Virtual Try On System On Clothes with Realistic 3D Model Generation

Abstract

The core objective of this research is to design and develop robust models that facilitate the creation of an accurate and user-friendly 3D VTON system. To achieve this, we employ various techniques in computer vision and machine learning to convert 2D user images into 3D representations. This involves sophisticated image processing algorithms and 3D modeling techniques to ensure the avatar accurately reflects the user's appearance and the clothing's fit and style.

Furthermore, this study conducts a comparative analysis of existing 3D VTON models to evaluate their performance, accuracy, and usability. By comparing our developed models with current solutions, we aim to identify strengths and weaknesses, offering insights into how 3D VTON technology can be improved and effectively implemented in e-commerce. The results of this comparative study will provide a comprehensive understanding of the capabilities and limitations of current 3D VTON systems, guiding future advancements in virtual try-on technology.

Introduction

The rapid advancement of technology has profoundly impacted various aspects of our lives, transforming how we interact, work, and shop. E-commerce platforms, in particular, have evolved significantly, offering unparalleled convenience and a wide range of products at our fingertips. As consumer expectations continue to rise, these platforms are constantly seeking innovative ways to enhance the shopping experience. One such innovation is the virtual try-on (VTON) system, which is becoming indispensable in both online shopping and physical retail environments. VTON systems allow users to visualize how garments will look on them before making a purchase, thereby increasing customer satisfaction, reducing return rates, and fostering greater consumer confidence.

The increasing development of machine learning (ML) and artificial intelligence (AI) technologies has been pivotal in the advancement of virtual try-on models. These technologies have enabled the creation of sophisticated VTON systems capable of processing user inputs, such as a single 2D image, to generate detailed and realistic 3D avatars. These avatars can simulate the user wearing selected clothing items, providing a

more immersive and interactive shopping experience. The capabilities of ML and AI in image recognition, pattern analysis, and data processing have been instrumental in overcoming previous limitations of VTON systems, making them more accurate and accessible.

As the VTON technology trend continues to evolve, both paired and unpaired 2D models and 3D priors are being incorporated to enhance performance and accuracy. This study focuses on exploring various state-of-the-art models and techniques to understand their strengths and limitations.

The 2D models studied include SDVTON[1], Size Does Matter[2], Pix2Surf[3] and Pasta-GAN[4] On studying the 3D model , we examined models such as GET 3D Humans[5], SMPL-X[6], Pifu-HD[7], M3D-VTON[8], and EVA3D[9] and Magic123[10]. Additionally, we explored pose detection and parsing algorithms like OpenPose[11], 2D Human Parse[12], and Graphonomy[13]. Based on our comprehensive study, we developed a robust pipeline optimized for GPU devices, enabling the efficient generation of high-quality 3D avatars for virtual try-on applications.

