RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Phase II Report On

POSGEN360: Virtual Try-On

Submitted in fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Business Systems

By

Celina Elizabeth Jacob (U2109021)
Aadarsh Suresh (U2109001)
Amal Thomas (U2109008)
Hathik H (U2109028)

Under the guidance of

Ms. Ancy C A

Computer Science and Business Systems
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
November 2024

CERTIFICATE

This is to certify that the project report entitled ***"POSGEN360: Virtual Try-On"*** *is a bonafide record of the work done by* ***Celina Elizabeth Jacob (U2109021),Aadarsh Suresh (U2109001),Amal Thomas (U2109008),Hathik H(U2109028)****, submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Business Systems during the academic year 2024-2025.*

Ms. Ancy C A	Dr. Nikhila T Bhuvan	Dr. Divya James
Project Guide	Project Coordinator	Head of the Department
Assistant Professor	Associate Professor	Associate Professor
Department of CSBS	Department of CSBS	Department of CSBS
RSET	RSET	RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Rev. Dr. Jaison Paul Mulerikkal CMI**, Principal of RSET, and **Dr. Divya James**, Head of the Department of Computer Science and Business Systems for providing us with the opportunity to undertake our project,**"POSGEN360: Virtual Try-On"**.

We are highly indebted to our project coordinator, **Dr. Nikhila T Bhuvan**, Associate Professor, Department of Computer Science and Business Systems, for her valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Ms. Ancy C A** for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express my sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Celina Elizabeth Jacob

Aadarsh Suresh

Amal Thomas

Hathik H

Abstract

In today's fashion and e-commerce landscape, the integration of digital and physical experiences is becoming increasingly important. This project seeks to develop a system that transforms a single static image into a dynamic 360-degree video, allowing users to virtually try on clothes in a highly realistic manner. Built using Python, our approach combines advanced machine learning techniques, such as semantic correspondence models from StableVITON and image animation models from MagicAnimate, to seamlessly generate these virtual try-on experiences. The commercial potential of this technology is immense. By enabling customers to see how garments move and fit in a lifelike video, e-commerce platforms can significantly reduce return rates and enhance the online shopping experience. Additionally, this technology has the potential to revolutionize virtual fashion shows and online fitting rooms, offering a new level of interactivity and engagement. To achieve this, we face several challenges. These include generating accurate postures from static images, ensuring that the video frames are smoothly and consistently animated, and making sure that the clothes adapt naturally to the user's movements. We are addressing these challenges by using a combination of U-Net models (label each pixel in an image with a specific category) for image processing, diffusion models (gradually transforming a simple, easy-to-sample distribution [like Gaussian noise] into a complex data distribution [like an image] through a series of steps) for realistic animation, and frame-by- frame reconstruction techniques to create the videos. This process demands powerful GPUs for real-time processing and access to large datasets of human poses and garments for model training. By overcoming these technical challenges, we aim to set a new standard in virtual try-on technology, providing users with a more immersive, interactive, and lifelike experience.

Contents

List of Abbreviations

AR - Augmented Reality

CNN - Convolutional Neural Network

cGAN - Conditional Generative Adversarial Network

GAN - Generative Adversarial Network

IDE - Integrated Development Environment

SMPL - Skinned Multi-Person Linear

TPS - Thin-Plate Spline

UML - Unified Modeling Language

U-Net - Universal Network

ST-VTON - Style Transfer Virtual Try-On

List of Figures

List of Tables

Chapter 1

Introduction

This project aims to create an immersive virtual try-on platform that transforms static images of clothing into dynamic, 360-degree video representations. By enabling realistic try-ons that show garment movement and fit, the platform addresses key challenges in online fashion retail, including sizing uncertainty, high return rates, and customer engagement.

1.1 Background and Significance

In today's fashion and e-commerce landscape, bridging the gap between digital and physical experiences is essential. While online shopping offers convenience, it lacks the tactile and visual elements of in-store experiences, leading to high return rates and customer dissatisfaction due to uncertainties about sizing and fit.

Traditional product photos and static 3D models fail to convey how clothing drapes, moves, and fits, leaving consumers unsure about their purchases. This project develops an advanced virtual try-on system that converts static clothing images into dynamic, 360-degree videos. Using machine learning techniques like semantic correspondence models and image animation frameworks, the project aims to provide an interactive experience that demonstrates how garments move and fit in real time.

