Evaluating Generative Models for Virtual Try-On Applications: Performance Across Diverse Conditions

Shubham Singh, Xu Zhou, Inder Khatri

New York University, Tandon School of Engineering, New York, NY

Abstract

This research investigates advanced generative models for virtual try-on applications, with a specific focus on clothing. By evaluating these models under various lighting conditions, angles, and backgrounds, the study aims to identify the most effective solutions for practical use. Our comprehensive analysis includes a detailed comparison of model performance in different scenarios to determine their robustness and accuracy. The findings are expected to provide valuable insights into the application of generative models in the fashion industry, enhancing user experience and paving the way for future advancements in virtual try-on technology. Code base: https://github.com/ssnyu/ECE-GY-7123/tree/main/FinalProject

Introduction

The advent of generative models has revolutionized various industries, with virtual try-on applications emerging as a prominent use case in the fashion sector. These applications leverage sophisticated algorithms to allow users to visualize how clothing items would look on their bodies without physically trying them on. This technology not only enhances the online shopping experience but also addresses challenges related to fit and style selection.

Despite significant advancements, the performance of generative models in virtual try-on applications can be influenced by factors such as lighting conditions, viewing angles, and background settings. Variations in these parameters can affect the accuracy and realism of the rendered images, thereby impacting the overall user experience. As such, it is crucial to evaluate these models under diverse conditions to identify the most reliable and effective solutions for practical deployment.

The structure of this report is as follows: first, we review the existing literature on generative models and virtual try-on technology. Next, we outline our methodology for evaluating the models, including the experimental setup and criteria for assessment. We then present the results of our analysis, followed by a discussion of the findings.

Literature Survey

Image-based virtual try-on is a popular research topic in the field of AI-generated content (AIGC), specifically in the domain of conditional person image generation. It enables editing, replacement, and design of clothing image content, making it highly applicable in various domains such as ecommerce platforms and short video platforms. In particular, online shoppers can benefit from virtual try-on by obtaining try-on effect images of clothing, thereby enhancing their shopping experience and increasing the likelihood of successful transactions. In addition, AI Fashion has also emerged on short video platforms, where users can edit the clothes worn by characters in images or videos according to their own creativity. This allows users to explore their sense of fashion and produce a wide range of engaging images and videos.

The concept of virtual try-on was proposed as early as 2001, which uses a pre-calculated generic database to produce personally sized bodies and animate garments on a web application. Virtual try-on methods can be divided into three categories: physical-based simulation, real acquisition and image generation.

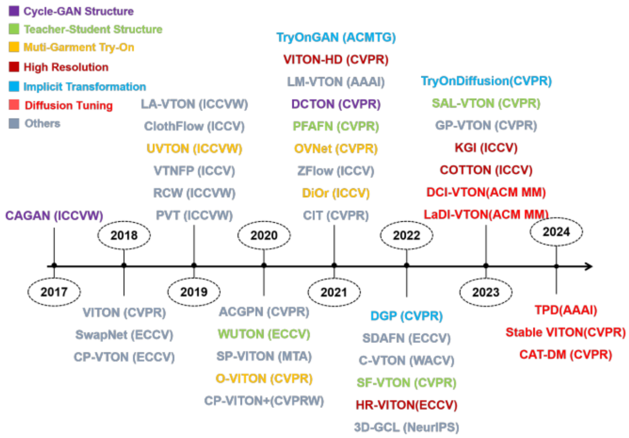

Figure 1: A concise timeline of image-based virtual try-on milestones. Different colors indicate the main characteristic of method.

Tremendous efforts have been made and Figure 1 show some representative methods on a timeline. In 2017, CAGAN gave the first try by employing CycleGAN to overcome the lack of training triplet data, i.e., (original person image, in-shop clothing image, try-on image), but the generation quality is far from satisfactory. Subsequently, VITON creatively proposed clothing-agnostic person representation by human parsing to make up the lack of supervised training data. They constructed the basic network framework of “Try-On Indication + Cloth Warping + Try-on”, laying the foundation for further improvement on generation quality in subsequent works.

Dataset

The Viton dataset is a large-scale collection of fashion product images, comprising 352,551 images across 164 categories, such as dresses, tops, shoes, and bags. The dataset was created by Zalando, a European e-commerce company, to support the development of machine learning models for fashion product classification. The images, with a resolution of 256x256 pixels, were collected from Zalando’s website and labeled using a combination of human annotation and automated processes.

