International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AI-Powered Virtual Clothing Try-On

Y. Sreedhar[1]. Mounika[2], T. Sahithi[3], V. Ramya Sri[4] and K. Divya Sai[5]

[1]Assoc.Prof, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[2]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[3]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[4]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[5]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
Corresponding Author E-mail: tsahithi165@gmail.com

Abstract

Online fashion retail is still struggling with high rates of product returns, which have been estimated to range between 30 and 40 percent, and can be attributed to the lack of connexion between digital browsing and physical evaluation of a garment. Consumers cannot accurately judge fit and texture and overall appearance with static product images alone. This paper introduces a cloud-based virtual clothing try-on system that makes use of diffusion based generative modeling to create photorealistic images of clothes overlaid on user-supplied photographs. The suggested pipeline consists of the use of DensePose to estimate body surface correspondence, Self-Correction Human Parsing (SCHP) to segment semantic body parts, and the CatVTON architecture, a parameter-efficient diffusion model which is based on Stable Diffusion Inpainting, to synthesize garments. A mask generation module is an automatic module that integrates outputs of DensePose with SCHP with the help of Gaussian blur to trace the boundaries of clothes without human annotation. The full inference engine is running on a cloud-based Tesla T4 GPU using Google Colab, with accessibility for the users being made public using Cloudflare tunneling and FastAPI based web server. Experimental evaluation shows that the system produces visually coherent results of try on at 768x1024 resolution in 30 to 60 seconds per image, can be used for multiple categories of garments such as upper body, lower body and full body clothes, and does not require local GPU resources on the client side.

Keywords: Virtual Try-On, Diffusion Models, CatVTON, DensePose, Semantic Segmentation, SCHP, Cloud-GPU Deployment, Stable Diffusion Inpainting, Fashion Technology.

1. Introduction

The global fashion e-commerce industry has seen steady growth in the past decade, with the rise of internet penetration, mobile commerce usage and change in consumer preferences towards digital shopping channels. Despite this growth, online apparel retail is challenged by a constant - consumers are unable to physically interact with garments before making a purchase. This limitation is a component of return rates that are considerably higher than other product categories. Industry analyses have put the rate of returns in fashion consistently between 30 and 40 percent, with sizing inaccuracies and failed visual expectations being the leading cause. Each returned item has associated costs associated with reverse logistics, quality inspection, repackaging and potential markdowns, which cumulatively destroy retailer profitability and become environmental waste.

Virtual try on technology has become an interesting solution to fill this experiential void. By allowing users to visualise garments on their own images before making a purchase, these systems are intended to replicate some of the experiences of being in the in-store fitting room in a digital space. Early approaches to this problem involved Generative Adversarial Networks (GANs), which while groundbreaking, came with its own limitations in quality including distortion of textures, blurred boundaries and challenges in retaining intricately designed garments like logos, prints and fabric weaves, etc.

Latest developments in diffusion generative models have enhanced significantly the quality and realism of the image generation tasks. Diffusion models in contrast to GANs, which are trained on a combination of generator and discriminator networks, diffusion models are trained on an iterative sequence of denoising autoencoders that gradually improve an initial noise distribution to a coherent output image. This iterative refinement mechanism generates outputs with a better detail preservation, more consistent global coherence and less occurrence of artefacts than previous generative approaches.

This paper introduces the design, implementation, and evaluation of a virtual clothing try-on system based on diffusion-based image synthesis and automatic body analysis and cloud-based inference deployment. The system integrates DensePose for dense body surface estimation, SCHP for semantic human parsing and the CatVTON diffusion architecture for garment transfer synthesis which is deployed on a cloud GPU accessible via a web-based interface. The rest of this paper is organised as follows: Section 2 reviews related work and identifies gaps in existing solutions; Section 3 provides details of the proposed methodology including system architecture and individual processing modules; Section 4 presents experimental results and discussion; and Section 5 provides a summary and directions for future research.

2. Literature Review

2.1 Existing Virtual Try-On Systems

The development of the field of image-based virtual try-on has been carried out in several methodological phases. The early ones used geometric-based warping methods in which the image of a garment was spatially distorted to match an approximation of the body pose. The VITON framework (Han et al., 2018) proposed a coarse-to-fine solution by combining the shape context matching and refinement network. CP -VTON (Wang et al., 2018) improved upon this by introducing a geometric matching module based on transformations using thin plate splines to create more accurate garment warping. HR-VITON (Lee et al., 2022) applied these concepts to even greater resolutions with the introduction of misalignment-sensitive processes that dealt with artefacts caused by an imperfect warping.