By enhancing customer engagement and boosting purchase confidence, this technology can reduce return rates in online fashion retail. Additionally, it can be applied in digital fashion shows, virtual fitting rooms, and immersive marketing, marking a significant advancement in the digitalization of fashion. This project promises to set a new industry standard for online shopping, aligning with the evolving needs of modern consumers.

1.2 Problem Definition

Current online fashion platforms fail to provide customers with a realistic and interactive way to visualize how clothing will fit and move on their bodies. This limitation creates uncertainty in purchasing decisions, lowers customer confidence, and contributes to high return rates.

1.3 Scope and Motivation

The scope of this project encompasses the development of an advanced virtual try-on platform that transforms static clothing images into dynamic, 360-degree video representations. It includes the implementation of machine learning techniques for semantic correspondence and image animation, enabling realistic garment visualization. The platform will cater to various applications within the online fashion retail sector, focusing on enhancing customer engagement and reducing return rates. Additionally, it will explore integration with digital fashion shows and virtual fitting rooms to create immersive shopping experiences. Overall, the project aims to set a new standard for virtual try-ons, addressing the evolving demands of the fashion industry.

The motivation behind this project stems from the growing challenges faced by online retailers in providing satisfactory shopping experiences. High return rates and customer dissatisfaction due to sizing uncertainty underscore the need for innovative solutions in the e-commerce landscape. By bridging the gap between online and in-store shopping experiences, this project seeks to enhance consumer confidence in their purchases. Furthermore, the shift towards digitalization in fashion emphasizes the importance of creating interactive and engaging platforms. Ultimately, this project aims to redefine online fashion retail, making it more appealing and effective for modern consumers.

1.4 Objectives

- Create an advanced platform that converts static clothing images into dynamic, 360-degree video representations to enhance the online shopping experience.
- Utilize semantic correspondence models and image animation frameworks to accurately depict garment movement and fit in real time.
- Increase user interaction and satisfaction by providing a realistic and immersive try-on experience that mimics in-store shopping.
- Minimize product returns by offering consumers a better understanding of how garments will fit and move before purchase.
- Investigate the use of the virtual try-on platform in digital fashion shows, virtual fitting rooms, and immersive marketing campaigns.
- Establish benchmarks for quality and interactivity in virtual try-on technologies to meet the evolving needs of modern consumers in the fashion industry.

1.5 Challenges

The project faces several challenges, including ensuring accurate garment representation during the transformation of static images to dynamic videos and addressing variations in user input for personalized fit. Additionally, mitigating potential issues related to internet connectivity and processing times for video rendering is crucial to maintain a smooth user experience. Establishing a reliable feedback mechanism for continuous improvement and addressing potential integration difficulties with existing e-commerce platforms will also be key to the project's success.

1.6 Assumptions

We assume that users will provide accurate measurements and preferences for effective garment fitting, and that stable internet connectivity will support seamless real-time interactions. Additionally, we expect the system to be compatible across various devices, ensuring accessibility for all users. Lastly, we anticipate that positive user adoption of the new features will enhance engagement and satisfaction with the platform.

1.7 Societal / Industrial Relevance

The societal and industrial relevance of this project is profound, as it addresses critical challenges in the evolving landscape of online fashion shopping. As e-commerce continues to expand, consumers increasingly seek reliable and engaging shopping experiences that replicate the tactile elements of in-store browsing. By offering a realistic virtual try-on platform that allows users to visualize how garments fit and move, this project significantly enhances consumer confidence in their purchasing decisions, ultimately leading to reduced return rates. This not only benefits retailers economically but also contributes to more sustainable fashion practices by minimizing waste associated with returns.

From an industrial perspective, integrating this technology into the fashion retail ecosystem represents a substantial advancement in customer interaction. Utilizing machine learning and advanced visualization techniques enables retailers to deliver personalized experiences that resonate with consumer preferences, fostering stronger brand loyalty. Moreover, the application of the virtual try-on system extends beyond traditional retail; it can be utilized in digital fashion shows, virtual fitting rooms, and immersive marketing campaigns, redefining how fashion brands engage their audiences and creating dynamic, interactive environments that captivate consumers.