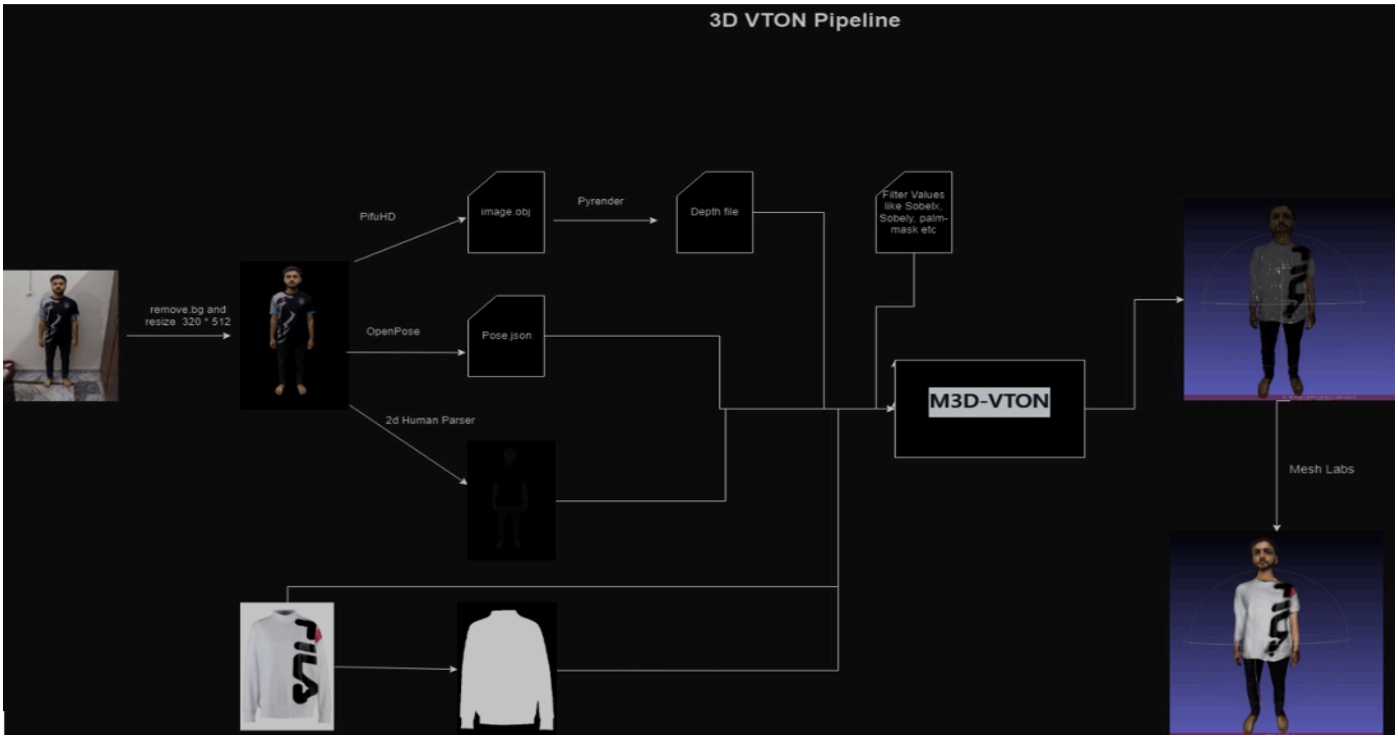
Literature Review

In our search for existing 3D VTON systems, we identified only one notably functional system: M3D-VTON[8]. This system was tested on a custom dataset to evaluate its performance and methodology for creating 3D avatars. The M3D-VTON[8] system employs a multi-step process to generate a 3D avatar, which includes several key components and technologies.

Firstly, the **M3D-VTON[8]** model utilizes PIFuHD[7] to generate a 3D mesh of the user. PIFuHD[7], or Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization, is a state-of-the-art method for creating detailed 3D human shapes from 2D images. This mesh forms the foundational structure of the avatar. Additionally, pose detection is performed using the OpenPose[11] model, which is renowned for its accuracy in detecting human poses and body landmarks. This step ensures that the avatar accurately reflects the user's pose and body orientation.

The system also incorporates segmentation generated from 2D Human Parse, a technique that segments the input image into different clothing and body parts. This segmentation is crucial for accurately mapping the selected clothing onto the avatar. The generated mesh is then refined using MeshLab, an open-source system for processing and editing 3D triangular meshes. MeshLab enhances the mesh quality and ensures that the final output is more visually appealing.

Despite these advanced techniques, the overall results of the M3D-VTON[8] system were suboptimal, even after testing with multiple settings and adjustments. The generated avatars did not consistently achieve the desired level of realism and accuracy. Below is detailed pipel



Building on our research, we explored methods for converting clothing to 3D representations and found **Pix2Surf[3]** to be a promising candidate. Pix2Surf[3] is designed to convert 2D images of garments into 3D models, providing a surface representation that can be applied to virtual try-on systems. To evaluate its effectiveness, we tested the model according to its claims.

