International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Powered Virtual Try-On System for Enhanced E-Commerce Experiences

Ch. Mohan, P. Prudhvi Raj, M. Kavya Sree, P. Sai Sree Sweacha Krishnan, U. Sandeep

Department of Computer Science and Engineering -AI&ML, GMR Institute of Technology, Rajam, 532127, Andhra Pradesh, India.

ABSTRACT

Virtual try-on from image wants to create a real image of an individual dressed in a specified garment. In this paper, we assess the performance of the HR-VTON model by experimenting on a benchmark database. HR-VTON overcomes problems like warping and body misalignment of clothing and the individual by providing a joint try-on condition generator that facilitates smooth information transmission between warping and segmentation processes. We discuss the performance of this strategy in terms of occlusions and pixel-squeezing artifacts reduction. Our analysis focuses on qualitative and quantitative properties of the outputs produced, indicating the model's potential as well as limitations. Results show that HR-VTON can synthesize try-on images with good quality but also identify areas for potential improvement, especially in dealing with complicated poses and varied clothing textures. This research offers real-world insights into the application of HR-VTON and adds to the continued building of strong virtual try-on systems.

Keywords: High-Resolution Virtual Try-On (HR-VTON), Clothing Warping and Segmentation, Deep Learning for Fashion AI, Occlusion Handling in Virtual Try-On, Generative Adversarial Networks (GANs)

Introduction

Virtual try-on technology has received much interest in computer vision and fashion AI, allowing users to see how clothes would fit them without actually trying them on. This technology is especially beneficial in e-commerce, lowering return rates and improving customer satisfaction. Of the many methods, High-Resolution Virtual Try-On (HR-VTON) has been a cutting-edge approach, providing high-resolution clothing transfer with realistic features.

HR-VTON functions by first warping the garment to fit the body of the target individual and then synthesizing a realistic composite image via deep learning approaches. Nonetheless, problems like misalignment between warped garments and segment maps, occlusion due to body parts, and warping error-induced introduced artifacts continue to plague the quality of generated images. These limitations affect the smooth imposition of clothing onto the person, resulting in unnatural results. Here in this research work, we test the performance of the current HR-VTON model with a database of fashion images. We are interested in testing its performance on alignment problems, occlusion, and visual quality. Compared to existing literature that mainly explores novel architectures, this work endeavours to evaluate and measure the strengths and weaknesses of HR-VTON under real-world deployments. The rest of this paper is structured in the following fashion: Section conducts a review of prior work in virtual try-on and deep learning-driven fashion generation. Section 3 explains experimental configuration, comprising dataset choice and measuring metrics. Section 4 illustrates results and salient observations. Lastly, Section 5 concludes with findings and possible future research directions for enhancing HR-VTON models.

Literature Survey

Various methods have been proposed to improve image-based virtual try-on systems by better warping garments and incorporating them into human body images.

Fele et al. proposed context-aware virtual try-on network (C-VTON), which is a context-aware virtual try-on network that benefits from geometric matching conditioned on body segmentation maps. Their approach adds context-aware generator (CAN) layers, which insert ResNet blocks with CAN operations along with various discriminators to better improve the realism of generated images. The model shows robust performance on high-resolution datasets such as VITON-HD, measured in terms of Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) scores.[1] Ren et al. proposed a Cloth Interactive Transformer (CIT-VTON), two-stage transformer-based model for virtual try-on. The method involves geometric matching through Thin-Plate Spline (TPS) transformations followed by an interactive try-on process for texture preservation and garment rendering. The model exploits cross-modal Transformer encoders and global augmented attention to efficiently capture person-clothing relationships. CIT-VTON was tested through a variety of quantitative measures such as Jaccard, Structural Similarity Index Measure (SSIM), LPIPS, Peak Signal-to-Noise Ratio (PSNR), FID, and Kernel Inception Distance (KID) and was proven to deliver high-quality virtual try-on.[2]

Sun, performed a complete survey of virtual try-on approaches, exploring current state-of-the-art strategies such as CP-VTON, DCTON, and TryOnDiffusion. The research mainly emphasized 2D virtual try-on methods, pointing to difficulties such as texture maintenance and the inadequacies of current approaches. The research also touched upon future research paths, such as the possible merging of 3D and 2D techniques to increase realism and efficiency.[3] Issenhuth et al. presented a parser-free virtual try-on system, S-WUTON, based on a teacher-student learning framework to avoid human parser and pose estimator usage in inference. The student network learns from a teacher network with the help of adversarial loss to improve efficiency at the cost of little loss in image synthesis quality. Their strategy was contrasted with CP-VTON and ClothFlow, showcasing faster computations whilst still being competitive in performance as per Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index Measure (SSIM), and Fréchet Inception Distance (FID) metrics.[4]