Furthermore, as the fashion industry faces increasing scrutiny regarding sustainability and ethical practices, this project supports a more responsible consumption model. By enabling informed purchasing decisions and reducing the likelihood of ill-fitting garments, the platform encourages mindful shopping behavior. Overall, this project aligns technological innovation with consumer needs and industry demands, paving the way for a more connected, informed, and responsible fashion retail landscape that enhances customer satisfaction while addressing key industry challenges.

Chapter 2

Literature Survey

[1] K. Sun, P. Zhang, J. Zhang, and J. Tao, "DGM-Flow: Appearance flow estimation for virtual try-on via dynamic graph matching," Knowledge-Based Systems, vol. 302, pp. 112377, 2024. DOI: 10.1016/j.knosys.2024.112377

The paper introduces DGM-Flow, a method for enhancing appearance flow estimation in virtual try-on applications. The goal is to accurately transfer garments onto user images, accounting for garment flexibility, structure, and detailed texture alignment. Traditional methods often suffer from low accuracy in cases of complex deformations, occlusions, or alignment issues due to limitations in feature matching. DGM-Flow addresses these issues by leveraging dynamic graph matching to better capture the geometric and structural relationships between garment and body features. This advanced modeling enables a more realistic and immersive virtual try-on experience, particularly in complex scenarios.

Comparison with Previous Work: DGM-Flow is compared to prior approaches like TPS-based VTON and CP-VTON, which use thin-plate spline transformations but struggle with accurately modeling flexible garment deformations. More recent methods, such as ClothFlow and PF-AFN, utilize appearance flows for garment warping but are limited by noise interference and inaccurate feature matching in complex scenarios. DGM-Flow advances upon these by incorporating multi-layer graph convolution and cross-graph matching, which reduce mismatched features and improve alignment accuracy, outperforming existing methods in both standard and high-resolution fitting conditions.

Dataset: The experiments utilize the VITON and VITON-HD datasets, commonly used in virtual try-on research. VITON contains 14,221 training image pairs, while VITON-HD consists of high-resolution images, suitable for complex and detailed gar-

ment fitting tasks. Both datasets include images with occlusions, intricate textures, and large deformations, testing the model's robustness in real-world scenarios.

Model Selection: The DGM-Flow model is selected for its ability to dynamically match garment features to body features using a multi-order graph neural network (GNN) architecture. Unlike traditional methods that compare static features, DGM-Flow's dynamic approach allows it to adapt to varying textures and complex pose transformations, essential for high-resolution applications. This model also integrates a global graph matching module (GGM) for large-scale deformations and a local matching module (LM) for fine detail adjustments.

Methods and Steps:

- Preprocessing: Human pose and garment features are extracted and represented as pyramid feature maps, capturing both global and detailed garment structures.
- Graph Matching: Global Graph Matching (GGM): Computes a structural similarity matrix between garment and body features using dynamic graph embedding, accounting for higher-order geometric relationships.
- Cross-Graph Matching: Identifies mismatched features to refine the similarity matrix, improving alignment by filtering irrelevant features.
- Training: The model is trained using an L1 loss for pixel accuracy, perceptual loss to enhance visual realism, and smoothing constraints to avoid artifacts in garment flow.
- Virtual Try-On Module: Adopts a Res-UNet framework to integrate warped garments onto user images, refining alignment and structure for a final try-on result.

Results: The DGM-Flow model achieves superior results on both VITON and VITON-HD datasets, especially in handling complex deformations and occlusions. Quantitative metrics (SSIM, PSNR, LPIPS, FID) confirm improved structural fidelity, realism, and accuracy over baseline models. DGM-Flow excels in retaining garment details and aligning high-resolution textures, with notable performance in complex scenarios where other

models falter.

Advantages

- Enhanced Accuracy: Dynamic graph matching allows for precise garment alignment, even under complex deformations and occlusions.
- Realism: Improved texture handling and semantic structure preservation offer a more realistic virtual try-on experience.
- High-Resolution Compatibility: The model performs well on high-resolution images, making it practical for commercial applications.

Disadvantages

- Computational Complexity: Multi-layer graph convolution and dynamic matching increase computational costs compared to simpler models.
- Background Interference: The model may struggle with background noise in images, affecting garment and body segmentation accuracy.
- Layering Limitations: Current design has difficulty accurately simulating the sequential layering of garments, which could impact scenarios requiring multiple layers.