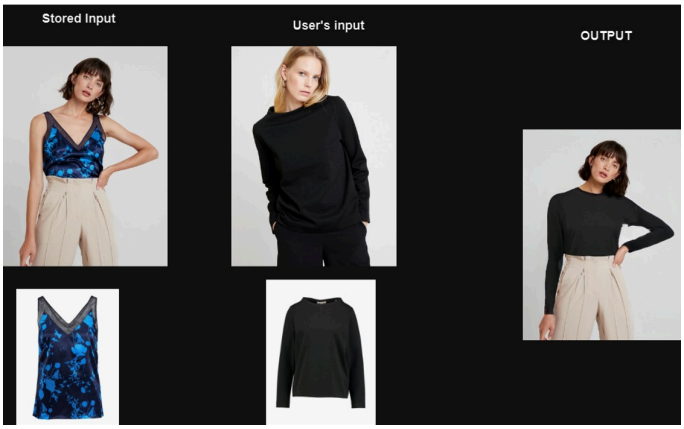
However, our testing revealed significant limitations. On a 16GB GPU, Pix2Surf[3] was constrained to generating outputs with a maximum resolution of 200px by 200px. This resolution is insufficient for creating detailed and realistic 3D garment models required for high-quality virtual try-on experiences. Additionally, we discovered that Pix2Surf[3] is restricted to a predefined set of 16 garment types. This constraint means that all garments of a particular type, such as t-shirts, are represented by a generic body shape, rather than adapting to the unique body shape of each user.

This lack of customization is a critical drawback, as our goal is to generate 3D avatars that accurately reflect the individual user's body and the specific fit of each garment. Despite our efforts to optimize and adapt Pix2Surf[3] for our needs, these limitations rendered the model unsuitable for our application. Below is an example of the output generated by Pix2Surf[3], illustrating the challenges encountered.

Due to the unsatisfactory results from our initial exploration of 3D models, we extended our research to 2D VTON models, with the intention of converting their outputs into 3D models. We identified two main categories of 2D VTON models: those that work with paired datasets and those that work with unpaired datasets.

For paired datasets, we examined **SD-VTON**, a state-of-the-art 2D VTON model. SD-VTON requires both an image of the user and an image of the garment they are wearing in addition to the garment they desire to wear. When tested with paired datasets, SD-VTON performed exceptionally well, producing highly accurate and realistic virtual try-on results. However, our project requirements necessitate using a single 2D image of the user without requiring a separate image of the garment. This limitation means that despite SD-VTON's impressive performance with paired datasets, it is not suitable for our needs due to its inability to work with unpaired datasets.

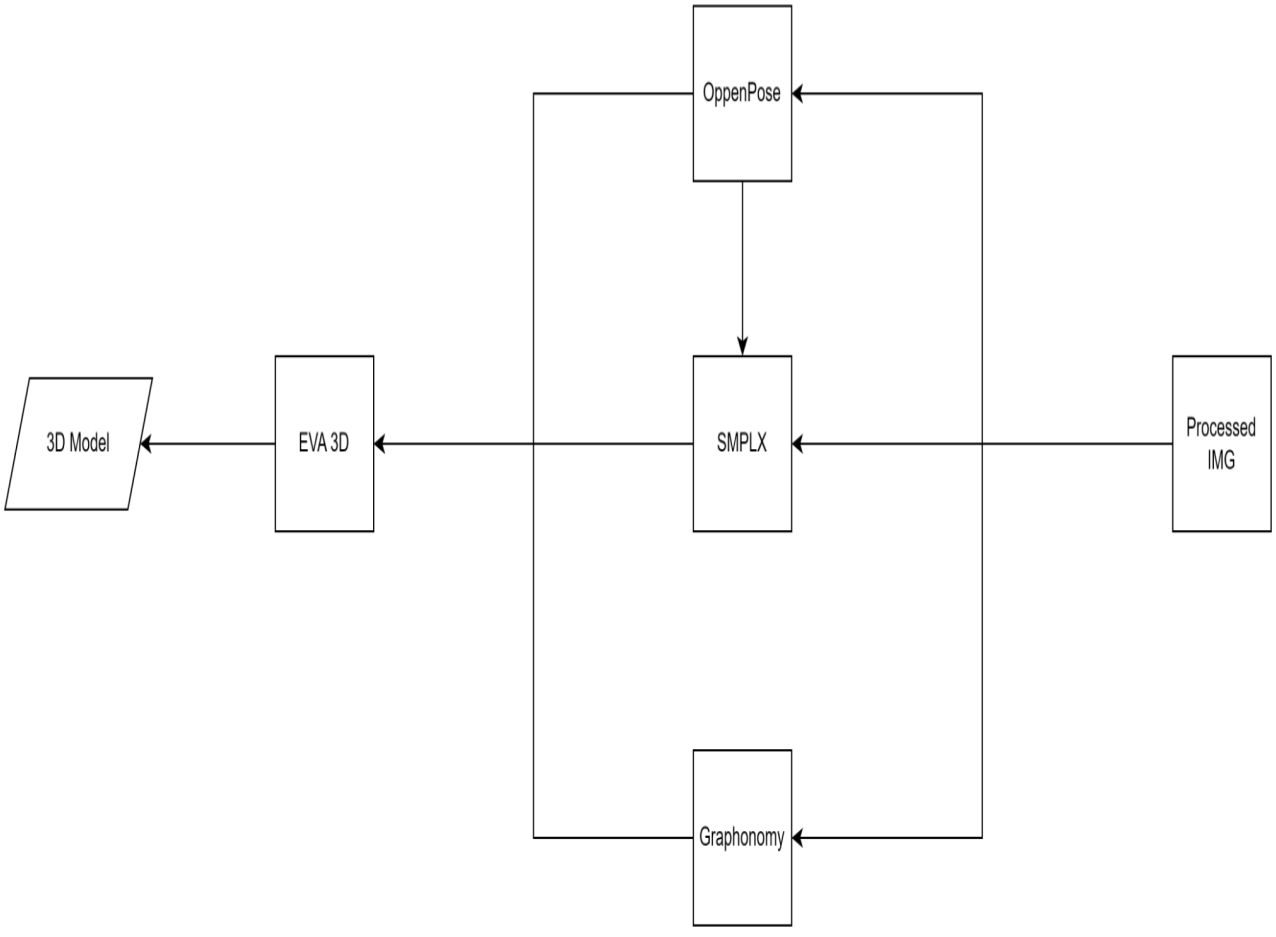
Consequently, we had to dismiss SD-VTON as a viable solution for our project, even though it delivered good results within its operational constraints. Below is a detailed description of SD-VTON's workflow and its performance on paired datasets.