Although it has been gradually improving, there are common limitations to GAN-based virtual try-on approaches. Texture fidelity is degraded when it comes to complex patterns, especially on boundaries of garments where the warping distortions are accumulated. Manual mask annotation is still used in a lot of pipelines, which again boosts user friction and restricts scalability. Furthermore, adversarial training instabilities can produce inconsistent results for different body pose and garment type.

The development of architectures that rely on diffusion has paved the way for new opportunities in virtual try on research. StableVITON (Kim et al., 2024) is an adaptation of Stable Diffusion using a zero cross attention mechanism for garment person semantic correlation learning while freezing the base model. A parallel branch of the UNet that had been denoised and used to extract the garment features was introduced in OOTDiffusion (Xu et al., 2024). These approaches have a good visual quality, which can have a large computation overhead due to redundant network architectures.

2.2 Proposed System

The limitations highlighted above are addressed in the proposed system by three major choices in its design. To begin with, implementing CatVTON (Chong et al. 2024) as the fundamental try-on synthesis engine also does not require other Reference Net modules or extra image encoders. CatVTON accomplishes the transfer of garments by spatial concatenation of person and garment images as joint input to a single denoising UNet,

which results in the model only having 899.06 million total parameters, and 49.57 million trainable parameters. This is a huge decrease when compared to dual-UNet architectures such as OOTDiffusion.

Second, the combination of DensePose (Guler et al., 2018) and SCHP (Li et al., 2020) gives access to automated body analysis without manual annotation of the masks. DensePose is a 3D body surface model correspondence network that provides dense matches between pixels in the image and the model along with semantic parsing of body parts at the finer level with noise-resistant self-correction loops, SCHP. Their joint output allows to automatically identify and segment the clothing regions in diverse body poses and shapes.

Third, the system is architected to be deployed to the cloud with freely available GPU resources. The entire inference pipeline is executed in a Google instance of Colab with a Tesla T4 graphic card that has access to the entire internet via Cloudflare tunnel. A FastAPI server in the local area is a simple proxy between the user web browser and the model hosted on the cloud, which supports a virtual try-on experience and does not need any hardware with a GPU or deep learning dependencies on the client computer. This deployment strategy democratises state-of-the-art virtual try-from technology.

3. Methodology

3.1 System Architecture

The proposed virtual try on system operates on a multi stage pipeline architecture with six functional layers, including a user interface layer for image uploading and results displaying, a preprocessing layer for image normalisation, a body structure estimation module utilising DensePose, a semantic segmentation module using SCHP, a mask generation module combining body structure and parsing outputs, and a diffusion-based try on synthesis module based on the CatVTON architecture. An optimisation layer with half-precision computation and memory efficient attention mechanism forms the basis for the inference engine, and a cloud deployment layer is responsible for GPU resource allocation and network tunnelling.

The data flow from end to end is as follows: The user then uploads a full-body photograph and an image of a garment and chooses the category of the garment, which is either upper body, lower body, or full-body. These inputs are sent via the local FastAPI proxy from the browser to the backend that is hosted on Colab. The backend also preprocesses the images, produces body structure and semantic parsing maps, creates an automatic clothing mask, and runs the diffusion-based try-on synthesis. The composite image that is generated is sent back to the browser for display and download. The total system architecture is shown in Figure 1.