[2] R. Velastegui, M. Tatarchenko, S. Karaoglu, and T. Gevers, "Image Semantic Segmentation of Indoor Scenes: A Survey," Computer Vision and Image Understanding, vol. 248, p. 104102, 2024. DOI: 10.1016/j.cviu.2024.104102.

This survey explores the landscape of semantic segmentation for indoor scenes, with an emphasis on deep learning-based architectures. The focus is on evaluating models not only by traditional accuracy metrics but also by assessing temporal consistency and robustness to image corruption. Unlike previous surveys, which primarily cover outdoor environments, this survey is dedicated to the challenges unique to indoor settings, such as cluttered scenes, noise, blur, and camera movement. This work aims to fill a gap in segmentation research, providing a comparative analysis of around 50 architectures under

various challenging conditions. The study offers new insights into developing segmentation models optimized for complex indoor environments.

Comparison with Previous Work: Previous surveys primarily focus on accuracy, which is insufficient for real-world applications involving videos, where temporal consistency is crucial. Other studies that assess segmentation under challenging conditions like noise or weather largely focus on outdoor datasets. This survey differs by exclusively examining indoor scenarios, where environmental factors and complex object layouts present unique difficulties. Additionally, this work evaluates models in controlled, consistent settings, improving comparability across methods and providing a more comprehensive analysis of robustness against real-world challenges.

Dataset: The survey utilizes both real-world and synthetic datasets for indoor scene segmentation:

- Real-World Dataset: Scannet, comprising over 2.5 million frames from diverse indoor settings, offers RGB images, segmentation masks, camera poses, and depth information. Its natural variability and real-world disturbances, such as noise and blur, present realistic challenges for model testing.
- Synthetic Dataset: InteriorNet provides around 20 million synthetic images with pixel-perfect labels and consistent scene structure across frames. While it lacks real-world imperfections, its extensive labeling quality and control over conditions make it ideal for systematic training and evaluation.

Model Selection: The survey evaluates approximately 50 segmentation models, spanning traditional fully convolutional architectures like FCN, pyramid-based models like PSPNet, and modern attention-based approaches such as SegFormer and FANH. The models vary by backbone architecture and complexity, allowing for a diverse analysis of segmentation accuracy, temporal consistency, and robustness against corruption. The inclusion of both classic and transformer-based models highlights advancements and variations in performance across different architecture types.

Methods and Steps:

- Preprocessing: Image inputs are preprocessed with pixel-level annotations and structured into hierarchical feature maps, accommodating the model requirements for both RGB and depth information.
- Model Training and Evaluation:
  - Training: All models are trained under consistent conditions using the MMSegmentation framework, with hyperparameters like batch size, learning rate, and optimizers standardized.
  - Testing: Models are evaluated on metrics for accuracy (IoU, pixel accuracy, DICE coefficient) and temporal consistency in segmentation over sequential frames.
- Evaluation Metrics:
  - Segmentation Accuracy: Traditional metrics like mIoU and pixel accuracy assess pixel-level classification.
  - Temporal Consistency: Measured between consecutive frames to ensure stable segmentation outputs over time.
  - Corruption Vulnerability: Robustness to image noise, blur, and distortions is assessed by applying perturbations to test images and measuring accuracy changes.

Results: The survey shows that attention-based and transformer models, such as FANH and SegFormer, achieve the highest segmentation accuracy and robustness against corruption, particularly for complex indoor scenes. Convolution-based models tend to exhibit better temporal consistency in frame-by-frame video segmentation but may be less resilient to noise and other corruptions. In scenarios involving high object entropy or extensive camera movement, the robustness of transformer models becomes especially evident, outperforming traditional methods in both synthetic and real-world datasets.

Advantages:

- High Accuracy and Robustness: Transformer-based models show enhanced performance under complex and corrupted conditions, making them suitable for real-world indoor applications.
- Consistent Evaluation: Standardized settings across models ensure fair comparisons, providing a reliable benchmark for future segmentation studies.
- Comprehensive Metric Coverage: The inclusion of temporal consistency and corruption vulnerability metrics allows a thorough assessment beyond traditional accuracy measures.

Disadvantages:

- Computational Demands: Transformer-based models, while robust, have high computational and memory requirements, which can limit their deployment in real-time applications.
- Synthetic vs. Real-World Domain Gap: Models trained on synthetic datasets may not generalize fully to real-world scenarios due to differences in environmental complexity and imperfections.
- Limited Adaptability to Dynamic Objects: Frame-by-frame segmentation methods may struggle with dynamic objects in video, as temporal consistency is measured in static settings.

