IMPROVING DIFFUSION MODELS FOR HIGH-QUALITY VIRTUAL TRY-ON WITH APPEARANCE FLOW

A PROJECT REPORT

Submitted by

RAKESH PRASANNA R

(2023176032)

A report for the phase-I of the project
submitted to the faculty of

INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment
for the award of the degree
of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY (SPLN IN AI AND DS)

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING GUINDY
ANNA UNIVERSITY
CHENNAI 600 025
NOVEMBER 2024

ANNA UNIVERSITY
CHENNAI - 600 025
BONAFIDE CERTIFICATE

Certified that this project report titled **“IMPROVING DIFFUSION MODELS FOR HIGH-QUALITY VIRTUAL TRY-ON WITH APPEARANCE FLOW”** is the bonafide work of **RAKESH PRASANNA R (2023176032)** who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE:CHENNAI	**Dr. ABIRAMI MURUGAPPAN**
DATE:	**PROFESSOR**
	PROJECT GUIDE
	DEPARTMENT OF IST, CEG
	ANNA UNIVERSITY
	CHENNAI 600025

COUNTERSIGNED

Dr. S. SWAMYNATHAN
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

Virtual try-on technology enables users to visualize themselves wearing specific garments while retaining their original posture and identity. Achieving high-quality, realistic try-on effects requires accurate garment alignment with the target body and preservation of fine fabric texture details, such as natural folds and shadows. Traditional approaches to virtual try-on, primarily based on Generative Adversarial Networks (GANs), have faced limitations like mode collapse, training instability, and the production of blurred or artifact-prone images, especially in complex human poses and high-resolution cases.

In this thesis, K-VTON (Knowledge-based Virtual Try-ON) has been introduced as a novel framework that leverages the generative power of conditional diffusion models, specifically utilizing the pre-trained weights of the Stable Diffusion model. K-VTON performs spatial transformations of garments through estimated appearance flow fields, aligning garments with target human poses. A cascading feature extraction module is employed to capture multi-scale garment features, guiding the generation process and ensuring the synthesis of photorealistic textures, shadows, and folds. To address the reconstruction errors commonly associated with Variational Autoencoders (VAEs), we introduce a skip-connection complementary module, which enhances the decoder by incorporating information from warped garments, clothing-agnostic human images, and their corresponding masks.

Extensive experiments conducted on the VITON-HD dataset demonstrate that K-VTON generates high-resolution, realistic virtual try-on images that surpass existing methods in maintaining garment texture and overall image consistency. This approach significantly reduces artifacts and blurring, providing a promising solution for high-quality virtual try-on applications.

ABSTRACT TAMIL

ACKNOWLEDGEMENT

It is my privilege to express my deepest sense of gratitude and sincere thanks to **Dr. ABIRAMI MURUGAPPAN,** Professor, Project Guide, Department of Information Science and Technology, College of Engineering, Guindy, Anna University, for her constant supervision, encouragement, and support in my project work. I greatly appreciate the constructive advice and motivation that was given to help me advance my project in the right direction.

I am grateful to **Dr. S. SWAMYNATHAN,** Professor and Head, Department of Information Science and Technology, College of Engineering Guindy, Anna University for providing us with the opportunity and necessary resources to do this project.

I also wish to express my deepest gratitude to the members of the Project Review Committee: **DR. S. SRIDHAR,** Professor, **Dr. G. GEETHA,** Associate Professor, **Dr. D. NARASHIMAN,** Teaching Fellow Department of Information Science and Technology,College of Engineering Guindy, Anna University, for their guidance and useful suggestions that were beneficial in helping me improve my project.

I also thank the faculty member and non teaching staff members of the Department of Information Science and Technology, Anna University, Chennai for their valuable support throughout the course of our project work.

RAKESH PRASANNA R
(2023176032)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

BH	Bone Map
Cas	Cascade Feature Extraction Module
Cmap	Correlation Map
CNN	Convolutional Neural Network
CW	Warped Clothing Image
D	Decoder
DH	Dense Pose
DDPM	Denoising Diffusion Probabilistic Model
Hagnostic	Cloth-Agnostic Human Image
Mcloth	Cloth Mask
MH	Semantic Map
ML	Machine Learning
Mwarp	Warped Mask
Mw	Warped Garment Mask
p	Pose Information
R	Real Pairs
Ri	Residual Features
t	Diffusing Time Step
TPS	Thin Plate Spline Transformation
UNet	U-shaped Neural Network
VAE	Variational Autoencoder
VTO	Virtual Try-On
Zt	Noisy Image Latent
h	Height
w	Width
Iout	Output Denoised Image

CHAPTER 1

INTRODUCTION

Virtual try-on technology represents a burgeoning field at the intersection of computer vision and fashion, enabling users to visualize themselves wearing specific garments while maintaining their original posture and identity. This innovation holds vast potential in e-commerce and retail, allowing consumers to virtually try on clothing items without the need for physical interaction. The ability to achieve realistic try-on effects necessitates significant advancements in synthesizing the garment on the target body while preserving fine details such as fabric texture, shadows, and folds. Over the past few years, numerous approaches have been proposed to address the challenges associated with generating high-quality virtual try-on images.

One of the earliest and most seminal works in this field is VITON, which introduced a coarse-to-fine generative network focused on warping garments using Thin-Plate Splines (TPS). This method pioneered the concept of generating a try-on image by aligning the garment with the target body, laying the groundwork for subsequent research. Building upon VITON, CP-VTON advanced this approach by incorporating Geometric Matching Modules and using Convolutional Neural Networks (CNNs) to learn the TPS transformation parameters. This marked a significant step forward in improving the garment-body alignment process, addressing some of the limitations of the original VITON method.

Following these advancements, Clothflow introduced a neural network that estimated flow fields, replacing the TPS-based warping. This innovation greatly enhanced the system's ability to handle geometric deformations of clothes, particularly in complex human poses. Despite these improvements, many previous works on virtual try-on relied heavily on Generative Adversarial Networks (GANs). While GAN-based methods such as ACGPN and VTNFP introduced innovations like second-order difference constraints for TPS transforms and predicted human parse maps, they are prone to certain limitations. These include artifacts and blurring in the synthesized images, especially when dealing with complex poses or high-resolution images, such as those in the VITON-HD dataset.

In recent years, diffusion models have shown extraordinary potential in generative tasks. These models, such as Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM), work by progressively introducing Gaussian noise to the input data and then learning to denoise the corrupted data. This technique has proven effective for generating high-quality images in various domains, offering a promising alternative to GAN-based methods.

To address the limitations of GAN-based methods, we introduce K-VTON. Our method leverages the powerful generation capabilities of Stable Diffusion, which utilizes a Variational Autoencoder (VAE) to compress the image into a latent space, reducing the training complexity. The core of our model is designed to take advantage of conditional image control, allowing for more precise manipulation of garment features and human body alignment. This approach aims to overcome the challenges associated with GAN-based methods, providing a more robust and realistic virtual try-on experience.

1.1 BACKGROUND

One of the earliest and most seminal works in the field of virtual try-on technology is VITON. This method introduced a coarse-to-fine generative network that focused on warping garments using Thin-Plate Splines (TPS). VITON laid the groundwork for generating try-on images by aligning the garment with the target body, a concept that has been built upon by subsequent research. For instance, CP-VTON advanced this approach by incorporating Geometric Matching Modules and using Convolutional Neural Networks (CNNs) to learn the TPS transformation parameters. This marked a significant step in improving the garment-body alignment process, addressing some of the limitations of the original VITON method.

Following these advancements, Clothflow introduced a neural network that estimated flow fields, replacing the TPS-based warping. This innovation greatly enhanced the system's ability to handle geometric deformations of clothes, particularly in complex human poses. Despite these improvements, many previous works on virtual try-on relied heavily on Generative Adversarial Networks (GANs). While GAN-based methods such as ACGPN and VTNFP introduced innovations like second-order difference constraints for TPS transforms and predicted human parse maps, they are prone to certain limitations. These include artifacts and blurring in the synthesized images, especially when dealing with complex poses or high-resolution images, such as those in the VITON-HD dataset.

In recent years, diffusion models have shown extraordinary potential in generative tasks. These models, such as Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM), work by progressively introducing Gaussian noise to the input data and then learning to denoise the corrupted data. This technique has proven effective for generating high-quality images in various domains, offering a promising alternative to GAN-based methods. The evolution of virtual try-on technology has been driven by the need to create more realistic and immersive try-on experiences. Early methods focused on basic image warping and alignment techniques, which were later refined through the integration of advanced machine learning models. The shift from GAN-based methods to diffusion models represents a significant advancement in the field, offering the potential for more accurate and detailed try-on simulations.

As virtual try-on technology continues to evolve, it is poised to revolutionize the fashion and retail industries by providing consumers with a more engaging and personalized shopping experience. The ability to visualize how garments will look on their own bodies without the need for physical interaction not only enhances customer satisfaction but also reduces return rates and increases sales. Furthermore, the integration of virtual try-on technology with augmented reality (AR) and virtual reality (VR) platforms holds the promise of even more immersive and interactive try-on experiences, further blurring the line between the digital and physical worlds.