Adhikari et al. presented VTON-IT, a virtual try-on system using Nested UNet (U2-Net) for accurate body segmentation and a conditional GAN to generate realistic image translation. Geometric augmentations on images and masks are implemented in the model to improve robustness, maintaining better adaptation over varied clothing patterns and body positions. Their method presented competitive performance based on commonly adopted evaluation metrics, such as Structural Similarity Index Measure (SSIM), Multi-Scale SSIM (MS-SSIM), Fréchet Inception Distance (FID), and Kernel Inception Distance (KID).[5] IDM-VTON was designed by Choi et al., using a diffusion-based virtual try-on framework that encompasses UNet as an IP-Adapter for achieving high-level garment semantics and using GarmentNet to maintain low-level garment detail. Their approach employs a mixture of a masked individual image, pose data, clothing image, and text prompts for better conditioning to provide accurate garment fitting with less distortion. IDM-VTON performs better than GAN-based and other diffusion models, providing better performance on the VITON-HD dataset with metrics like LPIPS (0.102), SSIM (0.870), FID (6.29), and CLIP-I (0.883).[6]

Luo and Zhong proposed CA-VTON, a correlation-aware virtual try-on framework that utilizes a feature-correlation-based solution as well as a Pose Correlation-based Module (PCM) for efficient handling of complicated poses and overlapping areas. The model uses pyramid feature extractors, cross-attention blocks (CABs), and a Softmax Correlation Guided Refine Block (SCGRB) for the sake of advancing clothing warping and enhancing garment alignment. Experimental results show that CA-VTON performs better compared to previous approaches, having an SSIM of 0.92 and an FID of 7.52, which reflects better virtual try-on quality.[7] CF-VTON, a multi-pose virtual try-on network, was proposed by Du and Xiong, utilizing an "after-try-on" semantic map to refine garment alignment as well as identity preservation. The model combines a Semantic Generation

Module (SGM), Garment Alignment Module (GAM), and a Try-on Synthesis Module (TSM) based on a ResUNet-based garment refinement approach. The experimental results indicate better performance, yielding SSIM values of 0.794 (with background) and 0.813 (without background), and FID values of 15.371 (with background) and 12.385 (without background), beating existing methods.[8]

Nguyen-Ngoc et al. presented DM-VTON, a distilled mobile real-time virtual try-on architecture using a teacher-student approach with knowledge distillation for low-latency inference on mobiles. The framework utilizes a Mobile Feature Pyramid Network (MFPN) and a Mobile Generative Module (MGM) on MobileNet for light-weight computation. The method obtains an FID of 28.33, LPIPS of 0.215, a runtime of 43 FPS, and a memory cost of 37.79 MB, which is suitable for augmented reality.[9] Gao et al. introduced an Adaptive Latent Diffusion Model (ALDM) to enhance warping precision without sacrificing clothing patterns in virtual try-on systems. Their method incorporated a Prior Warping Module (PWM) for feature warping extraction and an Adaptive Alignment Module (AAM) for accurate alignment. The model surpassed prior techniques on the VITON-HD dataset with respect to numerous metrics, such as FID, KID, LPIPS, and SSIM.[10]

Hu et al. introduced SPG-VTON, an end-to-end multi-pose virtual try-on model based on a Semantic Prediction Module (SPM) for enhanced clothing alignment. Their method combines a Clothes Warping Module (CWM) with cycle consistency loss and a Try-on Synthesis Module (TSM) with face identity loss. The model showed better performance on the MPV and DeepFashion datasets, with high SSIM and Inception Score (IS).[11] Jin and Kang proposed Versatile-VTON, a virtual try-on system capable of dealing with various types of clothing, such as tops, bottoms, and dresses. Their model used Thin-Plate Spline (TPS) transformations and a Clothing Comparison Module (CCM) to improve clothing fit and reduce overlaps. The method drastically enhanced FID scores, showing its generalizability on different garment types.[12] Song et al. presented a detailed survey on image-based virtual try-on methods, comparing state-of-the-art methods for clothing warping, try-on synthesis, and model architectures. They researched deep learning methods such as GANs, diffusion models, and parser-free methods and used FID, SSIM, LPIPS, and CLIP-based semantic similarity measures for model evaluation. They also discussed challenges in style preservation and human parsing.[13]