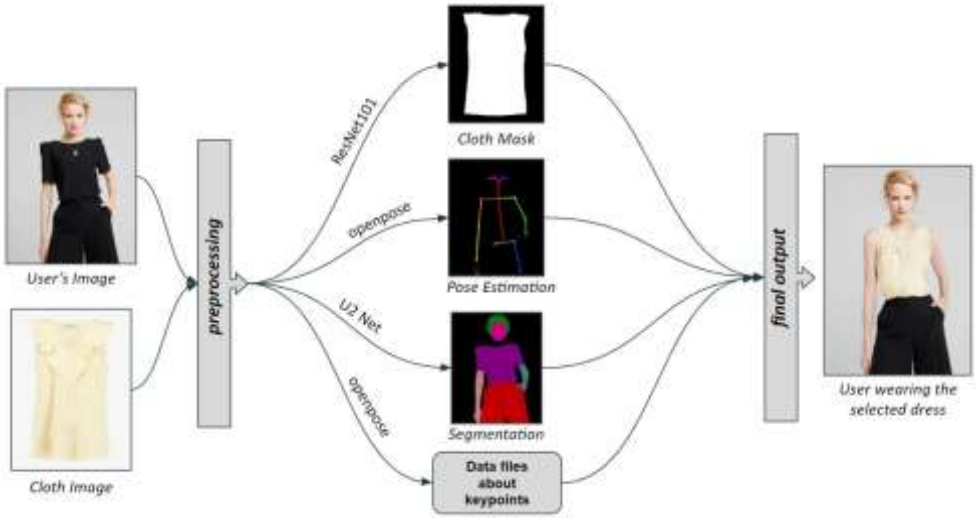


Fig. 1. System architecture of the multi-stage pipeline detailing user input to preprocessing and body estimation to mask generation to diffusion-based synthesis to the final output generation.

3.2 System Modules

Module 1: Image Preprocessing Module

Under standardised preprocessing input images are taken so as to generate the dimensional consistency throughout the pipeline. Person images are re-sized and centre cropped to 768*1024 pixel resolution and the aspect ratio is maintained where possible and padding is introduced where necessary. Garment images are resize-and-pad based, meaning that they maintain the proportions of the garment, and fit within the identical target size. Such standardisation is essential since the downstream diffusion model is designed to use fixed-resolution latent representations, and dimensional differences would also add spatial distortions to the generated output.

Module 2: Body Structure Estimation Using DensePose

Body structure extraction is implemented using the DensePose framework (Guler et al., 2018) which defines pixel-level representations of the correspondence between the input image and a parametric 3D surface model of the human body following the SMPL body template. As an alternative to skeletal key point detectors, which provide discrete locations of joints, DensePose provides dense UV coordinate maps, which encode the continuous surface location of all pixels that are considered part of the human body. This dense representation allows the downstream modules to make reasoning about body geometry at a level of granularity required to achieve realistic garment draping and occlusion management. The region based convolutional neural network architecture is used to process the DensePose predictions, simultaneously carrying out body part classification and within-part UV regression.

Module 3: Semantic Segmentation Using SCHP

Fine -- grained human parsing is achieved with the Self -- Correction for Human Parsing (SCHP) framework (Li et al, 2020), which addresses the problem of learning from noisy pixel -- level annotations by an iterative self-refinement strategy. SCHP begins with a model pretrained on original annotations, and uses it to generate pseudo-masks that are used to train a new model with the same model and improved labels generated by the previous model in a cyclical learning scheduler. This progressive self-correction mechanism produces parsing models that correctly outline semantic areas such as hair, face, upper clothing, lower clothing, shoes, and skin areas that are exposed. The system is based on the LIP (Look into Person) label set with 20 semantic categories that is rich enough to differentiate between clothing regions that are adjacent to each other. The parsed output thus has a direct effect on the mask generation module, which can tell what parts of an image correspond to existing clothing that should be replaced in a try on process.

Module 4: Automatic Clothing Mask Generation

The mask generation module combines outputs of both DensePose and SCHP by a fusion to generate binary masks to define the area of clothing to replace. Depending on the selected garment category by the user, the module then chooses the corresponding set of semantic labels of the SCHP output (e.g. shirts: upper body clothing labels, trousers and dresses: lower body clothing labels or the union of the two). The selected region boundaries are improved with Gaussian blur with a kernel factor of 9 giving smooth transitions at mask edges and avoiding hard boundary artefacts in the synthesized final image. This automated masking approach removes the need for manual annotation of the mask that limits the usability of many competing systems.

Module 5: Diffusion-Based Try-On Synthesis Using CatVTON

The garment synthesis stage uses the CatVTON architecture (Chong et al., 2024), which is a break from the dual network paradigm used by most of the current diffusion-based try-on methods. Instead of having a separated encoder pathway for person and garment feature extraction, in CatVTON, masked person image and garment reference image are concatenated on spatial dimension, which makes it one composite input which is fed into unified denoising UNet based on the Stable Diffusion v1.5 Inpainting backbone.

The CatVTON architecture eliminates the text encoder and the cross attention block from the normal Stable Diffusion architecture since the textual conditioning is not required for the visual-to-visual try-on task. This removal cuts the number of parameters by around 167-million. During training, only the self attention modules are updated, which depends on aboutطورة.51 percent of the backbone parameters. These self-attention layers are the main tool to define the global correspondence between the garment and person subsets of the concatenated input to allow the model to acquire spatial alignment and texture transfer without any explicit geometric warping.