Having determined the type of 2D VTON model that best suited our needs, we proceeded with Pasta-GAN++[4], a continually evolving 2D VTON model known for its robust performance. However, our subsequent challenge involved converting the 2D output of the VTON model into a 3D representation while preserving texture details.

In our quest for state-of-the-art solutions, we explored models such as Magic123[10] and EVA3D[9]. These models are renowned for their ability to convert 2D images into high-fidelity 3D representations. Due to hardware limitations—a 16GB GPU—we opted for EVA3D[9], which offers a comprehensive three-angle view of the 3D model. However, it's noteworthy that EVA3D[9] was primarily trained on female models, which posed a significant limitation for our research.

Upon testing EVA3D[9] with male models, we encountered disappointing results, indicating a significant performance gap compared to its effectiveness with female models. Below is the pipeline and outputs generated by EVA3D[9] when applied to female models, highlighting its capabilities within its intended scope.



Our Findings

After conducting a thorough evaluation of state-of-the-art models in the field, our findings directed the methodology for generating 3D human models. Our approach prioritizes efficiency, accuracy, and adaptability to real-world scenarios.

Our first observation emphasized the utilization of deforming models, which transform a basic shape into a human-based body pose defined by keypoints. We found that commencing with a neutral human shape before deformation leads to improved and expedited results, as evidenced by models like Smpl.

