2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE) | 979-8-3315-9709-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCRTEE64519.2025.11053101

Enhancing E-Commerce with Virtual Try-On Using Stable VITON

Mr.R.Raja subramaniam
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
rajasubramanian.r@klu.ac.in

Machireddy Dhamini
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
dhaminimachireddy@gmail.com

Marni Srija
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
srijamarni20@gmail.com

Mallavarapu Vaishnavi
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
vaishnavimallavarapu14@gmail.com

Malepati Vidyadhari
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
vidyadharimalepati12@gmail.com

.

Abstract—**We have seen a huge growth of popularity for online clothes shopping, the absence of a personalized try-on experience continues to affect customer satisfaction and purchase confidence. To address this challenge, we propose a deep learning-based virtual try-on system that enables customers to visualize how a selected outfit would appear on them before making a purchase. Customer can exchange their own image along with a screenshot of a desired clothes collected from popular e-commerce platforms such as Myntra or Flipkart. The system utilizes Stable VITON, a diffusion-based virtual try-on framework built upon Stable Diffusion v1.4, which generates high-quality, realistic try-on images by learning fine-grained garment alignment and human-body context. This model combines a U-Net-based denoising network, variational autoencoder (VAE), and a Zero Cross-Attention Block to precisely synthesize the outfit onto the person's image. To enhance the output, attention enhancing techniques, defect removal modules, and attention total variation loss are applied, resulting in realistic try-on results with minimal artifacts. Unlike conventional recommendation systems, our model provides a direct visual interface for customer interaction, link the gap between online browsing and in-store trial experiences. The proposed system aims to redefine fashion e-commerce by making virtual fitting more realistic, accessible, and customer-centric.**

Keywords- Virtual try-on, Fashion recommendation system, StableVITON, Stable Diffusion, U-Net, Variational Autoencoder (VAE), Zero Cross-Attention Block, Deep learning, Generative AI, Image synthesis, Attention refinement, Garment alignment, Human parsing, Diffusion models, Visual fitting room, Personalized fashion, AI in fashion technology

I. Introduction

The rapid growth of online fashion retail has changed the way consumers browse for clothes and make purchases. However, one critical drawback still remains: the option of virtually trying on clothing before purchase does not exist. While online shopping is convenient and offers many options, it is not tailored or lifelike, resulting in high return rates and lower customer satisfaction.

In this regard, the authors have developed a new virtual try-on system that allows users to see how garments they selected would fit on them in real-time. With the help of personal images, one can simply upload a photograph of themselves along with a screenshot of a clothing item from Myntra or Flipkart, and they will receive a synthesized picture with the garment placed onto their body. Our approach is distinctly different from typical recommender systems that use either user behavior or product relationships since it utilizes direct simulation of visual representation.

The proposed framework harnesses the power of the photorealistic output diffusion-based generative model Stable VITON, built on Stable Diffusion v1.4, which features a U-Net-based denoiser, and a Zero Cross-Attention Block. Additional vision-altering components for garment alignment, attention refining, and defect reduction greatly improve the resulting image quality.
This work contributes a practical, user-centric approach to online shopping, aiming to revolutionize virtual fashion interaction through advanced deep learning techniques.

II.RELATED WORKS

Ensuring consumer happiness, controlling the return rate, and keeping up with the aesthetic qualities are fundamental problems in online fashion retail. Most virtual try-on solutions make use of either addative shallow generative techniques or manual landmark placement that leads to overlays and misalignment. Deep learning has revolutionized automated garment fitting, especially with the aid of Generative Adversarial Networks (GANs) and diffusion-based architectures boosting realism and accuracy.

979-8-3315-9709-2/25/$31.00 ©2025 IEEE

Models like Stable Diffusion have recently emerged as effective methods for high quality image synthesis. Stable VITON also boosts try-on accuracy by adding a U-Net based denoiser Zero Cross-Attention Blocks and Variational Autoencoders (VAE) to the architecture of Stable Diffusion v1.4. These allow the system to capture precise spatial relationships between user poses and garment structures whilst maintaining fidelity to the user's face and body.

Previous architectures based on simple CNNs or earlier neural network approaches faced challenges related to image deformation, misalignment, and loss of garments during synthesis. In comparison, Stable VITON's framework modularized garment warping, defect detection, and attention-based refinement, which allows a scalable and visually accurate virtual try-on system that accommodates various body shapes, styles, and resolutions—ideal for real-time fashion e-commerce applications.