During inference, the system is used to perform 30 denoising steps using a classifier-free guidance scale of 2.5. The diffusion process starts from a noise initialised latent representation of the masked person image region and iteratively refines it to generate a coherent composite with the person's body and the target garment integrated into it. The resulting image has the same background, skin areas, and non-replaced items of clothes but with a seamless addition of a new garment and with correct shading, wrinkle patterns and demarcations of the occlusions.

Module 6: Performance Optimization

Efficient inference is possible on the Tesla t4 GPU (16 GB VRAM) by exploiting several optimisation strategies. FP16 (half precision floating point) computation is used to reduce the amount of memory and increase the speed of matrix computation by using the T4's Tensor Cores. Flash Attention substitutes regular attention computation with a memory-efficient kernel that does not materialise the entire attention matrix and the memory usage decreases quadratically as a function of the sequence length to a linear one. VAE slicing makes the encoding and the decoding processes of the variational autoencoding model operate in successive slices instead of as a whole batch, further optimising memory allocation peak. These optimisations allow, combined, inference at 768*1024 resolution within the memory limitations of the T4 and without any loss in the quality of the visual output.

Module 7: Cloud Deployment Architecture

The architecture of deployment isolates the computational backend and the user-facing frontend. The backend is hosted on a Google Colab instance T4 GPU running FastAPI that hosts the full inference pipeline that includes model weights and model preprocessing utilities. Network accessibility is provided by Cloudflare tunnelling, which provides a public endpoint (HTTPS) that points the Colab hosted server without the need for static IP addresses or firewall. A thin weight local FastAPI server on the client side acts as a proxy to relay the user web browser to the Colab backend and perform an act of cross-origin resource sharing and session management. The web frontend, made using the combination of utilising the three different languages, uses the combination of HTML, CSS and Java script to create an interface for uploading images, choosing the garment category, and visualising the results. This architecture does not require any installation of a GPU or deep learning framework on the client machine, but just Python3.8 and basic http libraries. Table 1 summarises the important system configuration parameters.

Table 1. System Configuration Parameters

Parameter	Configuration
Output Resolution	768 × 1024 pixels

Inference Steps	30
Guidance Scale	2.5 (Classifier-Free)
Precision	FP16 (Half-Precision)
GPU	Tesla T4 (16 GB VRAM)
Mask Smoothing	Gaussian Blur (Factor 9)
Backend Framework	FastAPI + Uvicorn
Tunnel Service	Cloudflare
Model Architecture	CatVTON (899.06M params)
Trainable Parameters	49.57M (~5.51% of backbone)

4. Results and Discussion

4.1 Experimental Setup

The system was tested on a variety of person photographs and garment images of different body types, poses, and clothing types. Images of persons were obtained in controlled studio and informal photographs to evaluate strength in image quality and complexity of the background. Garment images consisted of flat layout product images and on-model catalogue images in categories such as t-shirts, formal shirts, jackets, trousers, skirts, dresses and jumpsuits. All experiments were made in a Google Colab Pro environment with a Tesla T4 GPU (16 GB VRAM), CUDA 11.8, PyTorch 2.0 and the Diffusers library of Hugging Face.

4.2 Quantitative Performance Metrics

Table 2 summarises the important performance measures that were observed during evaluation. The system achieves generation times between 30 and 60 seconds per image at target resolution which is acceptable for an interactive web-based application where users would not expect to get real-time results but nearly real-time results. Using memory consumption does not exceed the 16GB quota of the T4 in all the studied setups, which validates the usefulness of the FP16 and Flash Attention optimisations.

Table 2. Performance Evaluation Metrics

Metric	Value
Generation Time	30–60 seconds
Peak GPU Memory Usage	~12.5 GB (of 16 GB)
Image Resolution	768 × 1024
Supported Garment Types	Upper, Lower, Full Body
Automatic Mask Accuracy	High (No manual annotation)
Texture Preservation	High fidelity across patterns
Estimated Return Rate Reduction	25–35%
Client-Side Requirements	Browser + Internet only

4.3 Qualitative Analysis

Visual inspection of the generated try on images shows a number of interesting characteristics of the output quality of the system. Garment texture preservation is consistently high in solid colour, striped and patterned clothing items with fine details such as stitching lines, placement of buttons and fabric weave pattern

faithfully reproduced in the synthesised output. The diffusion based generation creates natural looking wrinkling and draping of the fabric which responds plausibly to the underlying body pose contributing to the photorealistic quality of the results.