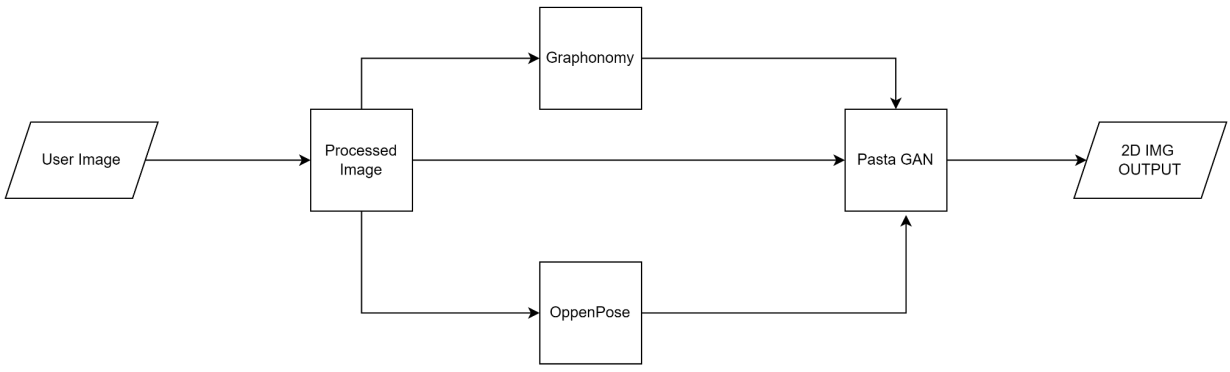
Furthermore, our analysis favored models generating point clouds over those employing vertex-based shapes, such as triangles. Point-based representations provide superior accuracy, enabling finer detail capture and smoother deformations, thus enhancing the fidelity of the 3D model.

Considering the limitations of paired datasets in representing real-world scenarios, our methodology places a strong emphasis on models capable of operating effectively without relying on paired data. We prioritize the development and evaluation of models capable of working with unpaired datasets, ensuring applicability to diverse user scenarios and clothing variations.

By incorporating these insights into our methodology, we aim to develop a robust and versatile approach for generating high-quality 3D human models suitable for virtual try-on applications. Our methodology underscores the importance of efficiency, accuracy, and adaptability in addressing the challenges inherent in virtual try-on systems.

Methodology

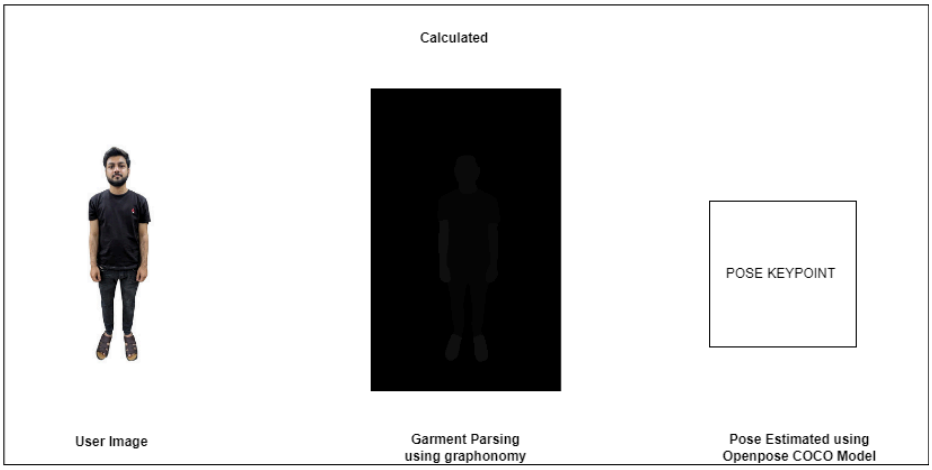
For our study, we employed the Versatile Unpaired Virtual Try-on via Patch-Routed Spatially-Adaptive GAN++[4] model to achieve the task of changing a user's clothing to desired attire. Our methodology involved a thorough examination of the input pipeline of the model to understand its operational framework.



As Shown Above the model utilizes a multi-step process to accomplish its task. Initially, it employs OpenPose[11] to calculate the body pose key points of the user. These key points provide essential information about the user's body posture and orientation.

Subsequently, the model utilizes Graphonomy[13], a state-of-the-art parsing model, to parse the input image and extract detailed information about the clothing worn by the user. Finally, armed with the information obtained from both OpenPose[11] and Graphonomy[13], the model executes the transformation process, seamlessly replacing the user's current clothing with the desired attire. This involves intricate adjustments and adaptations to ensure a natural and realistic appearance of the modified clothing on the user.

By meticulously examining the 2D output pipeline of our model, we gained valuable insights into its functioning and capabilities.