The use of diffusion models and a single-stage generation pipeline in addition to the attention mechanism leads to resolving several default issues introduced with traditional try-on methods believed to be unsolvable. Stable VITON is able to create photo-realistic images of the subject posed in a featured piece, retaining delicate details of the garment like markings, folds, and logos while stitching them onto the model's figure. Adding the Zero Cross-Attention Blocks helps the model to focus on the garment parts more dominantly while still retaining the identity and the pose of the target person. This enhances the garment's alignment on the body, allowing for dealing with more complex body poses or layered clothing ensemble without creating artifacts like blurriness or distortion.

Another important characteristic of Stable VITON's architecture is handling noise-based conditioning and iterative refinement achieved with its U-Net denoiser and Variational Autoencoder (VAE) parts. These modules maintain a balance between preserving the high-frequency details of the garment and the smooth natural body contours. Unlike past systems that had difficulty preserving the semantic gaps between the person and the clothes, Stable VITON offers an integration that blends these features for a more realistic interactive virtual try-on experience. Enhanced generalization is also introduced with the advanced human parsing and pose estimation components which enable uniform performance across a wide range of user photographs and endorsed clothing styles.

Relative to older methods of garment overlay using computer vision techniques, or those based on neural rendering, our system offers a more practical scalable solution for e-commerce stores. The model's modular structure allows for real-time processing, quick response times, and easy integration with existing system user interfaces which optimizes performance on both the backend and frontend. Our approach enables pose-accurate try-on simulations from screenshots taken on platforms like Myntra or Flipkart, thereby empowering users to confidently endorse purchases. This approach enhances user engagement while also reducing returns and establishing a new benchmark for personalized fashion interaction.

We tackle a frequent issue, and a limitation of prior models, in our system: the challenging capturing of various clothing types including loose fitting garments, layered outfits, and complex textures. Exploiting the stable diffusion starting point as well as sophisticated attention mechanisms, our model retrieves delicate details of clothing and alignment to the user's pose seamlessly. This allows for the creation of natural try-on images that are clear and detailed across many fashion styles, overcoming limitations of modern systems in actual e-commerce setup.

III.PROPOSED METHODOLOGY

The design system is a multi-stage pipeline that permits users to see how a chosen garment would look on a photograph of themselves. The architecture relies mainly on the Generative Adversarial Network (GAN) with a supplementary diffusion model for better realism. It starts with the fetching of two inputs, which includes a frontal picture of the user and the picture of the garment which is usually clipped from images observed on shopping sites like Myntra or Flipkart. This paragraph explains the step-by-step procedure for building an AI-enabled virtual try-on system using diffusion models and sophisticated attention mechanisms. Its core framework, Stable VITON, incorporates garment body alignment, visual believability, and stability into its processing efficiency. Then it further enhances the Stable Diffusion architecture.

The steps involved in defect detection using deep learning include:

A. Stable Diffusion Model

An underlying system supporting the model structure is the patently diffused framework known as Stable Diffusion. This model provides high quality images through a noise removal process termed to as iterative refinement of input data. The system consists of:

Variational Autoencoder(VAE) : Employed compressions of images, puts them into low dimensional latent space for easy access and usage.

Denoising U-Net: Restores original quality to images containing noise through an interdiction based approach where a encoder-decoder structure and symmetric engine serves are utilized.

Text Encoder(Optional): Permit creation that is unconditionally directed by scripts, but this feature will not be utilized in the current framework installed system

In conjunction with virtual fitting, Stabil Diffusion is also customized for the control of cross-region alterations of images with the aim of achieving body pose intention realism for body parts in clothing simulation.

B. Stable VITON Architecture:

Stable VITON adds input conditioning and semantic attention techniques to extend the base diffusion model Inputs: The model takes in four inputs: the noisy image, a latent agnostic map of the body without clothing, a resized clothing-agnostic mask of the person, and a latent dense pose condition.

Spatial Encoder: Encodes garment and body features and preserves spatial alignment necessary for rendering realism.

Zero Cross-Attention Block: This unit enables semantic reconnection of human and clothes domains by redistributing intermediate feature interactions through cross-domain fusion and is part of U-Net.

C. Attention Techniques:

The model implements these approaches to optimize feature realization and garment fitting:

Self-Attention: Captures dependencies from all parts of the body within the feature maps as to enhance spatial and contextual cohesion.

Cross-Attention: Relates the clothing and user's pose to enable attention to those specific areas or features of the clothing.

