Virtual Try on Application and Parameter Analysis Based on Clothing Warping and Masking Technology

Chunbo Feng[a]
Data Science and Big Data, South China University of Technology, Guangzhou, China

Keywords: Virtual Fitting, DCI-VITON, Masking Techniques, Visual Quality.

Abstract: Virtual Fitting represents a pivotal application of technology, facilitating seamless online clothing purchases by enabling users to visualize garments on their bodies. As a focal point within the image generation domain, it holds immense commercial value, garnering substantial attention. This paper delves into the examination of parameters and training sessions' impact on Diffusion-based Conditional Inpainting for Virtual Try-ON (DCI-VITON), elucidating the efficacy of individual blocks and the overall model. Leveraging garment warping and masking techniques, the method preserves coarse character image structures while harnessing the potent generative capabilities of diffusion models to refine results. Experimental findings highlight the nuanced effects of various parameters on model performance, showcasing its superiority over alternative virtual fitting methods, particularly in capturing clothing details and styles. The model's ability to produce realistic, intricately detailed morphed garments with consistent visual quality underscores its significance. This project serves as a crucial reference for future virtual fitting techniques grounded in diffusion models.

1 INTRODUCTION

Along with the growing community of users shopping online, the commercial value of image-generated virtual try-on is increasing. Virtual fitting allows users to very visually see the effect of goods on the body instead of going to a physical store to try on clothes. This reduces the cost of fitting and greatly improves the user's online shopping experience. Virtual fitting technology is based on images of people as well as images of clothing to generate the final effect. It is ideal for online platforms as it is able to edit and replace the content of clothing images directly online (Song, 2023).

Virtual Try-on based on image generation presents several challenges. First of all, the image should retain the original pose and the original shape of the body in a complete and natural way. Secondly, in order to achieve a better effect, the costume should be naturally deformed so that the posture of the costume can fit more closely with the original posture and body shape of the character. In addition, the parts of the body that are initially covered by the clothing should be rendered appropriately. Satisfying these requirements was challenging due to the initial mismatch between the clothing and character images.

Generative adversarial networks (GANs) have underpinned most previous virtual try-on techniques. Nikolay et.al. Bergmann proposed Conditional Analogical GAN (CAGAN) which suggested a U-NET-based GAN (Jetchev, 2017). However, the convergence of GAN's generator is heavily dependent on the choice of hyperparameters and parameters, making it unrealistic and imperfectly detailed. In 2018, VITON has innovatively proposed an image-based virtual try-on network that does not make use of any 3D information (Han, 2017). Subsequently, the proposal of VITON-HD (Choi, 2021) and HR-VITON (Lee, 2022) further advanced the development of virtual try-on, which is no longer limited to low resolution. Recently, as the capabilities of diffusion models have been tapped in the field of generative models, more and more people have begun to investigate virtual try-on techniques based on diffusion models. In 2023, Aiyu CUI utilized the diffusion model for the first time to provide virtual fittings to passersby on the street, expanding the direct beneficiaries of virtual fittings from merchants to every one of us (Cui, 2023). TryOnDiffusion (Zhu,

[a] https://orcid.org/0009-0001-9835-3431

Feng, C.
Virtual Try on Application and Parameter Analysis Based on Clothing Warping and Masking Technology.
DOI: 10.5220/0012937600004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 328-332
ISBN: 978-989-758-713-9

2023) and StableVITON (Kim, 2023), launched by Google based on the diffusion model, have achieved the best performance so far in 1024x1024 single-piece top fitting.

Enhancing clothing quality post-deformation by adjusting parameters in the deformation network and optimizing clothing details constitute the primary objectives of this study. Furthermore, this research is dedicated to improving the visualization of generated clothing on the human body in virtual fitting tasks. To achieve this, a multi-step approach is employed. Firstly, a morphing network predicts the clothes to be tried on and the models wearing them, facilitating accurate clothing morphing. Secondly, Diffusion models refine initial synthetic results. Thirdly, various models' predictive performance is analyzed and compared. Additionally, the optimal model is determined through iterative training sessions, hyperparameter adjustments, and latent space downsampling. Experimental findings elucidate the impact of different parameters on model performance.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

This study utilizes the High-Resolution VITON-Zalando Dataset (VITON-HD), specifically designed for high-resolution virtual try-on tasks (Choi, 2021). VITON-HD comprises original images featuring female characters adorned in various garments, alongside associated agnostic masks, cloth masks, densepose images, and more, derived from these base images. The dataset is meticulously divided into distinct training and test sets, with the test set containing 11,647 pairs and the training set comprising 2,032 pairs. This extensive dataset serves as a comprehensive resource for training and evaluating the performance of the proposed methods in virtual fitting tasks. The sample is shown in the Figure 1.