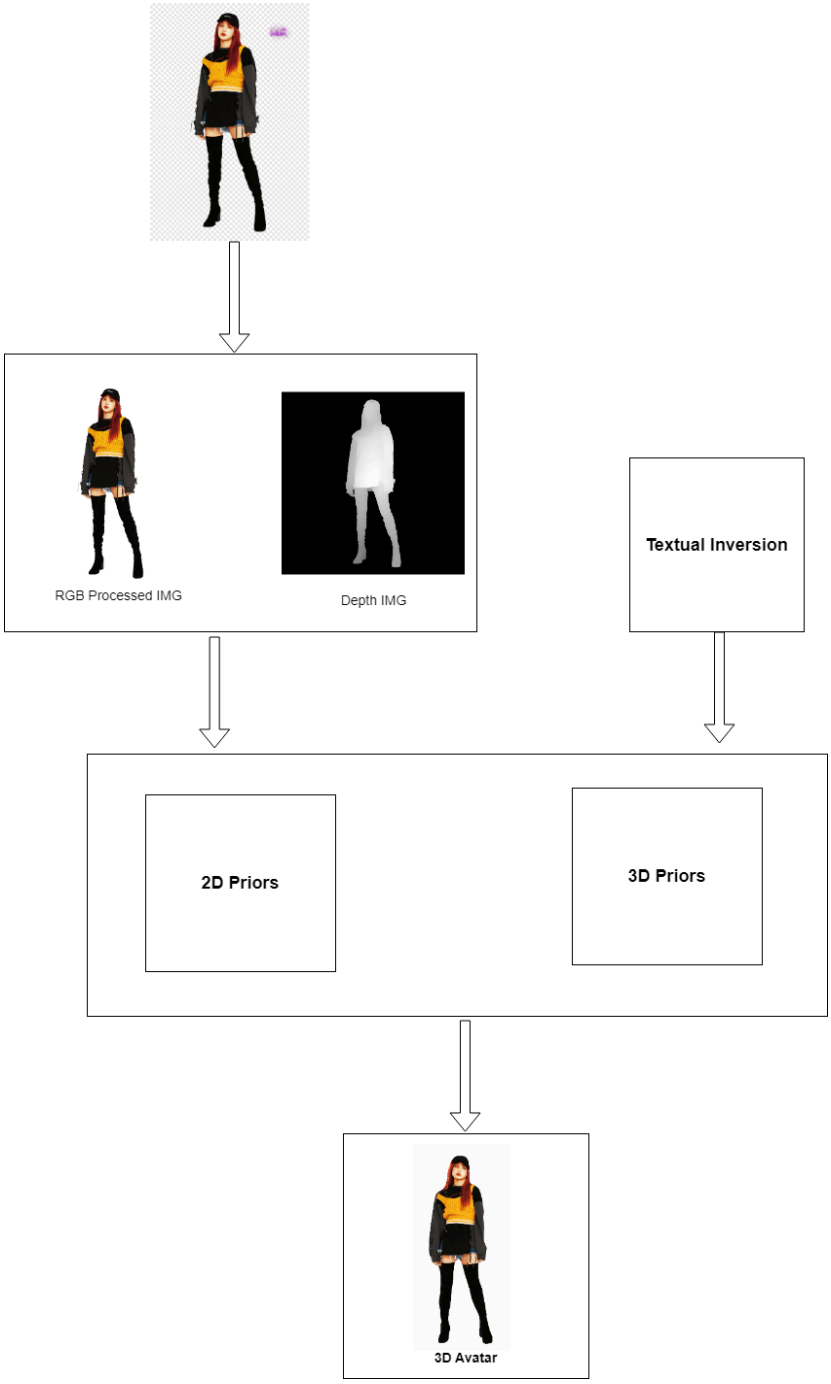


To facilitate the conversion of 2D images to 3D human representations, we selected Magic123[10] as our primary model. Magic123[10] is a state-of-the-art 3D model generation framework that leverages textual inversion alongside both 2D and 3D priors to enhance the quality of its output.

Magic123[10] offers a flexible approach, allowing us to tailor the conversion process based on the available hardware resources. Specifically, we have the option to enable or disable textual inversion depending on hardware constraints. Additionally, we can choose between utilizing only the 2D prior RealFusion or incorporating 3D priors such as Zero 1-to-3, depending on the specific requirements of the task at hand.