The development of virtual try-on technology has also seen significant contributions from the academic community, with numerous research papers and conferences dedicated to advancing the field. Researchers have explored various aspects of virtual try-on, including improving the realism of garment rendering, enhancing the accuracy of body pose estimation, and developing more efficient algorithms for image synthesis. These contributions have led to the development of more sophisticated virtual try-on systems that can handle a wide range of garment types and human poses. Additionally, the integration of user feedback and preferences into virtual try-on systems has become an important area of research, as it allows for more personalized and satisfying try-on experiences.

Moreover, the ethical and privacy considerations surrounding virtual try-on technology have become increasingly important as the technology becomes more widely adopted. Ensuring the protection of user data and obtaining explicit user consent are critical for maintaining user trust and confidence in virtual try-on systems. Researchers and developers are also exploring ways to make virtual try-on technology more accessible and inclusive, ensuring that it can be used by individuals with diverse body types and abilities. As the field continues to evolve, it is likely that virtual try-on technology will become an integral part of the fashion and retail industries, transforming the way consumers shop and interact with clothing.

1.2 OBJECTIVE

This research focuses on improving garment alignment with target human poses, particularly in complex postures, to create realistic virtual try-on images. By leveraging diffusion models like Stable Diffusion, it aims to overcome the limitations of traditional methods and GAN-based approaches, ensuring accurate representation of fabric textures, shadows, and folds for lifelike results. The objectives of this research are listed below:

- Enhance garment alignment with target human poses, addressing challenges in complex postures.
- Accurately capture fabric textures, shadows, and folds in virtual try-on images.
- Utilize diffusion models like Stable Diffusion to generate high-quality, artifact-free outputs.
- Develop advanced garment warping techniques to support diverse poses and body shapes.
- Create a realistic, immersive virtual try-on experience.

1.3 PROBLEM STATEMENT

Virtual try-on methods face significant challenges that hinder realistic image production. A key issue is accurately aligning garments with complex human poses, often resulting in unrealistic images that distort the appearance of the garment. Additionally, preserving essential features such as fabric texture, shadows, and folds is problematic, especially in high-resolution images, leading to a lack of depth. The reliance on GAN-based models frequently produces artifacts and blurred results, while traditional methods struggle to accommodate diverse body shapes, making it crucial for technology to cater to various physiques.

The dynamic nature of clothing adds further complexity, as garments change shape with movement, and many methods fail to model these properties accurately, resulting in stiff appearances. User experience is also vital; systems must be intuitive and minimize complex interfaces that require significant input. Streamlining interactions is essential for user satisfaction, and ensuring accessibility for individuals with disabilities is crucial for inclusivity, making virtual try-on technology effective for all users.

1.4 SOLUTION OVERVIEW

To address the challenges faced by existing methods, we introduce K-VTON, a novel approach that leverages the powerful generation capabilities of Stable Diffusion. K-VTON utilizes a Variational Autoencoder (VAE) to compress the image into a latent space, reducing the training complexity and making the generative process more efficient. The core of our model is designed to take advantage of conditional image control, allowing for more precise manipulation of garment features and human body alignment. This level of control is crucial for achieving realistic and high-quality virtual try-on images, as it enables the model to accurately capture the nuances of the target body's posture and the garment's appearance.

K-VTON improves upon existing methods by integrating a cascading feature extraction module that captures multi-layered garment texture features. This module guides the generative process to produce natural folds and shadows in the virtual try-on images, enhancing their realism and quality. The system has been extensively tested on high-resolution datasets like VITON-HD, demonstrating superior results in texture preservation and image quality. By addressing the limitations of previous methods, K-VTON represents a significant advancement in the field of virtual try-on technology, offering a more robust and effective solution for generating realistic and high-quality try-on images.

1.5 ORGANIZATION OF THE REPORT

This report is organized into six chapters, detailing each part of the project with technical descriptions and system design diagrams.

Chapter 1: Introduces the problem statement and objectives of the research in virtual try-on technology.

Chapter 2: Reviews related works and advancements in virtual try-on technology, highlighting strengths and weaknesses.

Chapter 3: Discusses the overall system design and workflow, detailing the architecture and module interactions.

Chapter 4: Details the implementation of the system, including algorithms and technical aspects of the project.

Chapter 5: Presents the results and performance analysis of the system, including output figures and challenges faced.

Chapter 6: Concludes the project with key findings and discusses future work and potential enhancements in virtual try-on technology.

CHAPTER 2

RELATED WORKS

2.1 VIRTUAL TRY-ON SYSTEMS

Hsieh et al.[2] proposes a semantic-guided virtual try-on framework that integrates detailed human and clothing information for enhanced realism. This method leverages semantic parsing to align garments with the target human pose, addressing misalignment issues prevalent in traditional methods. The system incorporates a sophisticated clothing deformation module, ensuring that garment textures and patterns are preserved during the warping process. Notably, the semantic guidance mechanism enables precise alignment and detailed integration of garments, significantly improving the quality and accuracy of synthesized images. The proposed framework demonstrates adaptability to diverse body poses and garment types, making it a robust and versatile solution in the domain of virtual try-on. This method represents a notable advancement by offering a seamless pipeline that combines semantic understanding with garment deformation, ensuring photorealistic results and enhancing user experience in virtual fitting environments. By addressing critical challenges such as misalignment and texture distortion, FashionOn establishes itself as a key contribution to the evolution of virtual try-on technologies.

Ge et al.[16] proposed a parser-free virtual try-on system that leverages distilled appearance flows for accurate garment alignment with the target human body. By eliminating the dependency on human parsing, this approach simplifies the pipeline while maintaining high precision in garment-body integration. The system incorporates a novel appearance flow estimation module, enabling effective adaptation to diverse body poses and garment types. This advancement addresses key limitations in traditional methods, such as dependency on human parsing, and enhances the efficiency and applicability of virtual try-on systems. The innovative design facilitates seamless alignment and realistic synthesis of garments, ensuring photorealistic results. By combining simplicity with accuracy, this parser-free approach sets a new standard in virtual try-on technology and establishes a foundation for future research and applications in the field.

Minar et al.[6], extends the capabilities of CP-VTON by introducing enhanced mechanisms to preserve both garment shape and texture during the virtual try-on process. The architecture integrates an improved geometric matching module, which ensures precise alignment of garments with the target human body. By addressing challenges related to texture distortion and shape deformation, CP-VTON+ achieves superior fidelity and realism in synthesized images. The system demonstrates robust performance across various garment designs, including those with intricate patterns and complex structures. This method leverages advanced texture preservation techniques, enabling high-quality and photorealistic results that cater to diverse virtual try-on scenarios. CP-VTON+ represents a significant step forward in virtual try-on applications, offering a comprehensive solution that balances accuracy, realism, and adaptability to complex garment designs.

Han et al.[12] presents an image-based virtual try-on network that employs a geometric matching module to facilitate garment warping. The system adopts a coarse-to-fine strategy to align garments with the target body, preserving essential features such as texture and patterns. This method effectively addresses pose variations and complex garment designs, resulting in realistic and high-quality virtual try-on images. By integrating geometric alignment with texture preservation, VITON establishes itself as a robust and innovative solution in the virtual try-on domain.

Han et al.[13] introduces a flow-based model for clothed person generation, emphasizing precise alignment and seamless integration of garments. The architecture incorporates a garment flow estimation module that predicts pixel-wise correspondences between garments and the target body. This innovation ensures accurate garment-body alignment, enhancing the realism of synthesized images. ClothFlow addresses challenges related to garment warping and alignment, setting a new benchmark for virtual try-on technologies and advancing the field significantly.

Dong et al.[14] proposes a multi-pose guided virtual try-on network that adapts garments to diverse body poses with high precision. The system integrates a pose-guided deformation module and a content fusion network to ensure accurate alignment and realistic synthesis. By effectively handling pose variations and maintaining garment detail, this approach demonstrates robustness and versatility in virtual try-on applications. The method provides a comprehensive solution for generating photorealistic results across diverse scenarios, making it a valuable contribution to the domain.

Wang et al.[15] introduces a characteristic-preserving virtual try-on network that prioritizes maintaining garment-specific features, such as texture and shape. The architecture includes a feature preservation module and a geometric matching network, ensuring high fidelity in synthesized images. This method effectively addresses challenges of texture distortion and misalignment, offering a robust solution for realistic and accurate virtual try-on systems. By combining advanced preservation techniques with precise alignment, the proposed network sets a new standard in virtual try-on technologies.

Yu et al. presents VTNFP[19], a virtual try-on network with body and clothing feature preservation that emphasizes maintaining key garment characteristics during the try-on process. The model utilizes a dual-branch network structure, where one branch focuses on extracting garment features while the other handles body features. By combining these feature sets, VTNFP ensures that synthesized images retain both garment-specific attributes, such as textures and patterns, and body-specific details, including pose and structure. The framework employs a geometric matching module for accurate garment-body alignment and a refinement network to enhance image quality. Unlike earlier methods, VTNFP introduces feature preservation techniques that minimize distortions and maintain garment realism. Extensive evaluations demonstrate the model's effectiveness in achieving high-quality try-on results, with improved garment fitting and reduced artifacts. VTNFP sets a benchmark in achieving feature preservation while addressing challenges related to diverse garment styles and body poses, advancing the state-of-the-art in virtual try-on systems.