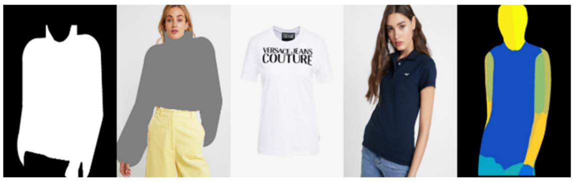

Figure 1: Images from VITON-HD dataset (Picture credit: Original).

2.2 Proposed Approach

This virtual try-on approach comprises two main components. Firstly, clothes are warped and merged with a masked image of the individual using a warping network. Secondly, the results from the initial step are refined using a diffusion model. Prior to this, the author preprocesses the individual's image to obtain segmentation results, dense pose, and clothes-agnostic data. To ensure compatibility between the clothing and the target individual, the deformation network predicts the appearance flow field, facilitating pixel-level matching between the clothing and the individual's torso image. Subsequently, the deformed clothing is integrated with the distorted character image, yielding an initial result. Following the addition of noise, the coarse output from the first step undergoes refinement through the diffusion model, resulting in an enhanced outcome after denoising. This method leverages the robust generative capabilities of the diffusion model, in conjunction with garment warping and masking techniques. The integration of these methodologies yields a seamless synthesis of the individual and the garment, thereby yielding excellent performance in virtual try-on tasks. The system's architecture is illustrated in Figure 2 below.

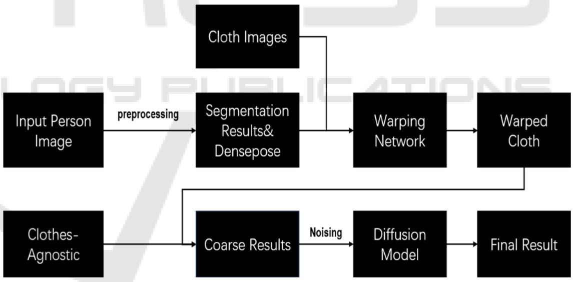

Figure 2: The pipeline of study (Picture credit: Original).

2.2.1 Warping Networks

There are two types of warping methods used for garment deformation, Thin Plate Spline deformation and appearance flow-based deformation. Thin Plate Spline (TPS) is one of the interpolation methods, which is a commonly used 2D interpolation method (Bookstein, 1989). TPS is often used in applications related to the deformation of image key points and so on. It will offset the control points of an image to achieve a specific deformation of the image through the control points.

Appearance flow-based deformation is based on the concept of appearance flow. Appearance flow refers to the difference in appearance between two images, which can be understood as the pixel-level correspondence between two images, as a way to deform and align images with each other (Gou, 2023). For this method, first, features are extracted from the two input images. Next, the correspondence between the two images is determined by calculating the difference in appearance between the two images. Using the estimated obtained appearance flow, feature points in one image are mapped to corresponding positions in the other image. In this way, as shown in Figure 4, the deformation between the images is realized. The process is shown in the Figures 3. and Figure 4.

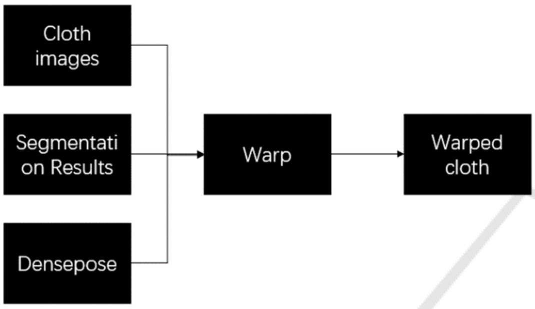

Figure 3: The process of warping network (Picture credit: Original).

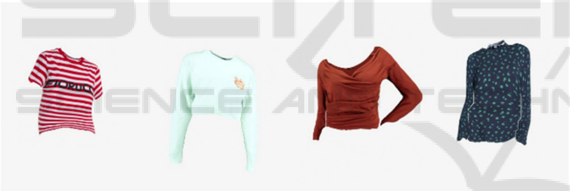

Figure 4: Warped clothes after passing through the warping network (Picture credit: Original).