The pipeline for utilizing Magic123[10] involves several key steps. Firstly, the model analyzes the input 2D image, extracting relevant features and characteristics. It then employs the selected priors and inversion techniques to generate a corresponding 3D representation of the human subject depicted in the image.

By adopting Magic123[10] as our chosen framework and delineating the specific parameters for its operation, we aim to achieve accurate and high-fidelity conversions from 2D images to 3D human models. This approach allows us to capitalize on the advanced capabilities of Magic123[10].



2D Prior:

Training a complete NeRF model with a single reference image is insufficient without priors. DreamFusion addresses this by using a 2D diffusion model as a prior, guiding novel views through score distillation sampling (SDS) loss. SDS leverages a 2D text-to-image diffusion model: it encodes the rendered view as a latent, adds noise, and predicts the clean novel view guided by an input text prompt. This process translates the rendered view into an image that respects both the content and the prompt. The SDS loss is formulated as:

$$L_{2D} = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(z_t; e, t) - \epsilon)\frac{\partial z}{\partial I}\frac{\partial I}{\partial \theta}\right],$$

where *I* is a rendered view, z_t is the noisy latent obtained by adding random Gaussian noise at time step ***t*** to the latent of ***I.*** Here, ϵ, $\epsilon\phi$, ϕ and θ resent the added noise, predicted noise, parameters of the 2D diffusion prior, and the parameters of the 3D model, respectively. θ can be MLPs of NeRF for the coarse stage or SDF, triangular deformations, and color field for the fine stage. DreamFusion also notes that the Jacobian term of the image encoder $\partial z/\partial I$ in Equation (5) can be eliminated, making the SDS loss more efficient in terms of both speed and memory. In our experiments, we utilize the SDS loss with Stable Diffusion v1.5 as our 2D prior. The rendered images are interpolated to 512×512512 \times 512512×512 as required by the image encoder.

Textural Inversion

The prompt e used for each reference image is not purely textual but involves textual inversion . Using purely textual prompts for image-to-3D generation often results in inconsistent textures and geometries due to the limited expressiveness of human language. For example, using “A high-resolution DSLR image of a specific human” can generate different geometries and colors that do not respect the reference image. Therefore, we follow RealFusion to use a special token <e> representing the object in the reference image. We use the prompt: “A high-resolution DSLR image of <e>”. This method enables Stable Diffusion to generate humans with textures and styles more similar to the reference image compared to results without textual inversion.

3D Prior:

Using only a 2D prior is insufficient to capture detailed and consistent 3D geometry. To address this, Zero-1-to-3 introduces a 3D prior solution by finetuning Stable Diffusion into a view-dependent version using Objaverse, the largest open-source 3D dataset with 818K models. Zero-1-to-3 takes a reference image and a viewpoint as input to generate a novel view from the given viewpoint, serving as a strong 3D prior for 3D reconstruction. The use of Zero-1-to-3 in an image-to-3D generation pipeline with SDS loss is formulated as:

$$L_{3D} = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(z_t; I_r, t, R, T) - \epsilon)\frac{\partial I}{\partial \theta}\right],$$

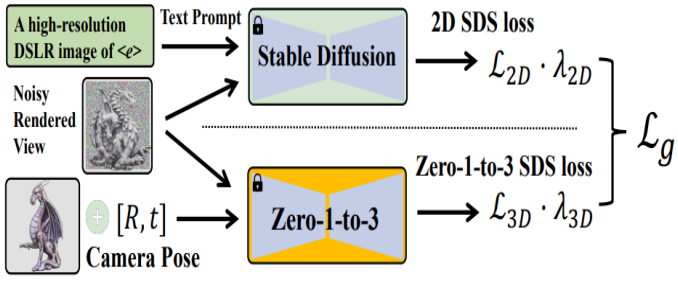
where ***R*** and ***T*** are the camera poses passed to Zero-1-to-3, the view-dependent diffusion model. Unlike the 2D prior, which uses text embedding as guidance, the 3D prior uses the reference view IrI_rIr with novel view camera poses, encouraging 3D consistency and utilizing more 3D information.

