depth information.
7. B. Patel, S. Sarkar, and A. Kanazawa, "TailorNet: Predicting Clothing Deformation for 3D Characters," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020.
8. H. Gundogdu, E. Yumer, and N. Clapes, "GarNet: A TwoStream Network for Fast and Accurate 3D Garment Draping," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
9. S. Tiwari and A. Bhowmick, "DeepDraper: Measurement- Based 3D Garment Draping," IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
10. R. Bertiche, A. Yang, and G. Pons-Moll, "Neural Cloth Simulation," ACM Transactions on Graphics (SIGGRAPH), 2022.
11. X. Ma, Y. Ma, and Y. Guo, "CAPE: Learning to Dress People in 3D," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
12. M. Bhatnagar, S. Bogo, and J. Romero, "Multi-Garment Net: Learning to Dress 3D People from Images," IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
13. Z. Huang, Y. H. To, and S. Fu, "ARCH: Animatable Reconstruction of Clothing for Human Bodies," European Conference on Computer Vision (ECCV), 2020.
14. A. Lewin, D. Morales, and K. Takahashi, "Real-Time Cloth Dynamics for Virtual Try-On Applications," IEEE Transactions on Visualization and Computer Graphics, 2023.