2.2.2 Diffusion Model

Denoising Diffusion Probabilistic Models (DDPM) (Ho, 2020) is a method that changes the input image into pure Gaussian noise and then recovers it back to the image, utilizing the forward process and Gaussian distribution variance to generate the image.

However, because the Denoising Diffusion Probabilistic Model itself is a Markov chain process, it is impossible to avoid the iterative process, thus leading to its inference being too slow. Therefore, Robin Rombach proposed the latent space diffusion model (LDM). This model is a good combination of diffusion process and latent variable model, aiming at capturing the underlying structure and dynamic evolution in the data, which greatly improves the efficiency and quality of the generation (Rombach, 2021). By modeling latent variables and the diffusion process, the latent diffusion model can be used to learn structures and patterns in the data and used to generate new data samples.

In this approach, the diffusion model generates more refined results based on the initial results generated in the first part to make the clothing fit the portrait better and to process the details of the clothing to make the overall effect more realistic. There are two main parts to the training process for this module: the reconstruction and the refinement (Gou, 2023). In the reconstruction part, the model first performs forward diffusion processing on the target image, and then gradually adds noise to improve the robustness and generalization ability of the model. The model uses a latent space diffusion model to reduce computation time and improve model performance. The image is first inserted into the latent space with a pre-trained encoder. Then the image is reconstructed by the pre-trained decoder. In the refinement part, the model improves the similarity between the model predictions and the results generated by the deformation network by adding noise to the inputs. The training process of the model is done by optimizing the objective function of both branches while optimizing both parts of the model.

2.2.3 Loss Function

For the appearance flow, the total variation loss function is optimal because it should be regularized to ensure the smoothness of the appearance flow. Total-variation (TV) loss is a loss function commonly used in image processing and computer vision tasks. It is used to smooth an image to avoid excessive noise or discontinuities in the image. TV loss is often used as a regularization term in tasks such as image reconstruction, denoising, and super-resolution to help maintain the smoothness and continuity of an image. Given a 2D image I, the TV loss can be expressed as:

$$TV(I)=\sum_{i,j}\left(\left|I_{i+1,j}-I_{i,j}\right|+\left|I_{i,j+1}-I_{i,j}\right|\right) \quad (1)$$

where I_{ij} denotes the pixel value at position (i, j) in image I. This formula calculates the differences between adjacent pixels in the image and smoothes the image by minimizing these differences

In addition, the L1 loss and the perceptual loss are used to minimize the distance between the warped cloth and the pose of the person in the feature space. L1 loss is a loss function commonly used in regression models to achieve feature selection using L1 loss to simplify the model and improve generalization. The L1 loss is defined as:

$$\mathrm{L}_{L_1} = \sum_{i=1}^{N} \left\| y_i - f(x_i) \right\| \quad (2)$$

Where the true value represented by y_i is the deformation mask and the predicted value represented by $f(x_i)$ is the deformed cloth.

Perceptual loss utilizes the ability of the convolutional layer to abstract high-level features to approximate the human eye's perception of image quality thereby comparing the similarity of the generated image to the target image. This loss function better captures the semantic content and structure of the image, not just pixel-level differences. Perceived loss is also applied to the diffusion model. The Perceptual loss can be defined as:

$$\mathrm{L}_{vgg} = \sum_{i=1}^{N} \left\| \phi(y) - \phi(\hat{y}) \right\|_1 \quad (3)$$

3 RESULTS AND DISCUSSION

For evaluating the performance of the model, this study uses measures that are commonly used in the field of image generation: structural similarity index (SSIM), perceptual distance (LPIPS). In addition to this Frechette's Perceptual Distance (FID) is also used. By comparing with other virtual try-on methods, as shown in Table 1, it is found that the improved method achieves good results for all evaluation criteria. The SSIM evaluation criteria show that the generated image has a very high similarity to the original image. The value of LPIPS reaches 0.092, which also indicates that the two images are very similar. Overall, the improved method generates more realistic and better images and compares well with other SOTA methods.

Table 1: Quantitative comparison with baselines.