[3] T. Islam, A. Miron, X. Liu, and Y. Li, "StyleVTON: A Multi-Pose Virtual Try-On with Identity and Clothing Detail Preservation," Neurocomputing, vol. 594, pp. 127887, 2024. DOI: 10.1016/j.neucom.2024.127887

The paper presents StyleVTON, a novel virtual try-on system designed for multi-pose applications, enhancing both garment fitting and the preservation of user identity in a virtual environment. Unlike traditional 2D try-on models, which are limited by fixed postures, StyleVTON enables pose transfer, allowing users to view garments from multiple perspectives. The approach combines three modules—segmentation, warping, and

pose transfer—to accurately fit clothing onto users, even with varying poses. This study highlights StyleVTON's potential to improve online shopping experiences by providing consumers with a more realistic and immersive try-on experience across different poses.

Comparison with Previous Work: Prior virtual try-on models, such as VITON, CP-VTON, and other GAN-based approaches, mainly focus on single-pose try-ons and often fail to preserve user identity when transferring poses. StyleVTON builds on these foundations by utilizing StyleGAN architecture for enhanced pose transfer while maintaining visual fidelity, especially around facial features. Unlike previous models that suffer from loss of detail and inaccurate garment fitting during complex poses, StyleVTON incorporates segmentation and dual discriminators to refine garment fitting and detail preservation, outperforming previous methods in terms of identity retention and multi-pose realism.

Method	**Dataset Used**	**Key Features**	**Limitations**
CP-VTON	VITON	Thin-Plate Spline (TPS) for warping	Limited to single pose
DGM-Flow	VITON-HD	Dynamic Graph Matching (GNN)	Computationally complex
StyleVTON	DeepFashion	Pose Transfer, Identity Preservation	Requires large training data
FVTN	VITON	Flow-Based Deformations	Background interference

Table 2.1: Comparison of StyleVTON with other Virtual Try-On Methods

Dataset:

- VITON-HD: This dataset includes 11,647 images of frontal-view models paired with clothing images, used primarily for training the segmentation and warping modules.
- DeepFashion: Consisting of 101,967 image pairs of models in various poses, this dataset supports the training of the pose transfer module, allowing StyleVTON to learn complex pose variations effectively.

Model Selection: StyleVTON was selected based on its capacity to handle both garment fitting and pose transfer simultaneously. It integrates a U-Net-based segmentation module, an STN-based warping module, and a StyleGAN-based pose transfer module, each optimized for different aspects of the try-on process. The choice of these modules

enables StyleVTON to retain garment detail, identity, and multi-pose compatibility, which traditional try-on models typically lack.

Methods and Steps:

- Segmentation Module: Uses U-Net to generate segmented labels for body parts, based on the target garment and user pose, ensuring correct spatial garment positioning.
- Warping Module: Implements a spatial transformer network (STN) followed by U-Net refinement to deform the garment according to the segmentation layout, allowing for accurate garment fitting.
- Pose Transfer Module: Employs StyleGAN and UV mapping to transfer the candidate's body parts into the desired posture, preserving identity and clothing detail. This module uses a coordinate completion model to ensure smooth appearance transfer across poses.
- Training: Each module is trained individually with specific loss functions (L1, GAN, and perceptual losses) to enhance visual realism, alignment, and detail retention across poses.

Results: StyleVTON demonstrates superior performance on metrics such as SSIM, FID, IS, and LPIPS across multiple datasets, including VITON-HD and Fashiontryon, when compared to baseline methods. The model effectively retains facial details, aligns clothing accurately, and provides realistic multi-pose images, outperforming other approaches, especially in complex poses. Qualitative results show that StyleVTON's images are visually superior, with fewer artifacts and enhanced garment and identity preservation.

Advantages:

- Multi-Pose Compatibility: Enables pose transfer, allowing users to view garments across different poses.

- Enhanced Detail Preservation: Dual discriminators and StyleGAN ensure realistic garment fitting and identity retention.
- Broad Applicability: Applicable in online fashion retail for enhanced user experience, offering a practical alternative to 3D try-on systems.