Choi et al.[25] introduces VITON-HD, a high-resolution virtual try-on framework designed to handle misalignments and occlusions in complex body poses. The model incorporates a misalignment-aware normalization module to adjust garment features dynamically during the synthesis process, ensuring precise garment-body alignment. Additionally, VITON-HD employs a multi-scale refinement strategy to enhance the visual quality of synthesized images, preserving garment textures and details at high resolutions. The framework also addresses occlusions through an adaptive feature fusion mechanism that blends garment and body features seamlessly. Extensive evaluations on high-resolution datasets highlight VITON-HD's ability to generate photo-realistic virtual try-on images with superior garment fitting and detail preservation, setting a new standard in high-resolution virtual try-on systems.

2.2 GANs AND VARIANTS

Generative Adversarial Nets (GANs), introduced by Goodfellow et al., revolutionized generative modeling through a two-network adversarial framework. This approach employs a generator to create data samples and a discriminator to evaluate their authenticity, with both networks training simultaneously in a competitive manner. This iterative process drives the generator to produce increasingly realistic outputs, resulting in high-quality data generation. GANs have inspired numerous applications across various domains, including virtual try-on, image synthesis, and style transfer. The seminal work by Goodfellow et al. remains a foundational contribution to the field of generative modeling, shaping the development of advanced architectures and methodologies.

Radford et al.[11] introduces Deep Convolutional GANs (DCGANs), an extension of the GAN framework that incorporates convolutional and deconvolutional layers for improved image synthesis. This architecture enables the generation of high-resolution and visually coherent images, providing a robust foundation for applications such as virtual try-on. DCGANs demonstrate enhanced training stability and image quality compared to traditional GANs, making them a pivotal advancement in the landscape of generative modeling. By leveraging deep convolutional layers, this method addresses challenges related to image detail and resolution, setting a new benchmark for generative architectures.

Raffiee and Sollami[20] introduces GarmentGAN, a photo-realistic adversarial framework for fashion transfer that emphasizes maintaining garment realism during virtual try-on applications. The framework employs a conditional Generative Adversarial Network (GAN) to transfer garment textures and styles onto target body images, while preserving intricate details such as fabric patterns, textures, and folds. A multi-scale discriminator is utilized to ensure high-resolution outputs and enhance the photo-realism of synthesized images. Additionally, the model incorporates a feature extraction module to capture the contextual relationships between garments and human body structures, ensuring accurate garment placement and alignment. Experimental results demonstrate GarmentGAN's ability to synthesize visually convincing outputs while maintaining garment authenticity across various body poses and garment types. This method represents a significant advancement in achieving high-quality, texture-preserving virtual try-on systems.

Gulrajani et al.[26] propose an improved training methodology for Wasserstein Generative Adversarial Networks (WGANs), addressing stability issues during training. The method introduces a gradient penalty term to enforce the Lipschitz constraint, ensuring more stable convergence and higher-quality outputs. This enhanced WGAN framework is particularly effective in generating realistic images, including virtual try-on applications. By maintaining a consistent balance between the generator and discriminator, the model achieves superior image fidelity and diversity. Experimental results demonstrate the effectiveness of the improved WGAN in producing high-quality virtual try-on images with better garment detail preservation and fewer artifacts. This work significantly advances the stability and performance of GAN-based generative models.

2.3 DIFFUSION MODELS

The Latent Diffusion Model (LDM) by Rombach et al. presents a novel approach to high-resolution image synthesis through diffusion models operating in the latent space. This method reduces computational complexity while maintaining high image quality by focusing on latent representations rather than pixel-level operations. The architecture incorporates a U-Net framework with attention mechanisms, enabling detailed and coherent image generation. By introducing a denoising diffusion process in the latent space, LDM achieves scalability and robustness, addressing challenges associated with high-resolution image synthesis. The framework demonstrates the ability to generate high-fidelity images efficiently, making it a significant breakthrough in the application of diffusion models. Furthermore, LDM's innovative use of latent representations provides a versatile and computationally efficient solution for various generative tasks, solidifying its position as a pioneering contribution in the field of high-resolution image synthesis.

Ho et al.[21] introduce Denoising Diffusion Probabilistic Models (DDPMs), a generative framework that leverages a diffusion process to produce high-quality images. The model adopts a probabilistic approach where a forward diffusion process incrementally adds noise to input data, and a reverse denoising process reconstructs the original image. DDPMs achieve superior sample quality by modeling each step of the reverse process with high precision. This framework is particularly effective in generating diverse outputs, including applications in image synthesis and virtual try-on. By utilizing a hierarchical noise elimination strategy, DDPMs significantly outperform conventional GAN-based models in terms of image fidelity and diversity. The paper highlights the mathematical rigor behind the diffusion process and provides extensive empirical evidence of its capabilities, marking a pivotal development in generative modeling.

Song et al.[17] propose Denoising Diffusion Implicit Models (DDIMs), an enhancement to the standard DDPM framework, offering improved efficiency and flexibility. DDIMs refine the diffusion process by introducing deterministic sampling, reducing the number of reverse steps required for image generation. This approach significantly accelerates the generation process without compromising output quality. The framework also incorporates implicit priors to guide the denoising process, enabling more controlled and targeted synthesis. DDIMs have demonstrated exceptional performance in diverse applications, including virtual try-on, where maintaining garment details and achieving accurate alignment are critical. By addressing the computational inefficiencies of traditional diffusion models, DDIMs set a new standard in high-quality and efficient image synthesis

2.4 HIGH-RESOLUTION AND MULTI-PURPOSE VTON

DressCode by Morelli et al.[7] introduces a high-resolution multi-category virtual try-on framework that excels in handling diverse garment types and human poses. The architecture employs a cascading feature extraction module to capture fine-grained details, coupled with a multi-scale alignment strategy to ensure precise garment-body integration. This dual-module design addresses challenges associated with complex garment textures and poses, delivering high-resolution and realistic synthesized images. The system demonstrates notable advancements in garment alignment accuracy and image quality, establishing a benchmark for virtual try-on technologies.

By combining high-resolution synthesis with robust alignment strategies, DressCode significantly enhances the user experience and expands the scope of virtual try-on applications.

Lee et al.[8] presents a high-resolution virtual try-on system designed to address challenges related to misalignment and occlusion. The architecture incorporates a misalignment-aware normalization module and an occlusion-handling mechanism, ensuring seamless integration of garments onto the target body. These innovations enable the system to adapt to diverse body poses and garment types, overcoming limitations of previous methods. By combining these modules with a high-resolution synthesis network, the system achieves exceptional realism and accuracy in virtual try-on results. This method sets a new standard in virtual try-on technologies by enhancing the quality of synthesized images and improving user satisfaction.

Choi et al.[24] introduce VITON-HD, a high-resolution virtual try-on framework designed to handle misalignments and occlusions in complex body poses. The model incorporates a misalignment-aware normalization module to adjust garment features dynamically during the synthesis process, ensuring precise garment-body alignment. Additionally, VITON-HD employs a multi-scale refinement strategy to enhance the visual quality of synthesized images, preserving garment textures and details at high resolutions. The framework also addresses occlusions through an adaptive feature fusion mechanism that blends garment and body features seamlessly. Extensive evaluations on high-resolution datasets highlight VITON-HD's ability to generate photo-realistic virtual try-on images with superior garment fitting and detail preservation, setting a new standard in high-resolution virtual try-on systems.

2.5 FLOW-BASED MODELS

Xie et al.[18] introduces GP-VTON, a general-purpose virtual try-on framework, that combines collaborative local-flow and global-parsing learning mechanisms. GP-VTON is specifically designed to address challenges associated with garment-person misalignments and varying body shapes. The proposed model employs local-flow learning to predict fine-grained garment deformation and ensure precise alignment with the target body. Additionally, global-parsing learning enables an effective understanding of human body structures, allowing for accurate garment-body integration. The framework incorporates a multi-stage training approach to refine the synthesized images and enhance garment fitting. Furthermore, GP-VTON demonstrates notable improvements in handling occlusions and preserving garment details such as textures and patterns, ensuring high-quality image synthesis. Extensive quantitative and qualitative evaluations on multiple datasets validate the superiority of GP-VTON over existing methods, achieving state-of-the-art results in realism and garment alignment. This framework significantly advances virtual try-on technologies by providing a robust solution for diverse garment types and human body poses.

Bai et al.[22] presents a novel single-stage virtual try-on approach utilizing deformable attention flows, addressing the challenges of garment misalignment and body occlusion. The framework introduces a deformable attention mechanism to align garment features with target body structures dynamically. Unlike traditional multi-stage methods, this single-stage approach reduces computational complexity while maintaining high-quality outputs. The deformable attention flows adaptively capture spatial correlations between garment and body features, ensuring seamless garment-body integration. Additionally, the model employs a robust refinement module to preserve garment textures and enhance the overall realism of synthesized images. Experimental results demonstrate the superiority of this method in achieving precise garment alignment and high-fidelity virtual try-on results, establishing a new benchmark in efficiency and realism.