Resolution	512*384		
Method	SSIM ↑	LPIPS ↓	FID ↓
PF-AFN	0.885	0.082	11.3
HR-VITON	0.8623	0.1094	16.21
CP-VITON	0.791	0.141	30.25
DCI-VITON	0.896	0.043	8.09

This method turns down the learning rate of the Adam optimizer compared to the original DCI-VITON model. Compared to the default learning rate of 0.001, problems such as gradient explosion or gradient vanishing can be avoided, although the model takes longer to converge. In addition to the β_1 of the Adam optimizer is adjusted to 0.5, which prevents the gradient decay rate from being too high and making it difficult to jump out of the local minimum.

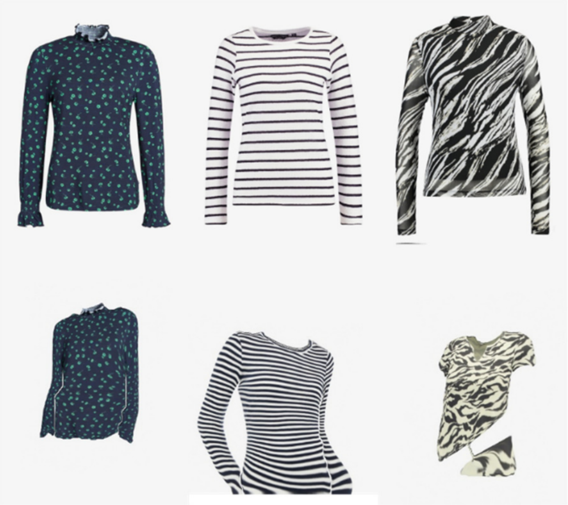

Figure 5: The comparison between the original garment and the deformed garment (Picture credit: Original).

As can be seen from the images generated in Figure 5, the first and second columns of clothing retain the original details (patterns and stripes) of the garments intact. There are no color differences and the style of the garments is consistent with the original garments. At the same time, the style of the deformed garment fits very well with the human body posture, structure and form. In terms of the color and pattern of the garment, although there are some deviations, the expected effect is generally achieved and the final effect is very realistic.

4 CONCLUSIONS

This study primarily focuses on analysing the structure of DCI-VITON and the impact of various parameters on its performance. The model's main innovation lies in the integration of the warping and refinement modules. The warping module combines warped clothing with masked characters to produce coarse results, while the refinement module employs the robust generative capabilities of the diffusion model to enhance image details after noise addition. The diffusion model training process involves two key stages: reconstruction and refinement, which collectively contribute to improved model performance. Additionally, adjustments to epochs and hyperparameters are made to compare performance effectively. The authors specifically concentrate on deformation networks, fine-tuning parameters to achieve accurate, realistic, and detail-preserving deformed clothing. Extensive experiments have been conducted to optimize the proposed method. Moving forward, the next phase of research will focus on implementing and enhancing diffusion models to generate complete virtual try-on figures. Furthermore, attention will be directed towards human accessories such as scarves, bracelets, and headbands, expanding the scope of virtual try-on applications.

REFERENCES

Song, D., Zhang, X., Zhou, J., Nie, W., Tong, R., & Liu, A. (2023). Image-Based Virtual Try-On: A Survey. arXiv, 2311.04811.

Jetchev, N., & Bergmann, U.M. (2017). The Conditional Analogy GAN: Swapping Fashion Articles on People Images. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp: 2287-2292.

Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L.S. (2017). VITON: An Image-Based Virtual Try-on Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp: 7543-7552.

Choi, S., Park, S., Lee, M.G., & Choo, J. (2021). VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp: 14126-14135.

Lee, S., Gu, G., Park, S.K., Choi, S., & Choo, J. (2022). High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. European Conference on Computer Vision.

Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., & Lazebnik, S. (2023). Street TryOn: Learning In-the-Wild Virtual Try-On from Unpaired Person Images. arXiv, 2311.16094.

Zhu, L., Yang, D., Zhu, T.L., Reda, F.A., Chan, W., Saharia, C., Norouzi, M., & Kemelmacher-Shlizerman, I. (2023). TryOnDiffusion: A Tale of Two UNets. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp: 4606-4615.

Kim, J., Gu, G., Park, M., Park, S.K., & Choo, J. (2023). StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. arXiv, 2312.01725.

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

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence, vol. 11(6), pp: 567-585.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. arXiv, 2006.11239.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp: 10674-10685.