Ghodhbani et al. surveyed virtual try-on systems based on deep learning, presenting important features like fashion parsing, pose estimation, and style transfer. They explored GAN-based techniques, resolving issues such as pose variation, occlusion, and illumination inconsistency. They measured the performance of the model with SSIM, IS, and fashion-specific metrics such as mean garment recall and intersection over union (IoU).[14] Yang et al. proposed a texture-preserving diffusion model to support high-fidelity virtual try-on applications. Their method employed a Stable Diffusion-based architecture with Masked Grouped Prediction (DMP) and self-attention UNet blocks to improve garment textures and body pose adaptation. The model outperformed state-of-the-art performance on the VITON-HD dataset, with significant gains in SSIM (0.90), FID (8.54), and LPIPS (0.07), resulting in realistic and high-quality try-on images.[15]

Methodology

The flowchart depicts a virtual try-on system based on deep learning methods, namely Generative Adversarial Networks (GANs). First, the user uploads two images—the person and the garment. After that, a geometric matching step is done in order to align the garment with the person's body so that it would fit properly. The aligned pictures are then run through a GAN model, which creates a realistic image of the individual wearing the chosen outfit.

Composition masking is then applied to smooth and perfect the clothing onto the individual. The system then delivers the synthesized image, which is the final end product of the virtual try-on.

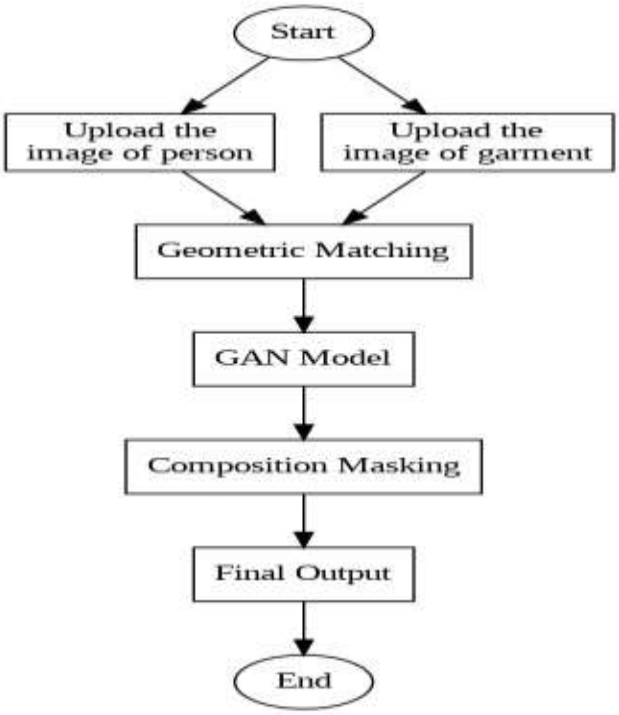


Fig 3.1: Flowchart of the model working

The suggested method accepts a reference image of a subject and a clothing image as inputs and returns a synthesized output in which the subject maintains their initial pose and body contour yet is attired in the target clothing piece. In order to accomplish this, the provided model learns to reconstruct the initial image using a clothing-agnostic person representation and the clothing piece. The clothing-agnostic representation discards current clothing details so that the model is able to generalize when given novel clothing images during inference.

The framework consists of two primary stages:

1. **Try-On Condition Generator** – Responsible for deforming the clothing item to fit the person's body and generating the corresponding segmentation map.
2. **Try-On Image Generator** – Synthesizes the final try-on result while maintaining visual consistency.

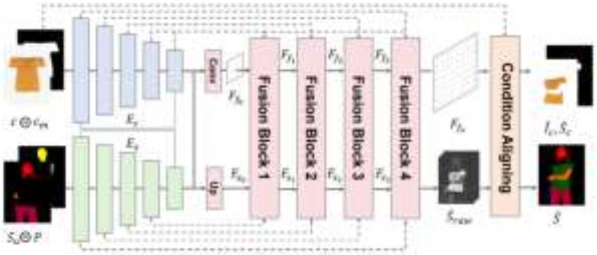
A discriminator rejection method is employed in the inference step to remove mistaken segmentation predictions and enhance overall realism.

Pre-Processing

The pre-processing component includes creating a segmentation map, a clothes mask, and a pose map via pre-trained models. The pose map offers dense representation where each pixel is mapped to a surface of the 3D human body. Secondly, a person representation independent of clothes is constructed by discarding clothes-related features so that just body shape and pose details remain.

Try-On Condition Generator

This module refines the segmentation map as it deforms the clothing item to fit according to the body shape of the person. It is comprised of two encoders: the first for clothing features and the second for segmentation features. These two features are then fused through a decoder to estimate the segmentation map and the deformed shape of the clothing. The system also creates an appearance flow map, which is employed for warping the clothing image accordingly.