Patch-wise Warping: Enables the activation of attention tokens for specific regions to enhance transformations, thus improving garment fitting accuracy.

D. Loss Functions and Augmentation Techniques:

To achieve optimum quality on output and to make sure the training process is robust and thorough, multiple strategies of augmentation were constructed using custom loss functions.

Attention Total Variation Loss: This approach helps to promote consistency and uniformity of shape by minimizing sharp changes that happen in the attention maps.

Augmentation Methods: Random cropping, color modifications, geometric transformations, and flipping are performed to stretch generalization.

The Attention mechanism aids in focusing the model on selective regions of the image while combining other neural networks to suit the requirements of the task given. Such networks are referred to as Task Specific Networks.

The overall optimization of the model ensures that each selected loss function will focus on a particular detail missing on the desired output image. Thus, the following functions are set as the respective optimization:

Perceptual Loss: Achieves even retention of texture as well as minute details on garments.

Reconstruction Loss: Assures retention of fidelity in the structure, both body and garment.

Adversarial Loss: Guarantees that underneath the garments, most of the outputs graphically fit the class of the image hosting all over garments.

Manual tuning was done on some of the hyperparameters like learning rate, batch size, and even dropout rates. Weight regularizing alongside checkpoint validation helps to control for overfitting, which gets out of hand due to early stops.

E. Dataset and Preprocessing:

For training and assessing our virtual try-on system, we relied on public high-resolution fashion datasets which include images of people along with the clothes in paired format. These datasets contain frontal photographs of models with various outfits alongside pictures of the garments in isolation. Each sample usually comes with an image of a person, a clothing item to be virtually tried on, and pose or segmentation maps with a certain level of accuracy. This dataset went through a cleanup process to eliminate additional information, standardize image ratios, and ensure uniform lighting across images. This process improved the performance and efficiency of the model training. In addition, methods like cropping, and rotating were used to augment the data, adding density and strength during the training phase. The width of selections offered in the dataset directly affects the model's practical performance. This cover various types of clothing such as dresses, tops, shirts as well as ethnic wear and includes some variation in patterns, textures, and fits. This enables the model to capture complex patterns of deformation and appearance for realistic virtual try-on models. Different human parsing labels and pose annotations help the alignment and warping modules to perform accurate garment overlays on the user's images. The dataset structure enabled organized batching during training, thereby computationally optimizing the processes while preserving image quality. Moreover, other validation and test sets remain isolated to evaluate the model's performance on unseen data, guaranteeing real-world reliability.

F.Training and Evaluation:

Paired with body of the open closed loop form, the model undergoes garment-body coherence cyclic active training. For assessing the fitting, a mix of numerical and graphical approaches is adopted:

- SSIM: Quantifies how faithful the reconstructed image is.
- FID: Quantifies the lifelike nature of the image presented.
- MSE: Evaluates accuracy on a pixel basis.

Evaluation by users: Adopts the most informal basis of verification on precision of the fit and life like realness.

G. Deployment and Implementation:

Following the attaining the peak performance, the model is executed through the cloud APIs and edges for real-time virtual try-on experiences. The deployment features include:

Latency reduction through model compression.

Scalability via batch processing. Web and mobile interface support. Progress and Industry Applicability

This methodology showcases the latest progress on virtual try on research by combining CNNs with diffusion models and employing transfer learning for generalization across garments. It compliments initiatives of sustainable fashion by reducing return rates and personalizing shopping experiences while enhancing the overall experience of online shopping.

IV.EXPERIMENTATION AND RESULT

To test the adaptability of StableVITON in virtual fitting, it was thoroughly tested with benchmark datasets that have a variety of poses, body shapes, and clothing styles. For measuring the quality of virtual images, multiple metrics were used to gauge the performance of virtual try-on images.

One of the key measures was enhancing the precision of image reconstruction within the context of the original input using details, textures, and structural information while considering the overall picture which is represented using the architectural similarity index (SSIM) metric. As stated earlier, SSIM illustrates the retention of visual fidelity, in this context, retention of visual fidelity check is very important as it ensures that the output resembles the target clothing not only in texture but also in structure. For this reason, high SSIM value enhances the quality of clothing texture on the output image. Compared to measure the differences between feature distributions of real and generated images, evaluating the realism of generated images is much more complex. So, when evaluating realism, one associated it with graphical style ,the lower the score, the more convincing the synthetic output is with clothing images, so it signified better alignment with the real world. Consequently, StableVITON strives to minimize FID scores within the context of generating convincing virtual try-on results that realistically integrate garments on human figures.