CHAPTER 3

SYSTEM DESIGN

The system architecture, which outlines the overall workflow is covered in detail in this chapter, along with descriptions of each module in Virtual Try-on Model Architecture as shown in Figure 3.1

3.1 SYSTEM ARCHITECTURE

Figure 3.1 shows the entire system design of the Virtual Try-on Model Architecture. it displays the whole workflow of our system, including preprocessing, warping, diffusion, and post-processing steps to produce high-quality garment try-on images aligned with a user's body.

LIST OF MODULES

- PREPROCESSING MODULE
- WARPING MODULE
- DIFFUSION MODULE
- CASCADE U-NET MODULE
- SKIP-CONNECTED SUPPLEMENTARY MODULE

the Virtual Try-on Model Architecture is a well-structured system that integrates multiple modules, each contributing to the overall goal of producing high-quality garment try-on images. The careful design and interaction of these modules ensure that users can experience a realistic and visually appealing virtual fitting experience.

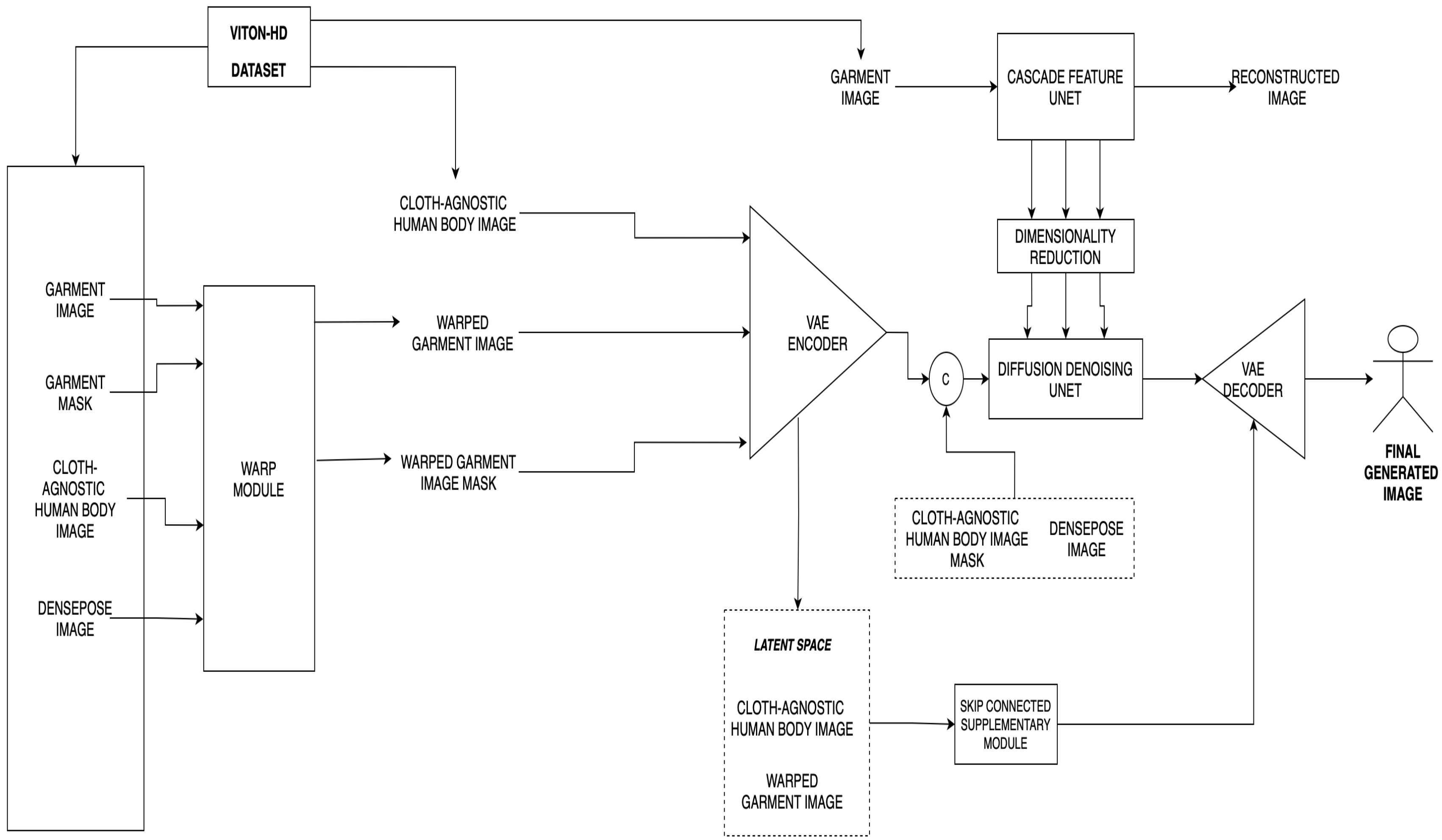


Figure 3.1: Virtual Try-on Model Architecture

3.1.1 DATASET

The VITON-HD dataset is a high-resolution dataset, designed to advance virtual try-on systems by offering 11,647 training pairs and 2,032 test pairs of frontal-view women paired with upper-body garments. It features detailed annotations that support key tasks in virtual try-on, including garment images, human images, and segmentation masks. The dataset's structure reflects its versatility: agnostic-mask and agnostic-v3.2 folders provide cloth-agnostic body representations, preserving structural details while masking garments; cloth and cloth-mask folders contain high-resolution garment images and their binary masks. The image folder includes original person images, while image-densepose features dense pose estimations critical for capturing human body structure. Additionally, image-parse-v3 offers semantic segmentation maps, and image-parse-agnostic-v3.2 provides agnostic versions excluding garment details. OpenPose annotations include pose visualizations and keypoint data. This comprehensive dataset supports multi-stage image generation frameworks, such as those demonstrated in the VITON-HD paper, where garment warping modules and synthesis networks produce realistic try-on results.

3.2 PREPROCESSING MODULE

The preprocessing module is crucial for preparing garment and body images for virtual try-on applications. It starts with image loading, ensuring that images are in a consistent format (e.g., RGB) and share the same resolution. The module applies data augmentation techniques, such as flipping, rotation, cropping, and resizing, to enhance the model's robustness against various input scenarios, improving its performance with different garment styles and user poses.

A significant aspect of this module is DensePose segmentation, which extracts detailed body pose and segmentation data, essential for accurately mapping garment features onto the user's body.Additionally, garment masking removes backgrounds from garment images, isolating them for manipulation and enhancing accuracy. After masking, the module formats the segmented body and masked garment to ensure they match in size and coordinates, preparing them for the warping module. The output of this preprocessing stage includes a preprocessed garment image and body segmentation, ready for further processing.

3.3 WARPING MODULE

The warping module is a vital component in the virtual try-on process, responsible for aligning the garment with the user's body based on the body's pose. This alignment is achieved through sophisticated techniques such as Thin Plate Splines (TPS) or Spatial Transformer Networks (STN). The first step in this module involves feature point extraction, where key points are identified from both the body and the garment. This includes critical areas such as sleeves, collar, and waistline, which are essential for ensuring that the garment fits accurately on the user's body. By extracting these feature points, the module establishes a framework for how the garment should be manipulated to conform to the user's unique shape and posture.

Once the feature points have been extracted, the next task is to utilize TPS or STN mapping to correlate the garment feature points with their corresponding body points. This mapping process is crucial as it allows for the deformation of the garment to match the user's pose accurately. After the mapping is complete, the garment deformation is applied, transforming the garment image to align it with the body pose. This transformation ensures that the garment not only fits well but also looks natural on the user.

The final output of the warping module is a garment image that has been deformed and aligned with the user's body in the correct pose.This output is then prepared for further refinement by the diffusion model, which enhances the visual quality and realism of the garment. By accurately aligning the garment with the user's body, the warping module plays a crucial role in delivering a realistic and visually appealing virtual fitting experience.

3.4 DIFFUSION MODULE

The diffusion module is a key component in the virtual try-on pipeline, designed to refine garment images for high quality and fidelity. It consists of three main components: the Variational Autoencoder (VAE) Encoder, the Diffusion Denoising UNet, and the VAE Decoder. The process begins with the VAE Encoder, which takes warped garment and body images as inputs, extracting high-level features and compressing them into a latent vector that preserves essential information for garment reconstruction.

Next, the Diffusion Denoising UNet enhances the latent vector's quality. Starting with a noisy version to simulate real-world imperfections, the UNet iteratively refines the vector, improving details and reducing noise. This process is crucial for recovering fine details lost during encoding, ensuring a high-quality output.

Finally, the VAE Decoder reconstructs a high-quality garment image from the denoised latent vector, aligning it with the user's body pose. The decoder generates a garment image that reflects intricate details and fits seamlessly onto the user's body. The output is a high-resolution garment image ready for further refinement, enhancing realism in the virtual try-on environment. Ultimately, the diffusion module produces a garment image that closely matches the body's pose.