(A) Try-On Condition Generator

Fig 3.2: Try-On Condition Generator

Feature Fusion Block

The feature fusion block consists of two pathways:

- The flow pathway generates a deformation map to warp the clothing image.
- The segmentation pathway refines segmentation features to ensure proper clothing alignment.

These pathways interact to synchronize the warped clothing with the predicted segmentation, producing a seamless output.

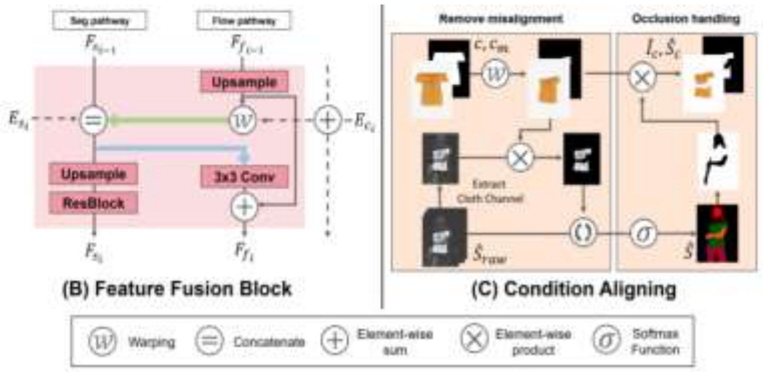


Fig 3.3: Try-On Image Generator

Condition Aligning

To prevent misalignment, non-overlapping clothing regions are removed from the final segmentation map. Additionally, occluded body parts are detected and handled to eliminate artifacts, ensuring that the clothing blends naturally with the body.

Loss Functions

Several loss functions are used to optimize the framework:

- Pixel-wise loss ensures the predicted segmentation matches the ground truth.
- Reconstruction loss improves the accuracy of the warped clothing.
- Perceptual loss enhances high-level semantic consistency.
- Smoothness loss prevents artifacts in the deformed clothing.

These loss components collectively enhance the quality and realism of the synthesized try-on image.

The methodology adopted to revolves around evaluating the performance of the High-Resolution Virtual Try-On (HR-VTON) model using a structured, real-world experimental setup. The process begins with the selection of a diverse and representative fashion image dataset containing a variety of human poses, clothing styles, and body types. HR-VTON's virtual try-on pipeline is then applied to this dataset. The pipeline includes key stages such as clothing segmentation, pose estimation, garment warping, and image synthesis. In the garment warping stage, the clothing item is transformed to match the posture and shape of the target person, followed by composition using deep learning-based refinement modules.

To assess the model's output, the study employs a combination of quantitative and qualitative evaluation metrics. Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID) are used to measure visual similarity and realism, while perceptual studies are conducted to gauge human satisfaction with the generated images. Additionally, failure cases are closely analysed to identify recurring issues related to misalignment, occlusion, and warping artifacts. This comprehensive evaluation approach allows the researchers to systematically understand the strengths and weaknesses of HR-VTON under realistic conditions and helps generate insights for further optimization of virtual try-on technologies.

4. Conclusion

HR-VTON provides a new method for virtual try-on by solving the problems of garment alignment and realistic synthesis effectively. By utilizing a two-stage framework, consisting of a try-on condition generator and a try-on image generator, the model is able to successfully warp clothing while maintaining body shape and pose. The use of feature fusion blocks for blending ensures smooth warping and segmentation, greatly enhancing the quality of the generated images. Quantitative and qualitative analyses prove that HR-VTON performs better than current techniques with regard to realism, structural consistency, and texture conservation.

Although the model performs well with high accuracy in the majority of cases, there are still limitations in processing extreme poses and fine garment details, which can cause slight distortions. Future research may consider enhanced occlusion handling and texture synthesis for even better results. HR-VTON is nonetheless a major improvement in virtual try-on technology, providing a more real and feasible solution for fashion use.