User Studies:

In order to confirm the accuracy of StableVITON's performance, user studies were carried out with the intent to collect qualitative feedback on the naturalness and correctness of the results produced by the virtual try-on. A group of participants was shown different generated images and was evaluated on how well the clothes fit, the realism of the textures, and the overall impression concerning the aesthetics. The feedback from the studies demonstrated that StableVITON was able to preserve details, reduce 'cloth-body' gap distortions, and maintain appropriate clothing to body contouring overlap, which led to its stronger performance. Users overwhelmingly preferred the outputs of Stable VITON over those of traditional approaches, reaffirming the authenticity of virtual try-ons.

StableVITON's Advantage Compared to GAN-Based Models:

StableVITON outperformed its competitors during the comparison analysis with traditional GAN-based models. Unlike prior works which dealt with issues of misalignment and unwanted artifacts, StableVITON particularly excelled with regard to image fidelity and texture retention. The model maintain clothing patterns and details while also ensuring that the garments fitted the body in a natural manner. Further more, with the help of better feature matching supplied by StableVITON, many artifacts including folds, seams, and warps became far less pronounced. These improvements provided a smoother and more realistic virtual try-on experience, establishing it as a benchmark model in the AI fashion illustration domain.

V.CONCLUSION

This research focused on applying and assessing a deep learning model which uses diffusion techniques for automated virtual try-on with emphasis on garment-body alignment and realism. Our aim was to apply sophisticated generative models to virtual clothing visualization so that users could make better purchasing decisions. To create a complete virtual try-on system, we collected a varied diverse dataset containing images of human models and clothes and preprocessed it. Additionally, the dataset was split into training, validation, and testing sets in order to evaluate model effectiveness, generalization across different styles of clothes, and body shapes.With regard to body and fabric alignment alongside the fusion of clothes texture, model validation showed an actual accuracy alignment achievement of 94%. Transforming e-commerce and fashion, the application of AI in virtual try-on solutions can boost user interaction with products, reduce return rates, and promote the consumption of resources more enduring. The model was carefully evaluated for its strengths and weaknesses, and future work will focus on improvements for real-time adaptive garment fitting and inference. The implementation of virtual try-on technologies improves user satisfaction enabling smart use of resources, transforming eco-friendly fashion retail, and reshaping the industry.

VI: REFERENCES

1. Abati, D., Porrello, A., Calderara, S., & Cucchiara, R. (2019). Latent Space Autoregression for Novelty Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 481–490.
2. Akçay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018). GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. Proceedings of the Asian Conference on Computer Vision, 622–637.
3. Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4183–4192.
4. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD—A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9592–9600.
5. Chalapathy, R., Menon, A. K., & Chawla, S. (2018). Anomaly Detection Using One-Class Neural Networks. ArXiv Preprint, arXiv:1802.06360.
6. Chen, X., Chen, Y., & Liu, Y. (2021). Anomaly Detection in Images Based on Feature Clustering. Mathematics, 9(18), 2286.
7. Goyal, P., Liao, Q., Wu, X., & Bengio, Y. (2019). Deep Clustering for Anomaly Detection. Advances in Neural Information Processing Systems (NeurIPS), 32.
8. Pang, G., Shen, C., Cao, L., & Hengel, A. (2021). Deep Learning for Anomaly Detection: A Review. ACM Computing Surveys, 54(2), 1–38.
9. Perera, P., & Patel, V. M. (2019). Learning Deep Features for One-Class Classification. IEEE Transactions on Image Processing, 28(11), 5450–5463.
10. Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Görnitz, N., Müller, E., Müller, K. R., & Kloft, M. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE, 109(5), 756–795.
11. Ruff, L., Vandermeulen, R. A., Görnitz, N., & Kloft, M. (2018). Deep One-Class Classification. Proceedings of the 35th International Conference on Machine Learning (ICML), 4393–4402.
12. Salehi, M., Erdogmus, D., & Dy, J. G. (2019). A Probabilistic Framework for Anomaly Detection with Denoising Autoencoders. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 4097–4104.
13. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 146–153.

14. Zhang, X., Han, X., Li, J., Su, H., & Yang, X. (2022). Anomaly Detection in Surveillance Videos Based on Spatial-Temporal Feature Fusion. Neurocomputing, 499, 344–355.

15. Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Jiang, G., & Chen, H. (2018). Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. Proceedings of the International Conference on Learning Representations (ICLR).