The Viton dataset is publicly available for research and commercial use, and its diversity and scale make it a valuable resource for the fashion industry and machine learning community. All the experiments were carried out on this dataset.

Technical Details

We are experimenting with 3 models, measuring performance/loss with the criteria: **FID (Fréchet inception distance)** and **IS (Inception Score)**; two popular metrics used to evaluate the quality of images generated by generative models.

Fréchet Inception Distance (FID)

FID measures the similarity between two sets of images by comparing the statistics of their features extracted by an Inception-v3 network. The main steps are:

1. **Feature Extraction:** Pass images from both the generated and real datasets through the Inception-v3 network to obtain features from an intermediate layer.
2. **Statistical Modeling:** Fit a multivariate Gaussian model to the features of each dataset. Calculate the mean (μ) and covariance (Σ) of these distributions for both generated (g) and real (r) images.
3. **Distance Calculation:** Compute the Fréchet distance between these Gaussian distributions using the formula:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}),$$

where μ_r, μ_g and Σ_r, Σ_g are the means and covariances of the real and generated images, respectively.

A lower FID indicates that the generated images are more similar to the real images, suggesting better quality and realism of the synthetic images.

Inception Score (IS)

Inception Score evaluates the quality of generated images based on diversity and the clarity of object representation in the images:

1. **Probability Estimates:** Use a pretrained Inception model to predict the class probabilities $p(y|x)$ for each generated image.
2. **Score Calculation:** The Inception Score is calculated as:

$$\text{IS} = \exp(E_x[\text{KL}(p(y|x)\|p(y))]),$$

where $p(y|x)$ is the conditional class distribution given each image, $p(y)$ is the marginal class distribution across the dataset, and KL represents the Kullback-Leibler divergence.

A higher IS indicates that the generated images are diverse and each clearly resembles a specific class of objects, reflecting higher quality of the generated images.

DCI-VTON-Virtual-Try-On

The DCI-VTON framework leverages a combination of warping and refinement modules powered by diffusion models to create high-fidelity virtual try-on images:

- **Warping Module:** It calculates an appearance flow field to align clothes to the target person’s pose. The output is combined with a clothes-agnostic image of the person to produce a coarse initial result.
- **Refinement Module:** This module refines the coarse result using a diffusion model that adds and then denoises noise iteratively. The module employs local and global conditions to ensure the preservation of clothing details and realistic integration with the human figure.

Deep Fashion

Deep Fashion: Achieving Photo-Realistic Virtual Try-On through Adaptive Content Generation and Preservation Virtual try-on, which involves transferring a target clothing image onto a reference person, has gained significant attention in recent years.

Existing approaches focus on preserving clothing characteristics, such as texture and logo, when warping it to fit various human poses. However, generating photorealistic try-on images remains a challenge, especially when dealing with large occlusions and human poses. To address this, we propose the Adaptive Content Generating and Preserving Network (ACGPN), a novel visual try-on network that predicts semantic layouts, determines content generation or preservation needs, and produces photorealistic try-on images with rich clothing details.

ACGPN consists of three main modules:

- **Semantic Layout Generation:** This module predicts the desired semantic layout after try-on using semantic segmentation of the reference image.
- **Clothes Warping:** This module warps clothing images according to the generated semantic layout, with a second-order difference constraint for stable training.

- **Inpainting Module for Content Fusion:** This module integrates all information (reference image, semantic layout, warped clothes) to adaptively produce each semantic part of the human body. Compared to state-of-the-art methods, ACGPN generates photorealistic images with superior perceptual quality and richer fine details.

Hyperparameters and Optimization

- **Lambda Values:** Different lambda values (λ) balance the importance of the loss components in the warping network.
- **Optimizer:** The warping network uses the Adam optimizer, while the refinement process employs the AdamW optimizer, with specific learning rates and training durations set for each.

Training Details Training is conducted separately for the warping and refinement modules, focusing on accurate alignment and effective refinement in latent space using a pre-trained encoder-decoder framework.

High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions

Our goal is to synthesize an image of a person wearing a given clothing item while maintaining their pose and body shape. We achieve this through a two-stage framework: a try-on condition generator and a try-on image generator. At test time, we use discriminator rejection to filter out incorrect segmentation map predictions.

Pre-Processing We obtain a segmentation map, a clothing mask, and a pose map from off-the-shelf models.

Try-On Condition Generator This stage generates the segmentation map of the person wearing the clothing item and deforms the item to fit the person's body.

Condition Aligning To prevent misalignment, we refine the segmentation map and handle occlusion.