Disadvantages:

- Computational Complexity: The use of multiple modules and discriminators increases the computational cost.
- Segmentation Inaccuracy: Occasional segmentation errors can lead to garment misalignment, impacting final output quality.
- Dataset Bias: Limited representation of diverse body types in training data may affect generalizability to varied user shapes.

[4] T. Wang, X. Gu, and J. Zhu, "A Flow-Based Generative Network for Photo-Realistic Virtual Try-On," IEEE Access, vol. 10, pp. 40899-40908, 2022. DOI: 10.1109/ACCESS.2022.3167509

The paper introduces the Flow-Based Virtual Try-On Network (FVTN), designed to create photo-realistic virtual try-on images by transferring clothing onto target person images with enhanced accuracy. Addressing limitations in existing methods, FVTN tackles challenges such as occlusion, spatial deformation, and feature misalignment through a flow-based deformation approach. Unlike traditional models that struggle with non-rigid clothing transformations and integration of human features, FVTN's multi-module architecture allows it to retain garment texture, color, and style while adapting clothing shapes to different poses. The proposed model improves the virtual try-on experience for online shopping by synthesizing high-quality images that align realistically with human bodies.

Comparison with Previous Work: Previous methods like VITON, CP-VTON, and ClothFlow often rely on affine or thin-plate spline transformations, which fail to accurately represent non-rigid clothing deformations. While ClothFlow introduced flow-based

deformations, it lacked unsupervised learning capabilities and struggled to handle complex spatial adjustments. FVTN advances beyond these methods by employing unsupervised multi-scale dense flow fields, offering superior garment fitting and human-body integration. Unlike approaches that use composition masks, FVTN's flow-based design preserves seamless garment-body transitions and mitigates boundary artifacts in synthesized images.

Dataset: The experiments are conducted on the VITON dataset, a popular dataset for virtual try-on tasks. The dataset contains 16,253 image pairs, including front-view images of women paired with clothing items, with images resized to 256×192 resolution. The data is divided into 14,221 pairs for training and 2,032 pairs for validation. For evaluation, the validation set images are rearranged into mismatched pairs to test the model's try-on capabilities.

Model Selection: FVTN was chosen for its ability to address multi-level challenges in virtual try-on, integrating three modules tailored for specific tasks: Parsing Alignment Module (PAM), Flow Estimation Module (FEM), and Fusion and Rendering Module (FRM). PAM provides semantic alignment by parsing and positioning garments on the target body, FEM employs a dense flow-based deformation model, and FRM synthesizes the final image by combining warped clothing and body features. These modules work together to capture complex deformations, ensuring accurate garment positioning and realistic image rendering.

Methods and Steps:

- Parsing Alignment Module (PAM): PAM uses a conditional generative adversarial network with U-Net to create a semantic parsing map, positioning the garment on the target person by aligning body parts and garment shapes.
- Flow Estimation Module (FEM): FEM performs unsupervised flow-based deformation by estimating multi-scale dense flow fields using a feature pyramid network and deformable convolution, enabling accurate correspondence between garment and body.

- Fusion and Rendering Module (FRM): FRM fuses the segmented clothing and body features using encoders for body, clothing, and head parts to generate the final try-on image. It employs an adaptive masking approach to synthesize realistic human features and garment details.
- Training: Each module is trained separately using the Adam optimizer with appropriate loss functions (cross-entropy, photometric, total variation, and perceptual losses) to maximize spatial alignment and visual fidelity.

Results: FVTN outperforms baseline models on quantitative metrics (SSIM, IS, FID, and LPIPS) and achieves high scores in human evaluations, demonstrating superior visual realism and garment detail preservation. Compared to existing models, FVTN maintains garment texture, fits clothing to varied poses, and reduces visual artifacts. Qualitative results illustrate that FVTN produces more seamless garment-body transitions, handling complex postures and occlusions effectively.

Advantages:

- Enhanced Garment Deformation: The flow-based design captures non-rigid clothing transformations, allowing FVTN to produce realistic results across diverse poses.
- High-Quality Image Synthesis: FRM's fusion of body and garment features provides seamless integration without visible artifacts, improving the overall image realism.
- Efficient Training and Inference: Unsupervised flow estimation and end-to-end training enable effective handling of complex deformations with minimal annotation requirements.

Disadvantages:

- Reliance on Accurate Segmentation: The model depends on accurate human segmentation, and any missegmentation can lead to unrealistic results.
- Computational Demands: FVTN's multi-module structure, especially flow estimation, requires significant computational resources, impacting its scalability.