The utilization of 3D priors demonstrates a commendable capacity for harnessing geometric details, resulting in significantly more accurate geometric representations compared to 2D priors. This precision is especially effective for objects commonly found within the pre-trained 3D dataset. However, the generalization capability of 3D priors is lower than that of 2D priors, potentially producing geometric structures that may appear implausible due to the limited scale of high-quality 3D datasets. For instance, in the case of uncommon objects, Zero-1-to-3 may yield overly simplified geometries, such as flat surfaces lacking detail in the back view.

Join loss for the 2d and 3d prior is:

$$\mathcal{L}_g = \mathbb{E}_{t_1,t_2,\epsilon_1,\epsilon_2}\left[w(t)\left[\lambda_{2D/3D}(\epsilon_{\phi_{2D}}(\mathbf{z}_{t_1};\mathbf{e},t_1)-\epsilon_1)+\lambda_{3D}(\epsilon_{\phi_{3D}}(\mathbf{z}_{t_2};\mathbf{I}^r,t_2,R,T)-\epsilon_2)\right]\frac{\partial \mathbf{I}}{\partial \theta}\right], \tag{7}$$

Here is a graphical representation of linking of 2d and 3d priors.



Conclusions

In this research endeavor, we embarked on a comprehensive exploration of virtual try-on systems, delving into both 2D and 3D modeling approaches to address the challenges of clothing visualization and customization in online shopping experiences.

Our investigation commenced with an analysis of state-of-the-art models in the field, revealing the diverse methodologies and technologies employed in virtual try-on systems. Through meticulous evaluation, we uncovered the strengths and limitations of various models, informing our subsequent methodological choices.

In our pursuit of realistic and versatile virtual try-on solutions, we formulated a methodology that integrates cutting-edge techniques and frameworks. Leveraging models such as Versatile Unpaired Virtual Try-on via Patch-Routed Spatially-Adaptive GAN++ and Magic123[10], we devised pipelines that harness the power of pose estimation, parsing, and textual inversion to achieve accurate clothing transformation and 3D human representation from 2D images.

Throughout our exploration, we encountered challenges such as hardware constraints, dataset realism, and model adaptability. However, by strategically navigating these obstacles and making informed decisions, we were able to develop methodologies that optimize model performance and accommodate practical considerations.

In conclusion, our research underscores the significance of continual innovation and interdisciplinary collaboration in advancing virtual try-on technology. By integrating insights from computer vision, machine learning, and graphics rendering, we contribute to the ongoing evolution of virtual try-on systems, paving the way for enhanced user experiences and increased engagement in online shopping environments.

Acknowledgments

This work was done during final year project of Punjab University College of Information Technology Under Supervision of Prof. Muhammad Farooq.

References

1. Sang-Heon Shim, Jiwoo Chung, Jae-Pil Heo, "Towards Squeezing-Averse Virtual Try-On via Sequential Deformation," arXiv preprint arXiv:2312.15861 (2023).
2. Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, Wen-Huang Cheng, "Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 7513-7522.
3. Aymen Mir, Thiemo Alldieck, Gerard Pons-Moll, "Learning to Transfer Texture from Clothing Images to 3D Humans," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
4. Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xiaodan Liang, "Towards Scalable Unpaired Virtual Try-On via Patch-Routed Spatially-Adaptive GAN," submitted on 20 Nov 2021.
5. Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, Xiaoguang Han, "Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model Using Pixel-Aligned Reconstruction Priors," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 9287-9297.
6. Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, Michael J. Black, "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975-10985.
7. Shunsuke Saito, Tomas Simon, Jason Saragih, Hanbyul Joo, "PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

8. Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, Xiaodan Liang, "M3D-VTON: A Monocular-to-3D Virtual Try-On Network," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 13239-13249.
9. angzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, Ziwei Liu, "EVA3D: Compositional 3D Human Generation from 2D Image Collections," in Proceedings of the International Conference on Learning Representations, 2023.
10. Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, Bernard Ghanem, "Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors," in Proceedings of The Twelfth International Conference on Learning Representations (ICLR), 2024.
11. Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," submitted on 18 Dec 2018
12. Bowen Wu, Fuwei Zhao, "2D-Human-Parsing," GitHub repository, 2021. [Online]. Available: [https://github.com/fyviezhao/2D-Human-Parsing]
13. Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, Liang Lin, "Graphonomy: Universal Human Parsing via Graph Transfer Learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.