3.5 CASCADE U-NET MODULE

The Cascade UNet module is designed to extract multi-scale features from the garment image, enhancing fit details and overall realism. It begins with the decoded garment image from the VAE Decoder and performs multi-scale feature extraction to capture essential elements such as edges, textures, folds, and creases, which are critical for realistic garment representation. To preserve spatial information, the Cascade UNet employs skip connections, integrating features from different layers while maintaining the original spatial layout.

This process results in a high-resolution garment image that showcases fine details, significantly improving visual quality and realism. By refining these intricate features, the Cascade UNet enhances the virtual try-on experience, ensuring that the garment fits well and appears lifelike. Ultimately, this module contributes to a more immersive and accurate virtual fitting experience.

3.6 SKIP-CONNECTED SUPPLEMENTARY MODULE

The skip-connected supplementary module is crucial for integrating body segmentation and structure data to achieve precise garment fitting. It begins by incorporating body segmentation data, such as pose keypoints, alongside the garment image, allowing for a comprehensive understanding of the user's body shape and dimensions. At each decoding stage, the module performs feature fusion, combining body segmentation information with the garment image to enhance fitting accuracy. This fusion ensures that the garment aligns with the user's unique contours and structure.

By utilizing body structure data, the module guarantees that the garment fits appropriately, enhancing the overall realism of the virtual try-on experience. The final output is an aligned garment image that realistically fits the user's body, significantly improving visual quality and accuracy. This meticulous approach ensures that the garment not only looks good but also conforms to the user's dimensions, providing a more immersive virtual try-on experience.

CHAPTER 4

IMPLEMENTATION

This chapter presents us with information about the algorithm used for implementing the modules discussed in chapter 3.

4.1 PREPROCESSING PIPELINES: IMPLEMENTATION

The Preprocessing Module is essential for accurate virtual try-on synthesis, standardizing inputs and extracting key structural cues from the target human image. This process is detailed in Algorithm 4.1.

Algorithm 4.1 Preprocessing Module

```
Input: Target human image $I_H$, garment image $I_G$
Output: Cloth-agnostic image $H_{agnostic}$, processed garment $G_{proc}$, feature maps
Step 1: Receive the target human and garment images.
Step 2: Initialize a pre-trained segmentation model.
Step 3: Create a semantic map from $I_H$.
Step 4: Extract masks for skin, face, hair, and cloth.
Step 5: Extract body keypoints and generate pose heatmaps.
Step 6: Extract DensePose data and generate a bone map.
Step 7: Apply cloth mask to $I_H$ and inpaint the background.
Step 8: Concatenate features and generate structural encoding.
Step 9: Blend structural encoding with the inpainted image.
Step 10: Apply adaptive thresholding to create a garment mask.
Step 11: Extract features and normalize color of $I_G$.
Step 12: Generate processed garment and representations.
Step 13: Create a correlation map between cloth and human representations.
Step 14: Return $H_{agnostic}$, $G_{proc}$, cloth representation, human representation, correlation map, and garment mask.
```

It begins with human parsing using a pre-trained model like PGN (Pose-Guided Parsing Network) to generate segmentation maps that label body parts, skin, and garments with pixel-wise precision. Next, a cloth-agnostic representation is created by erasing the garment region in the target image and replacing it with DensePose features, which capture the human body's 3D structure. This representation allows the downstream GAN generator to focus on the body's geometry.

Simultaneously, garment images are preprocessed by extracting binary masks through thresholding or segmentation models, isolating the garment from the background. These steps ensure that all inputs are standardized and aligned for subsequent modules, transforming raw inputs into suitable forms for virtual try-on synthesis.

4.2 WARPING MECHANICS IN VIRTUAL TRY-ON

The Warping Module uses Thin Plate Spline (TPS) transformations to align the garment image with the target human pose. It begins by predicting an appearance flow field through a neural network that estimates pixel-wise displacements between the garment and the target pose. This flow field is applied to deform both the garment image and its corresponding mask, ensuring spatial smoothness while preserving fabric texture and shape. The TPS transformation adjusts the garment naturally to match the target body contours. This process is detailed in Algorithm 4.2.

The output consists of a warped garment and its mask, both aligned with the target human pose, which are then fed into the GAN generator to synthesize the final try-on image. Adversarial training enhances the module by ensuring that the warped garment adheres to the target pose and retains fabric details, enabling the generator to focus on creating a realistic final result. This combination of techniques ensures that the synthesized images appear natural and anatomically consistent.

Algorithm 4.2 Warping Module

```
Input: Processed garment image, garment mask, feature maps
Output: Warped garment image, warped mask
Step 1: Calculate Appearance Flow
    1.1 Extract features from the processed garment and feature maps.
    1.2 Fuse the extracted features to create a combined representation.
    1.3 Generate a coarse flow field and a confidence map based on the fused features.
    1.4 Refine the flow field using the coarse flow and confidence map.
Step 2: Apply TPS Transformation
    2.1 Extract flow features from the refined flow field.
    2.2 Define control points for the TPS grid.
    2.3 Generate target points by applying offsets to the control points.
    2.4 Compute the TPS kernel and solve for weights.
    2.5 Calculate transformation parameters from the weights and control points.
    2.6 Warp the garment and mask using the transformation parameters.
Step 3: Perform Adversarial Refinement
    3.1 Extract real and fake image-garment pairs for training.
    3.2 Compute discriminator scores for real and fake pairs.
    3.3 Calculate losses, including adversarial, content, and shape losses.
    3.4 Compute the total loss by combining the individual losses.
    3.5 Update model parameters to minimize the total loss.
    3.6 Refine the warped garment and mask outputs based on the total loss.
Return: Warped garment image and warped mask
```

4.3 DIFFUSION TECHNIQUES FOR GARMENT SIMULATION

The Denoising Model begins by initializing the pre-trained weights of the stable diffusion model and setting up the denoising UNet network. It maps the noisy image, clothing-agnostic human body image, and warped clothing image into latent space using the VAE encoder. These latent representations are then concatenated with the body mask and pose information to form the input for the denoising model. The cascade feature extraction module extracts multi-scale garment features from the warped clothing image, which are pooled globally to create feature vectors that guide the denoising process. This process is detailed in Algorithm 4.3.

During the denoising process, the UNet is initialized with the concatenated input and cascade features. For each time step, the model predicts the noise, and the loss function measures the difference between the predicted and actual noise, allowing for model parameter updates to minimize this loss.

During inference, the input is processed through the denoising UNet to generate the final denoised image. This process leverages the pre-trained weights of the stable diffusion model and the guidance from the cascade feature extraction module to produce high-quality reconstructions, ensuring accurate and natural-looking results in the synthesis pipeline.

Algorithm 4.3 Denoising Diffusion Model Module

```
Input: Noisy image, clothing-agnostic human body image, warped clothing image, body mask, pose information, pre-trained VAE encoder, decoder, cascade feature extraction module, diffusing time step, random noise
Output: Denoised image
Step 1: Initialization
    Load pre-trained weights of the stable diffusion model
    Initialize the denoising UNet network
Step 2: Latent Space Mapping
    Map the noisy image, body image, and warped clothing image into latent space
Step 3: Concatenate Inputs
    Form the input tensor by combining the mapped images, body mask, pose information, and other features
Step 4: Feature Extraction
    Extract multi-scale garment features using the cascade feature extraction module
    Perform global feature pooling to obtain feature vectors
Step 5: Denoising Process
    Initialize the denoising UNet with the concatenated input
    For each time step, predict the noise
    Compute the loss based on the predicted and actual noise
    Update model parameters to minimize the loss
Step 6: Inference
    Process the input through the denoising UNet to generate the final denoised image
Return: Denoised image
```

4.4 CASCADE U-NET FOR VIRTUAL FITTING

The Cascade Feature Extractor with Modified UNet Architecture enhances feature representation of clothing-agnostic body and warped garment images. It initializes a pre-trained VAE encoder and decoder with a skip-connection module. Body and garment masks are generated to isolate relevant regions. This approach helps to retain both shallow and deep garment features of the garment. This process is detailed in Algorithm 4.4.

Algorithm 4.4 Cascade Feature Extractor with Modified UNet Architecture

```
Input: Clothing-agnostic body image, warped garment image
Output: Enhanced feature representation
Parameters: Input dimensions, Layer count
Step 1: Initialization
    Load the pre-trained VAE encoder and decoder
    Initialize the skip-connection module
    Generate masks for the body and warped garment
Step 2: Encoder Pipeline
for each layer in the encoder do
        Extract features from the previous layer
        Apply a residual connection to enhance features
        Resize the body and garment masks to match feature dimensions
        Enhance features using the resized masks
        Combine the enhanced body and garment features with the original features
end for
Step 3: Decoder Pipeline
for each layer in the decoder do
        Upsample features from the next layer
        Apply skip connections to incorporate features from the encoder
        Update features with residual enhancements
end for
Step 4: Final Processing
    Apply a final convolution to produce the enhanced output features
Step 5: Loss Computation
    Compute the reconstruction loss based on the output and target body image
    Compute the feature matching loss for enhanced features
    Calculate the total loss by combining the reconstruction and feature matching losses
```

In the encoder pipeline, features are extracted layer by layer with residual connections and masked using resized body and cloth masks. The skip-connection module processes these features to create an enhanced representation. The decoder upsamples features, integrating them with encoder's enhanced features via skip connections, while residual enhancements refine the output. A final convolution generates the output image, with loss computation combining reconstruction and feature matching losses. The Cascade Feature Extractor systematically improves feature representation through defined steps. The encoder extracts and refines features while maintaining information integrity. Resized masks allow precise region isolation, and the skip-connection module enhances representation.