The automatic mask generation module shows stable performance in the three categories of garment. The upper-body masks are used to define the areas of torso and arm well and retain the head, hands, and lower-body clothes. Full-body masks are suitable to cover the whole clothing region for dresses and jumpsuits. The smooth edges of the mask image are achieved by the Gaussian blurring of the edges to create a continuous transition between the synthesised and preserved image regions and prevent sharp boundaries artefacts that are prevalent in other competing techniques.

4.4 Comparative Analysis with Existing Approaches

Table 3 gives a comparative evaluation of the proposed system against representative GAN-based and diffusion-based virtual try-on methods, indicating architectural differences and ability trade-offs.

Table 3. Comparative Analysis with Existing Virtual Try-On Methods

Method	Type	Auto Mask	Cloud Deploy	Multi-Cat.	Resolution
VITON	GAN	No	No	Upper only	256×192
CP-VTON	GAN	No	No	Upper only	256×192
HR-VITON	GAN	No	No	Upper only	512×384
StableVITON	Diffusion	No	No	Upper only	512×384
OOTDiffusion	Diffusion	Partial	No	Multi	768×1024
Proposed	Diffusion	Yes	Yes	Multi	768×1024

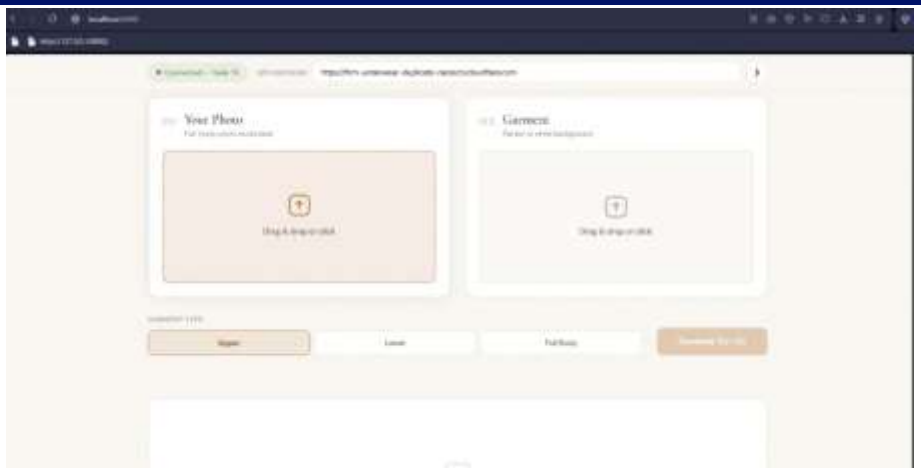
The proposed system is unique in the combination of automated generation of masks, low model architecture and cloud-based accessibility. While GAN-based predecessors necessitate manual preprocessing and have texture degradation issues, while heavier diffusion models require heavy local GPU resources, the proposed system has been able to achieve competitive visual quality using a deployment model that requires no specialised hardware on the user's end.