- Pose Limitations: FVTN struggles with uncommon poses and certain viewpoint transformations, leading to potential artifacts in challenging scenarios.

[5] H.-J. Lee, B. Koo, H.-E. Ahn, M. Kang, R. Lee, and G. Park, "Full Body Virtual Try-On With Semi-Self-Supervised Learning," Electronics Letters, vol. 57, no. 24, pp. 915–917, 2021. DOI: 10.1049/ell2.12307

This paper proposes the Full Body Virtual Try-On (FB-VTON) system, a method for generating realistic full-body virtual try-on images that handle both top and bottom garments simultaneously. Unlike previous models limited to single garments, FB-VTON is designed to align and fit multiple garments on a full-body model. The system addresses the challenge of limited training data for multi-garment alignment by introducing two innovative training strategies, allowing FB-VTON to generalize across diverse outfits. Consisting of three modules—Clothing Guide Module (CGM), Geometric Matching Module (GMM), and Try-On Module (TOM)—the proposed system synthesizes photo-realistic try-on images that preserve garment details such as logos, patterns, and fit, achieving significant improvements in realism and garment alignment.

Comparison with Previous Work: Prior works, including CP-VTON and LA-VITON, have focused on virtual try-ons for single garments using transformation methods like thin-plate spline (TPS) and geometric matching. However, these approaches struggle with maintaining garment characteristics when applied to multiple garments. FB-VTON introduces a multi-garment system using CGM to predict a clothing guide map (CGMap) for top and bottom garments, an approach that overcomes the data limitations of previous models. Unlike Neuberger et al., who required additional optimization to align multiple garments, FB-VTON's semi-self-supervised training leverages existing datasets more effectively, providing a practical solution for realistic multi-garment virtual try-on.

Dataset: The dataset comprises 2,813 model-top pairs from the MPV dataset and 4,135 model-bottom pairs collected from online sources, with each image sized at 192×256. The dataset lacks paired images of models wearing both top and bottom garments, posing a challenge for full-body virtual try-on training. FB-VTON addresses this by generat-

ing pseudo-triplets of top-bottom pairs to simulate full-body images, ensuring the model learns to handle multiple garments.

Model Selection: FB-VTON consists of three primary modules—CGM, GMM, and TOM. CGM predicts a clothing guide map (CGMap) to describe the garment shape on a model, which guides the alignment in subsequent modules. GMM performs geometric matching for both top and bottom garments, while TOM synthesizes the final try-on image by integrating the aligned garments with the CGMap. The structure of FB-VTON enables precise garment alignment and natural blending in try-on images, handling complex full-body outfits better than previous models.

Methods and Steps:

- PClothing Guide Module (CGM): CGM generates a CGMap that outlines the top and bottom garments on a model using self-supervised learning. It leverages two training strategies: creating pseudo-triplet data by pairing model images with randomly transformed garment sections, and exposing CGM to varied in-shop garment combinations to reduce training bias.
- Geometric Matching Module (GMM): GMM aligns the top and bottom garments separately using two self-attention-based networks, each trained to warp garments according to the body's pose and shape as defined in CGMap.
- Try-On Module (TOM): TOM synthesizes the final try-on image by combining warped garments with the person's features. The module includes two generators, one inspired by SPADE, to balance garment details and model features seamlessly.
- Training: CGM, GMM, and TOM are trained for 100 epochs, using cross-entropy and GAN losses to refine garment shape, position, and appearance.

Results: FB-VTON achieves high-quality results, outperforming previous methods in terms of garment alignment, detail preservation, and visual realism. Quantitative evaluations show that FB-VTON is preferred in user studies, with 83.8% of responses favoring its images over those produced by LA-VITON. The qualitative results display well-aligned

outfits that maintain natural proportions, realistic garment boundaries, and high visual fidelity even with varied garment styles.

Advantages:

- Realistic Full-Body Outfits: FB-VTON effectively aligns both top and bottom garments, providing a complete try-on experience.
- Innovative Training Strategy: The semi-self-supervised approach overcomes data limitations, allowing the model to generalize across garment types without extensive labeled datasets.
- Enhanced Detail Preservation: The use of CGMap enables FB-VTON to preserve garment-specific features like logos, textures, and proportions.