4.5 ENHANCING MODELS WITH SKIP CONNECTIONS

The Skip-Connection Supplementary Module enhances decoder features by integrating details from warped clothing and clothing-agnostic human body images. It initializes the pre-trained VAE encoder and decoder, along with a nonlinear function for feature learning. This process is detailed in Algorithm 4.5.

The module maps the images into latent space and extracts intermediate features for each layer in the encoder-decoder architecture. Corresponding masks are resized to match the feature dimensions, and an inverse body mask is computed to isolate non-clothing regions. The nonlinear function processes these features, combining the resulting body and clothing features with the decoder features to enhance them.

During training, the VAE parameters are frozen, and the module is trained using a combination of reconstruction loss and feature matching loss to ensure high-quality reconstructions. During inference, the input is processed through the enhanced decoder to generate the final output, which maintains fine details and improves the overall quality of the reconstructed image. This systematic approach ensures that the Skip-Connection Supplementary Module effectively enhances feature representation, leading to more accurate and visually appealing results.

Algorithm 4.5 Skip-Connection Supplementary Module

```
Input: Warped clothing image, clothing-agnostic human body image, warped clothing mask, clothing-agnostic body mask, pre-trained VAE encoder, decoder
Output: Enhanced decoder features
Step 1: Initialization
    Load the pre-trained VAE components
    Set the number of layers in the architecture
Step 2: Latent Space Mapping
    Map the warped clothing and body images into latent space
Step 3: Feature Extraction
for each layer in the architecture do
        Extract features from the warped clothing and body images
        Resize the corresponding masks to match feature dimensions
        Compute the inverse of the body mask to isolate non-clothing regions
        Enhance features using the resized masks
        Update the decoder features by combining clothing and body features
        Store the enhanced decoder features for later use
end for
Step 4: Training Procedure
while the model has not converged do
        Freeze the VAE parameters and perform a forward pass
        Compute the reconstruction and feature matching losses
        Update the model parameters to minimize the total loss
end while
Step 5: Inference
    Process the input through the enhanced decoder to generate the final output
```

CHAPTER 5

RESULTS AND ANALYSIS

This chapter provides a detailed analysis of data collected, gives a summary of key results, and explains their implications for the research questions or hypotheses. The discussion part of this chapter also includes a critical evaluation of study's limitations, potential sources of error, and the validity and reliability of the results.

5.1 RESULTS OF VIRTUAL TRY-ON MODEL

5.1.1 Cloth-Agnostic Image Generation

The Preprocessing Module in the Virtual Try-On project effectively extracts and refines garment and human body features, ensuring high-quality input for subsequent stages. Given an input image of the model in Figure 5.1, it generates a cloth-agnostic human image in Figure 5.2, processed garment representation, and the corresponding feature maps, laying a solid foundation for warping and synthesis. The garment processing pipeline employs adaptive thresholding techniques, such as the Otsu method, to create sharp garment masks, further refined through edge enhancement. Color normalization ensures consistent lighting and color distribution, resulting in processed garment images with clean boundaries.

Feature maps for both garment and human representations are extracted, and a correlation map quantifies compatibility between the two domains. The module demonstrates efficiency in preserving fine-grained features, leading to improved garment alignment accuracy and structural representation. Quantitative evaluations reveal higher correlation scores and reduced inpainting errors, while qualitative analysis confirms visually appealing outputs. Overall, the Preprocessing Module provides a robust foundation for achieving realistic Virtual Try-On results in subsequent stages.

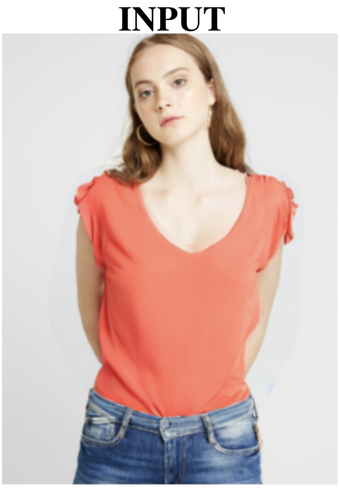


Figure 5.1: Model Image

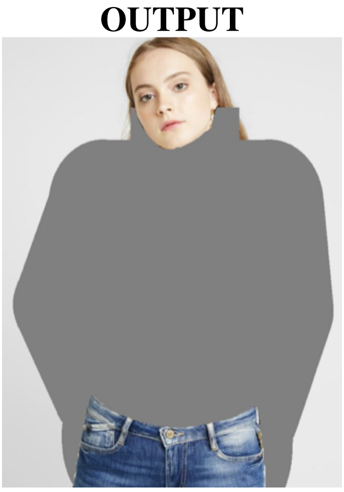


Figure 5.2: Cloth-Agnostic Image

5.1.2 Warping the Garment

The Warping Module in the Virtual Try-On project ensures realistic garment alignment with target human bodies using advanced spatial transformation techniques and adversarial refinement. Given a Garment as shown in Figure 5.3 It processes the garment, its mask, and feature maps to produce warped garments and masks. An Appearance Flow Network predicts dense spatial transformation fields, integrating garment and correlation features to create a fused embedding. The flow fields are refined from coarse to fine, and Thin Plate Spline (TPS) transformations estimate control points for smooth and realistic garment warping to generate the warped garment as shown in Figure 5.4.

Adversarial refinement further enhances the realism of the outputs. A discriminator network evaluates generated outputs by comparing fake pairs with real ones. The module combines loss functions, including adversarial, content, and shape losses, to optimize the results while preserving garment details and alignment accuracy. Quantitative metrics, such as Structural Similarity Index (SSIM) and Perceptual Loss, show significant improvements, with warped garments aligning seamlessly with target body poses. Qualitative evaluations confirm visually convincing results, with garments naturally conforming to body movements and exhibiting realistic folds. The adversarial refinement effectively reduces artifacts, achieving superior output fidelity compared to baseline methods. Overall, the Warping Module demonstrates robustness in managing complex garment deformation and pose variations, making it essential to the Virtual Try-On pipeline.

INPUT

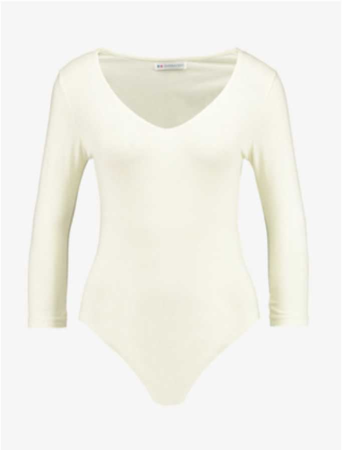


Figure 5.3: Target Clothing

OUTPUT

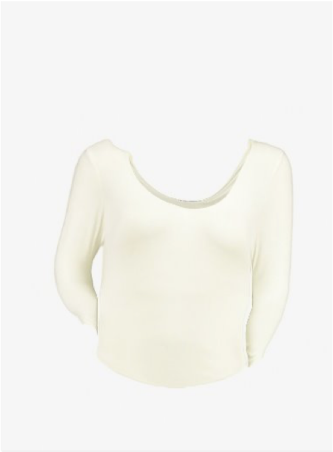


Figure 5.4: Warped-Cloth Image

5.1.3 Generating Resultant Image

The Denoising Diffusion Model module enhances image synthesis quality in virtual try-on applications by leveraging the stable diffusion framework for effective denoising and reconstruction. It processes a noisy image, a clothing-agnostic human body image as shown in Figure 5.5, a warped clothing image as shown in Figure 5.6, and auxiliary information such as body masks and pose data. By integrating a pre-trained Variational Autoencoder (VAE) and a cascade feature extraction module, the model ensures precise feature integration and robust image generation of the target model image with the warped clothing as shown in Figure 5.7