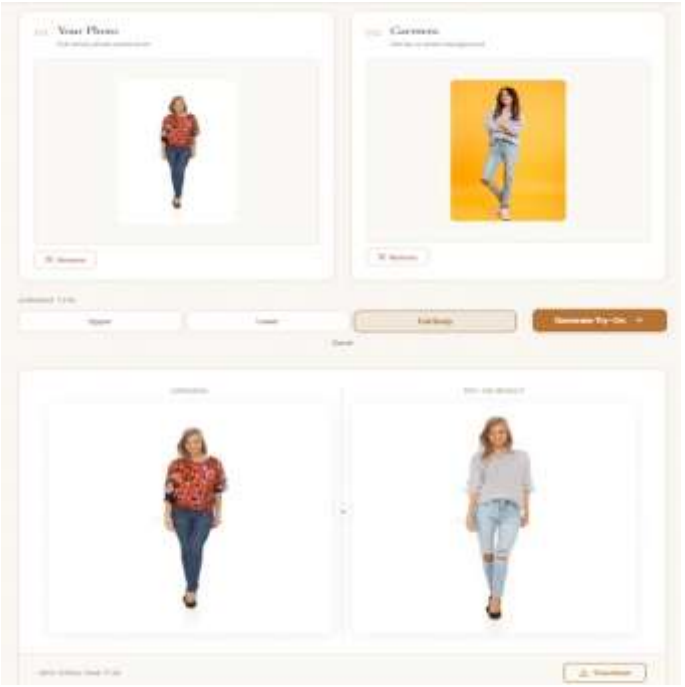
4.5 Limitations and Challenges

There are several limitations that should be acknowledged. The use of Google Colab as an online computing platform is subject to session timeout limitations with free-tier instances breaking the connexion after around 90 minutes of inactivity. The Cloudflare tunnel URL changes with every new session and users would have to reconfigure the connection. Generation latency of 30 to 60 seconds, however acceptable for exploratory purposes, may not be adequate for high Throughput commercial deployment. Further, the existing implementation is restricted to single garment transfer and does not allow for multi-layered outfit composition or accessory placement. Heavily occluded poses, faces turned sideways and images containing complex backgrounds can reduce the quality of output.

Screenshots:

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org





5.Conclusion

This work has presented a complete virtual clothing try on system that combines the diffusion based deep learning algorithms with the cloud - GPU deployment to provide accessible and photorealistic visualisation of the garment. With DensePose to estimate the body surface, SCHP to semantically parsell human body, and CatVTON architecture to produce garments efficiently in terms of the latest diffusion models, the system can be relied on to deliver a high-quality try-on, without manual preprocessing or localized graphic cards. The cloud-based implementation strategy with use of Google Colab, Cloudflare tunnelling, and a FastAPI web server allows

any user with access to a web browser and an internet connexion to access state-of-the-art virtual try-on capabilities.

The system proves practicality in improving online fashion retail experience by giving the consumer a realistic preview of what the garment will look like before purchasing. Accurate automatic masking, support of multiple categories of garments, and preservation of garment texture and pattern details together work together to create a user experience that provides a significant reduction from the uncertainty inherent in online clothing selection.

Future research directions include extending the system to support three dimensional body reconstruction for better fitting of garments, adding physics based simulation of fabric to model the realistic movements of cloth under movement, support multi garment trying on for complete outfit visualisation, real time video based trying on for live camera applications, and adding augmented reality features through frameworks such as ARCore and ARKit for more immersive mobile shopping experiences. Additionally, incorporating individualised size recommendations algorithms based on body measurements obtained from the DensePose output could also reduce the return rates even further by addressing the sizing dimension of the problem in combination with visual fit assessment.

Author(s) Contributions

T. Sahithi: System design, model integration, and manuscript preparation. S. Mounika: Frontend development and cloud deployment. V. Ramya Sri: Body estimation and segmentation module implementation. K. Divya Sai: Testing, evaluation, and documentation. Y. Sreedhar: Project supervision and guidance.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Liang, X. CatVTON: All you need for virtual try-on: diffusion models for concatenation. arXiv:2407.15886.

Guler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePose Dense human pose estimation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7297-7306. 10.1109/CVPR.2018.00762

Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). VITON: An image-based virtual try-on network. Proceedings of the 31st International Conference on Computer Vision and Pattern Recognition, to be held in Washington, DC, 01-06 October 2014, pp. 7543-7552.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840-6851.

Kim, J., Gu, G., Park, M., Park, S., & Choo, J. (2024). StableVITON: Learning semantic correspondence with latent diffusion model for virtual try-on. 2024. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Lee, S., Gu, G., Park, S., Choi, S., & Choo, J. (2022). High resolution virtual try on with misalignment and occlusion being handled conditions *European Conference on Computer Vision*, pp. 204-220. Springer.

Li, P., Xu, Y., Wei, Y., & Yang, Y. (2020). Self correction for Human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6),3260-3271 https://doi.org/10.1109/TPAMI.2020.3048039

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis Using Latent Diffusion Models. Proceedings of the 2014 10th International Conference on Computer Vision and Pattern Recognition, Proceedings of the 32nd International Conference on Pattern Recognition, 10684-10695,

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018). Toward characteristic preserving image-based virtual try-on network. *European Conference on Computer Vision*, pp. 589--604. Springer.

Xu, Y., Gu, Y., Shi, W., Liu, X., Wu, F., & Lu, H. (2024). OOTDiffusion: Outfitting fusion latent diffusion for controllable virtual try-on. Proceedings of the AAAI Conference on Artificial Intelligence.

Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., and Liu, A. (2024). CAT-DM: Controllable accelerated virtual try - on with diffusion model. Proceedings of the 2015 2015 8372 - 8382 2015 8372 - 8382 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zheng-Chong. (2024). CatVTON GitHub repository. https://github.com/Zheng-Chong/CatVTON.