Loss Functions We use pixel-wise cross-entropy, L_1, perceptual, and total-variation losses to guide the training.

Try-On Image Generator This stage synthesizes the final try-on image using the clothing-agnostic image, warped clothing image, and pose map.

Discriminator Rejection We employ a discriminator rejection method to filter out low-quality segmentation map predictions.

Results

In this study, we evaluated advanced generative models for virtual try-on applications, focusing on their performance under diverse lighting conditions, angles, and backgrounds. Our findings indicate significant advancements in model capabilities, particularly with DCI-VTON, HR-VITON, and Deep Fashion.

The DCI-VTON model, which combines warping and refinement modules powered by diffusion models, demonstrated superior performance with a Fréchet Inception Distance (FID) of 0.1473 and an Inception Score (IS) of 2.948. This model excels in aligning clothing to the target person's pose and maintaining realistic details, resulting in highly accurate and visually appealing try-on images.

Deep Fashion, with its Adaptive Content Generating and Preserving Network (ACGPN), also showed strong performance. This model focuses on preserving clothing characteristics such as texture and logos, and adeptly handles large occlusions and varied human poses. The ACGPN generates photorealistic images with rich details by predicting semantic layouts, warping clothes according to these layouts, and using an inpainting module for content fusion. Despite slightly lower scores compared to DCI-VTON, Deep Fashion's emphasis on detail preservation and photorealism makes it a formidable solution for virtual try-on applications.

HR-VITON, another notable model in our evaluation, achieved a Fréchet Inception Distance (FID) of 0.1783 and an Inception Score (IS) of 1.432. This model demonstrates promising performance in virtual try-on tasks, particularly in handling diverse lighting conditions and backgrounds. With further refinement, HR-VITON has the potential to compete with leading models like DCI-VTON and Deep Fashion.

These findings underscore the progress made in virtual try-on technology and highlight the diverse strengths of state-of-the-art generative models. Future research should continue to explore novel techniques to further improve model robustness and performance under various challenging conditions.

Model Name	Fréchet Inception Distance (FID)	Inception Score (IS)
Deepfashion	0.1618	1.016
DCI-VTON	0.1473	2.948
HR-VITON	0.1783	1.432

Table 1: Model comparisons

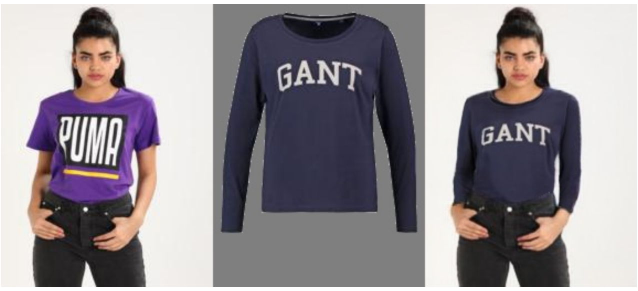

Figure 2: dci-vton result

Conclusion

In this study, we evaluated advanced generative models for virtual try-on applications, focusing on their performance. Our findings indicate that models like DCI-VTON, Deep Fashion, and HR-VITON significantly outperform earlier methods in generating high-fidelity, photorealistic try-on images.

DCI-VTON, with a lower Fréchet Inception Distance (FID) of 0.1473 and a higher Inception Score (IS) of 2.948, demonstrated superior performance in aligning clothing to the target person's pose and maintaining realistic details. Deep Fashion also showed strong results with its Adaptive Content Generating and Preserving Network (ACGPN), excelling in semantic layout generation and inpainting for content fusion. Additionally, HR-VITON achieved a FID of 0.1783 and an IS of 1.432, showing promising performance in virtual try-on tasks.

These insights highlight the advancements and potential of current generative models for practical virtual try-on applications in the fashion industry.

Future research should continue to address challenges such as large occlusions and varied human poses to further enhance realism and user satisfaction.

References

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021. Parser-Free Virtual Try-on via Distilling Appearance Flows. *arXiv preprint arXiv:2103.04559*.

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. Taming the Power of Diffusion Models for High-Quality Virtual Try-On with Appearance Flow. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10471–10480.

Lee, S.; Gu, G.; Park, S.; Choi, S.; and Choo, J. 2022. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. arXiv:2206.14180.

Yang, H.; Zhang, R.; Guo, X.; Liu, W.; Zuo, W.; and Luo, P. 2020. Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content. arXiv:2003.05863.

Appendix

Codebase:
https://github.com/ssnyu/ECE-GY-7123/tree/main/FinalProject