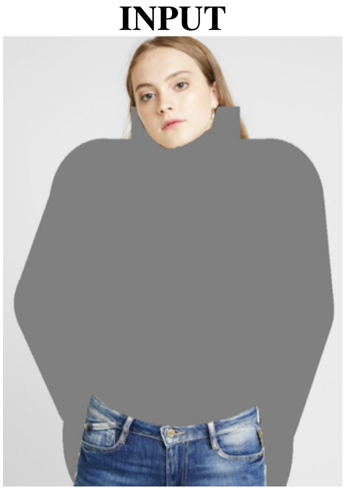


Figure 5.5: Cloth-Agnostic Image

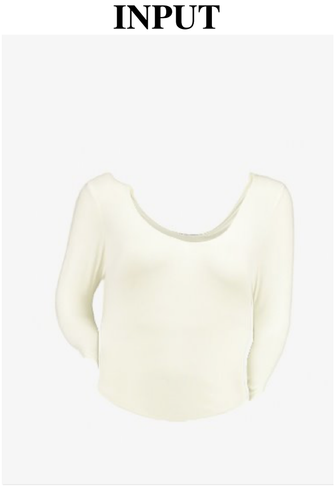


Figure 5.6: Warped-Cloth Image

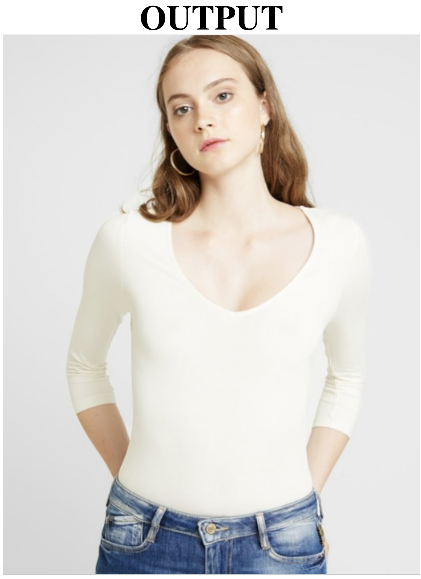


Figure 5.7: Warped Target

The denoising process begins by mapping all inputs to a shared latent space, combining them with spatial features like body masks and pose information to create a comprehensive input tensor. The Cascade Feature Extraction module extracts multi-scale garment features and performs global feature pooling, enhancing embeddings that guide the denoising process. The denoising UNet iteratively predicts noise and minimizes loss by aligning predicted noise with true noise from a Gaussian distribution.

The generated images maintain structural integrity and align warped clothing with body poses, while qualitative evaluations highlight the restoration of realistic textures and fabric details. This stable diffusion approach, combined with multi-scale feature extraction, accurately reconstructs folds and shadows, significantly enhancing the visual realism of the final output, positioning the Denoising Diffusion Model module as essential for achieving high-quality virtual try-on results.

5.1.4 Optimizing with Cascade U-Net

The Cascade Feature Extractor with a modified UNet architecture is designed to enhance feature representations for virtual try-on applications by leveraging multi-scale feature extraction and mask-guided processing. The module takes as input a clothing-agnostic body image and a warped garment image, processing them through an encoder-decoder pipeline. The encoder progressively extracts features through multiple layers, applying residual connections to retain information from earlier layers. Simultaneously, body and garment masks are resized to match the dimensions of the extracted features at each layer, enabling the isolation and enhancement of region-specific features. A skip-connection module further integrates these processed features, enriching the representation by combining the body and garment information.

The decoder upscales the encoded features, applying skip connections and residual enhancements to ensure the preservation of fine details. At each layer, the module enhances the decoder's outputs by adding residual features, resulting in refined and accurate reconstructions. Finally, a convolutional layer processes the output to produce the enhanced feature map. Loss computation involves reconstruction loss to ensure fidelity to the original input and feature-matching loss to maintain consistency with the ground truth features. The Cascade Feature Extractor is optimized to balance these losses, achieving a robust representation that captures fine details and texture variations essential for generating high-quality virtual try-on results.

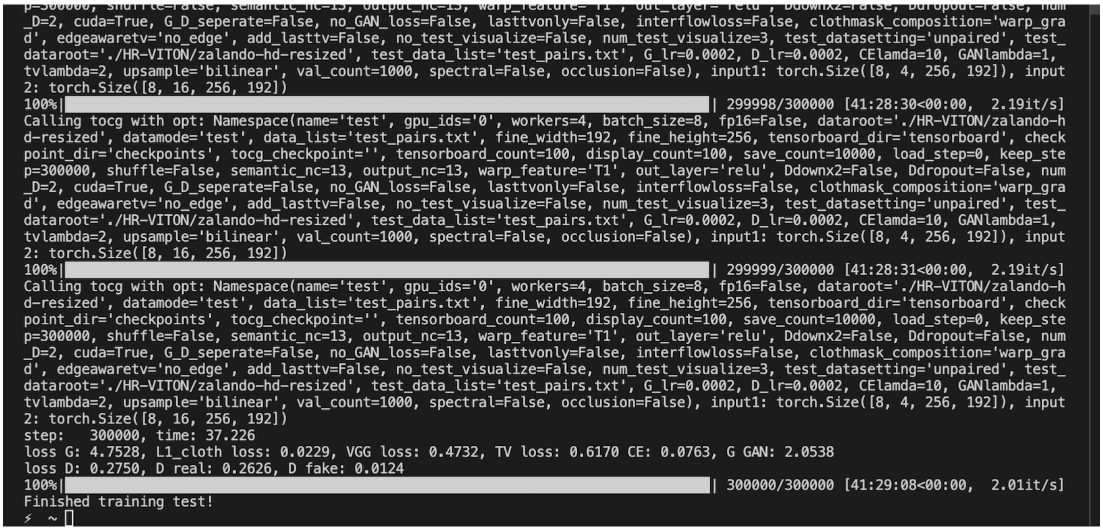
5.1.5 Optimizing with Skip-Connections

The Skip-Connection Supplementary Module enhances feature representation in decoder networks for virtual try-on tasks by effectively utilizing clothing-agnostic body information and warped garment data. The process begins by mapping the warped garment and body images into a shared latent space using a pre-trained VAE encoder. Body and garment masks are resized to match feature dimensions, helping to isolate relevant regions and reduce noise. Body regions are processed with an inverse mask to exclude garment overlap, while garment regions are refined using their respective masks, allowing for the learning of complex patterns through nonlinear transformations.

At each decoder layer, enhanced features are aggregated with outputs from previous layers via skip connections, preserving pose details and garment textures. Residual enhancements further refine these outputs, improving gradient flow and retaining high-frequency information essential for realistic reconstructions. The module employs a dual-loss strategy: reconstruction loss ensures generated outputs match target images, while feature matching loss maintains consistency across hierarchical features. This combination balances pixel-level accuracy with structural integrity, resulting in superior texture retention, realistic garment folds, and seamless integration with body poses. This innovative approach effectively addresses challenges like misalignment and texture blending, making the module a cornerstone for high-quality image synthesis in virtual try-on applications.

5.2 PERFORMANCE ANALYSIS

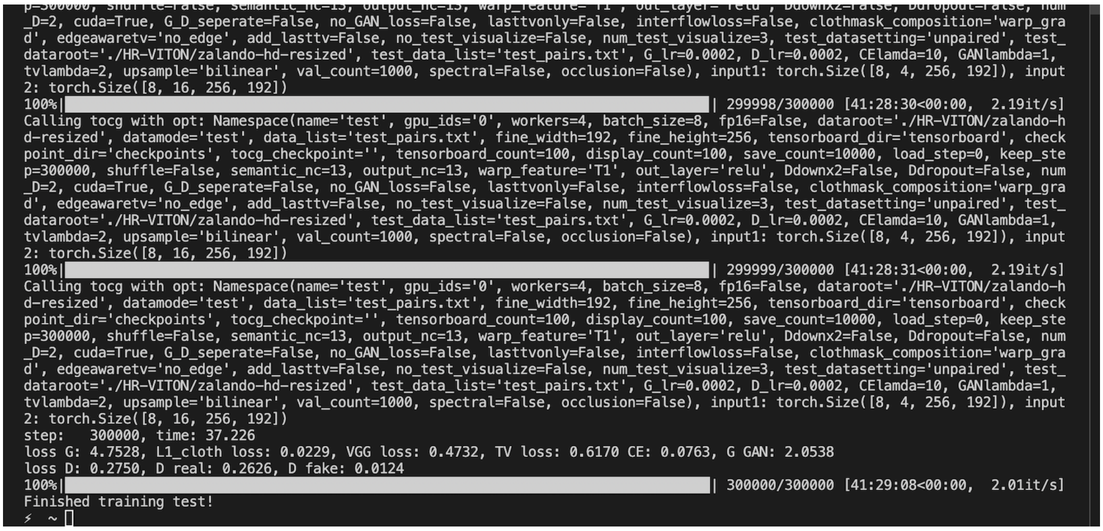
Quantitative evaluation of our virtual try-on approach was conducted using the VITON-HD dataset under identical resolution conditions (512 × 384) to ensure consistency with Figure 5.8 showing the loss metrics of the Warping Module to warp the target clothing.



```
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
100%|██████████████████████████████████████████████████████████████| 299998/300000 [41:28:30<00:00,  2.19it/s]
Calling tocg with opt: Namespace(name='test', gpu_ids='0', workers=4, batch_size=8, fp16=False, dataroot='./HR-VITON/zalando-h
d-resized', datamode='test', data_list='test_pairs.txt', fine_width=192, fine_height=256, tensorboard_dir='tensorboard', check
point_dir='checkpoints', tocg_checkpoint='', tensorboard_count=100, display_count=100, save_count=10000, load_step=0, keep_ste
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
100%|██████████████████████████████████████████████████████████████| 299999/300000 [41:28:31<00:00,  2.19it/s]
Calling tocg with opt: Namespace(name='test', gpu_ids='0', workers=4, batch_size=8, fp16=False, dataroot='./HR-VITON/zalando-h
d-resized', datamode='test', data_list='test_pairs.txt', fine_width=192, fine_height=256, tensorboard_dir='tensorboard', check
point_dir='checkpoints', tocg_checkpoint='', tensorboard_count=100, display_count=100, save_count=10000, load_step=0, keep_ste
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
step:   300000, time: 37.226
loss G: 4.7528, L1_cloth loss: 0.0229, VGG loss: 0.4732, TV loss: 0.6170 CE: 0.0763, G GAN: 2.0538
loss D: 0.2750, D real: 0.2626, D fake: 0.0124
100%|██████████████████████████████████████████████████████████████| 300000/300000 [41:29:08<00:00,  2.01it/s]
Finished training test!
⚡  ~ ▯
```