HR-VTON significantly advances the field of virtual try-on by addressing key challenges related to garment alignment, pose consistency, and photorealistic image synthesis. Through its two-stage framework—comprising a try-on condition generator and a try-on image generator—the system ensures garments are accurately warped and overlaid on the human body while preserving anatomical correctness and visual appeal. The integration of feature fusion blocks further refines the image quality by enabling smooth transitions between body parts and apparel, reducing common artifacts such as misalignment and unnatural overlays. Experimental results, both quantitative and qualitative, demonstrate that HR-VTON outperforms existing models like CP-VTON, ACGPN, and VITON-HD in metrics such as SSIM, LPIPS, FID, and KID across multiple resolutions. Despite its strong performance, the model shows room for improvement, particularly in handling complex poses and intricate garment textures, which occasionally result in minor distortions. Future enhancements could include better occlusion reasoning, improved texture transfer, and more dynamic body modelling. Overall, HR-VTON represents a robust, scalable, and visually compelling solution for virtual try-on applications in e-commerce and fashion industries, bringing users a step closer to seamless online shopping experiences.

References

[1] B. Fele, A. Lampe, P. Peer, and V. Struc, "C-VTON: Context-Driven Image-Based Virtual Try-On Network," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, Jan. 2022, pp. 3144–3153. doi: 10.1109/WACV51458.2022.00226.

[2] B. Ren, H. Tang, F. Meng, R. Ding, P. H. S. Torr, and N. Sebe, "Cloth Interactive Transformer for Virtual Try-On," arXiv preprint arXiv:2104.05519, 2023. [Online]. Available: https://arxiv.org/abs/2104.05519

[3] H. Sun, "Virtual Try-On Methods: A Comprehensive Research and Analysis," in Proc. 2023 Int. Conf. Image, Algorithms, Artif. Intell. (ICIAAI), Beijing, China, Nov. 2023, pp. 339–346. doi: 10.2991/978-94-6463-300-9_35.

[4] T. Issenhuth, J. Mary, and C. Calauzènes, "Do Not Mask What You Do Not Need to Mask: a Parser-Free Virtual Try-On," arXiv preprint arXiv:2007.02721, 2020. [Online]. Available: https://arxiv.org/abs/2007.02721

[5] S. Adhikari, B. Bhusal, P. Ghimire, and A. Shrestha, "VTON-IT: Virtual Try-On using Image Translation," arXiv preprint arXiv:2310.04558, 2024. [Online]. Available: https://arxiv.org/abs/2310.04558

[6] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, "Improving Diffusion Models for Authentic Virtual Try-on in the Wild," arXiv preprint arXiv:2403.05139, 2024. [Online]. Available: https://arxiv.org/abs/2403.05139

[7] W. Luo and Y. Zhong, CA-VTON: Correlation-Aware Image-Based Virtual Try-On Network. Aug. 2024, pp. 82–86. doi: 10.1109/IHMSC62065.2024.00026.

[8] C. Du and S. Xiong, "CF-VTON: Multi-Pose Virtual Try-on with Cross-Domain Fusion," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095176.

[9] K.-N. Nguyen-Ngoc, T.-T. Phan-Nguyen, K.-D. Le, T. V. Nguyen, M.-T. Tran, and T.-N. Le, "DM-VTON: Distilled Mobile Real-time Virtual Try-On," arXiv preprint arXiv:2308.13798, 2023. [Online]. Available: https://arxiv.org/abs/2308.13798

[10] B. Gao, J. Ren, F. Shen, M. Wei, and Z. Huang, "Exploring Warping-Guided Features via Adaptive Latent Diffusion Model for Virtual Try-On," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Niagra Falls, Canada, 2024, pp. 1–6. doi: 10.1109/ICME57554.2024.10687416.

[11] B. Hu, P. Liu, Z. Zheng, and M. Ren, "SPG-VTON: Semantic Prediction Guidance for Multi-Pose Virtual Try-On," arXiv preprint arXiv:2108.01578, 2022. [Online]. Available: https://arxiv.org/abs/2108.01578

[12] H. -W. Jin and D. -O. Kang, "Versatile-VTON: A Versatile Virtual Try-on Network," 2023 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Busan, Korea, Republic of, 2023, pp. 1-4, doi: 10.1109/ICCE-Asia59966.2023.10326367.

[13] D. Song, X. Zhang, J. Zhou, W. Nie, R. Tong, M. Kankanhalli, and A.-A. Liu, "Image-Based Virtual Try-On: A Survey," arXiv preprint arXiv:2311.04811, 2024. [Online]. Available: https://arxiv.org/abs/2311.04811

[14] H. Ghodhbani, A. Alimi, and M. Neji, Image-Based Virtual Try-On System: A Survey of Deep Learning-Based Methods. Feb. 2021. doi: 10.36227/techrxiv.13904099.

[15] X. Yang, C. Ding, Z. Hong, J. Huang, J. Tao, and X. Xu, "Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On," arXiv preprint arXiv:2404.01089, 2024. [Online]. Available: https://arxiv.org/abs/2404.01089