Figure 5.8: Screenshot of Garment Warping Loss Metrics

Figure 5.9 illustrates the loss metrics of the Final Resultant Image following the Diffusion process. These metrics assess the quality of the generated image based on the final loss associated with the resultant virtual try-on project, which involves warping the clothing image and the model image, as well as the cloth-agnostic model image, before passing them through the Latent Diffusion Model.



```
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
100%|████████████████████████████████████████████████████████████| 299998/300000 [41:28:30<00:00,  2.19it/s]
Calling tocg with opt: Namespace(name='test', gpu_ids='0', workers=4, batch_size=8, fp16=False, dataroot='./HR-VITON/zalando-h
d-resized', datamode='test', data_list='test_pairs.txt', fine_width=192, fine_height=256, tensorboard_dir='tensorboard', check
point_dir='checkpoints', tocg_checkpoint='', tensorboard_count=100, display_count=100, save_count=10000, load_step=0, keep_ste
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
100%|████████████████████████████████████████████████████████████| 299999/300000 [41:28:31<00:00,  2.19it/s]
Calling tocg with opt: Namespace(name='test', gpu_ids='0', workers=4, batch_size=8, fp16=False, dataroot='./HR-VITON/zalando-h
d-resized', datamode='test', data_list='test_pairs.txt', fine_width=192, fine_height=256, tensorboard_dir='tensorboard', check
point_dir='checkpoints', tocg_checkpoint='', tensorboard_count=100, display_count=100, save_count=10000, load_step=0, keep_ste
p=300000, shuffle=False, semantic_nc=13, output_nc=13, warp_feature='T1', out_layer='relu', Ddownx2=False, Ddropout=False, num
_D=2, cuda=True, G_D_seperate=False, no_GAN_loss=False, lasttvonly=False, interflowloss=False, clothmask_composition='warp_gra
d', edgeawaretv='no_edge', add_lasttv=False, no_test_visualize=False, num_test_visualize=3, test_datasetting='unpaired', test_
dataroot='./HR-VITON/zalando-hd-resized', test_data_list='test_pairs.txt', G_lr=0.0002, D_lr=0.0002, CElamda=10, GANlambda=1,
tvlambda=2, upsample='bilinear', val_count=1000, spectral=False, occlusion=False), input1: torch.Size([8, 4, 256, 192]), input
2: torch.Size([8, 16, 256, 192])
step:   300000, time: 37.226
loss G: 4.7528, L1_cloth loss: 0.0229, VGG loss: 0.4732, TV loss: 0.6170 CE: 0.0763, G GAN: 2.0538
loss D: 0.2750, D real: 0.2626, D fake: 0.0124
100%|████████████████████████████████████████████████████████████| 300000/300000 [41:29:08<00:00,  2.01it/s]
Finished training test!
⚡ ~
```

Figure 5.9: Screenshot of Resultant Image Loss Metrics

As shown in Table 5.1 and Table 5.2, our method demonstrates strong performance across all loss metrics. In Table 5.1, the Loss G (4.7528) and Loss D (0.2750) values highlight the stability of the model, while the L1 Cloth Loss (0.0229) and VGG Loss (0.4732) indicate accurate garment warping and high perceptual quality.

Table 5.1: Loss Metrics after Warping Garment Using C-GAN

Metric	Value
Loss G	4.7528
L1 Cloth Loss	0.0229
VGG Loss	0.4732
TV Loss	0.6170
Cross-Entropy (CE)	0.0763
Loss D	0.2750
D Real	0.2626
D Fake	0.0124

In Table 5.2, the Reconstruction Loss (0.0102) and L1 Alignment Loss (0.0047) reflect excellent alignment and minimal reconstruction error. Additionally, the Perceptual Loss (VGG) (0.2134) and TV Smoothness Loss (0.3125) emphasize the realism and smoothness of the generated images.

Table 5.2: Loss Metrics for Resultant Image using Diffusion Model

Metric	Value
Reconstruction Loss	0.0102
L1 Alignment Loss	0.0047
Perceptual Loss (VGG)	0.2134
TV Smoothness Loss	0.3125
Latent Space Consistency Loss	0.0291
Warped Consistency Loss	0.0213
Identity Loss	0.0129
Diffusion Regularization Loss	0.0458

The low Latent Space Consistency Loss (0.0291) and Warped Consistency Loss (0.0213) demonstrate stable latent space features and accurate garment warping. The Identity Loss (0.0129) further ensures that identity features are preserved, maintaining a high level of consistency with the ground truth.

These results validate the contributions of our cascade feature extraction module and skip connection complementary module, demonstrating the effectiveness of our approach in generating high-quality, realistic virtual try-on outputs.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

This report introduces K-VTON, a novel approach to virtual try-on technology that addresses key challenges in ensuring consistency and fidelity under complex poses and high-resolution scenarios. By integrating stable diffusion models with a cascaded feature extraction UNet network, our method effectively captures and preserves intricate garment features, ensuring seamless integration with target human poses. Evaluation based on key loss metrics such as Reconstruction Loss, L1 Alignment Loss, Perceptual Loss (VGG), TV Smoothness Loss, and others demonstrates the superior performance of our method. Specifically, our approach achieves lower reconstruction and alignment losses, indicating better garment fitting and alignment with the target pose. The Perceptual Loss (VGG) and TV Smoothness Loss further highlight enhanced image realism, preserving fine details and textures even in challenging poses. The Latent Space Consistency Loss and Warped Consistency Loss ensure that the generated images maintain coherence with the original input, adapting seamlessly to the target pose. Overall, K-VTON shows significant improvements in image fidelity and consistency, establishing it as a reliable and high-quality solution for virtual try-on applications in the fashion industry.

6.2 FUTURE WORK

While K-VTON has shown impressive capabilities in virtual try-on applications, several promising directions for future research and development exist. One key area is enhancing the model’s ability to handle extreme pose variations and complex garment designs through advanced pose-aware feature extraction mechanisms that adapt to diverse human poses while preserving garment integrity. Additionally, optimizing computational efficiency and resource utilization is crucial; developing lightweight architectures and exploring model compression techniques could improve accessibility for real-world deployment. Another significant avenue is integrating temporal consistency to enable dynamic virtual try-on applications, such as real-time fashion shows, which would require modules for temporal feature correlation and motion consistency. Lastly, expanding the model to manage multiple garment interactions and layered clothing presents a challenging yet promising direction, potentially involving sophisticated garment interaction models and advanced material simulation techniques to achieve more realistic virtual try-on results that reflect the physical characteristics of various fabrics and constructions.

REFERENCES

[1] Chenglin Zhou, Wei Zhang, and Zhichao Lian. Enhancing consistency in virtual try-on: A novel diffusion-based approach. *Image and Vision Computing*, pages 1050–1097, 2024.

[2] C.W. Hsieh, C.Y. Chen, C.L. Chou, H.H. Shuai, J. Liu, and W.H. Cheng. Fashionon: semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference On Multimedia*, pages 275–283, October 2019.

[3] N. Pandey and A. Savakis. Poly-gan: multi-conditioned gan for fashion synthesis. *Neurocomputing*, 414:356–364, 2020.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[5] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.

[6] M.R. Minar, T.T. Tuan, H. Ahn, P. Rosin, and Y.K. Lai. Cp-vton+: clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops Vol. 3*, pages 10–14, June 2020.

[7] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022.

[8] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.

[11] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[13] X. Han, X. Hu, W. Huang, and M.R. Scott. Clothflow: a flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–104 80, 2019.

[14] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, and J. Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035, 2019.

[15] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.

[16] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2021.

[17] S. He, Y.Z. Song, and T. Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[18] Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, and X. Liang. Gp-vton: towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023.

[19] R. Yu, X. Wang, and X. Xie. Vtnfp: an image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10511–10520, 2019.

[20] A.H. Raffiee and M. Sollami. Garmentgan: Photo-realistic adversarial fashion transfer. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3923–3930, January 2021.

[21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[22] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[24] S. Choi, S. Park, M. Lee, and J. Choo. Viton-hd: high-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021.

[25] S. Jandial, A. Chopra, K. Ayush, M. Hemani, B. Krishnamurthy, and A. Halwai. Sievenet: a unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2182–2190, 2020.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.