Disadvantages:

- Dependence on CGMap Accuracy: The system's reliance on CGMap means that inaccuracies in CGMap generation can affect garment alignment and final image quality.
- Computational Overhead: FB-VTON requires extensive computational resources, particularly for GMM's separate self-attention-based networks for top and bottom garments.
- Limited Tucking Capability: The model does not yet handle garment tucking, which could further enhance realism for specific outfits.

[6] J. Lee, M. Lee, and Y. Kim, "MT-VTON: A Generative Approach to Virtual Try-On Using Multi-Level Feature Transformation," Applied Sciences, vol. 13, no. 11724, 2023. doi:10.3390/app13111724

This paper presents MT-VTON, a generative method designed to enhance virtual try-ons by accurately preserving both the overall context of clothing (like folds) and fine details (such as logos). MT-VTON improves the clothing warping process using a multi-level feature transformation approach, followed by pixel-level refinement for better image

quality.

Comparison with Previous Work: MT-VTON outperforms three baseline methods: ACGPN, Flow Style VTON, and HR-VITON. ACGPN uses an inpainting approach with STN for warping clothing, but struggles with finer details. Flow Style VTON leverages StyleGAN to warp clothing but has difficulty preserving logos. HR-VITON estimates flow fields and masks together, improving alignment but still lacking in fine details. In contrast, MT-VTON excels by preserving both contextual features and fine details like logos, providing a more realistic and accurate virtual try-on result.

Dataset: The dataset used for training and evaluation includes the AIHUB and VTON datasets. The AIHUB dataset is publicly available, providing a diverse set of images for virtual try-on tasks. The VTON dataset, accessible from GitHub, specifically focuses on virtual try-on images and serves as a crucial resource for training the model to handle various clothing types and poses. These datasets allow the model to be trained on real-world data, ensuring it performs effectively in diverse scenarios.

Model Selection: MT-VTON utilizes a generative model based on multi-level feature transformation. This method captures fine clothing features like folds and logos and integrates them into the human model using appearance flow. The approach includes refining the clothing fitting through pixel-level adjustments, ensuring better quality and realism.

Methods and Steps:

- Preprocessing: The dataset images (clothing and person) are preprocessed to extract features and align the clothing with the human model. This step involves detecting the clothing areas and preparing them for warping.
- Training: MT-VTON is trained from scratch using the preprocessing data, with a focus on preserving important features like logos and clothing context. The training process optimizes the model to ensure accurate clothing fitting onto the person image.
- Algorithm: The model uses a multi-level feature transformation technique to warp

clothing images. This is followed by pixel-level refinement, which adjusts the warped clothing and ensures it fits naturally on the human model.

- Flow Estimation: The method estimates the flow for warping the clothing image. This flow transformation helps match the clothing to the body image, preserving details like folds.
- Refinement: A refinement step is applied, merging the pixel flow to generate high-quality images that accurately integrate the clothing and human model.

Results: The proposed method outperforms previous approaches in most metrics, except for the Inception Score. It achieves a lower Fréchet Inception Distance (FID) and higher Inception Score (IS), suggesting better image quality and realism. The Learned Perceptual Image Patch Similarity (LPIPS) and Structural Similarity Index Measure (SSIM) metrics also show that the method produces more natural and visually appealing results. These improvements indicate that the approach successfully preserves both contextual features, such as cloth folding, and finer details like logos, which previous methods struggled with.

Advantages:

- Better Detail Preservation: MT-VTON excels in maintaining important clothing details like logos and folds, offering more realistic results.
- Natural Layering: Despite not using cloth-agnostic images during refinement, MT-VTON still generates realistic results with better clothing integration.
- High-Quality Results: The method produces high-resolution images with accurate clothing fitting.

Disadvantages:

- Layering Challenges: The approach still faces limitations in layering, as simply superimposing clothing onto the human model can sometimes lead to unrealistic results, particularly for inner clothes that are too large.

- Inception Score: The method did not perform as well in the Inception Score metric, indicating potential areas for improvement in generating realistic images according to this measure.

[7] Z. Muhammad, Z. Huang, and R. Khan, "A Review of 3D Human Body Pose Estimation and Mesh Recovery," Digital Signal Processing, vol. 128, 2022. [Online]. Available: https://doi.org/10.1016/j.dsp.2022.103628

3D human body pose estimation and mesh recovery aim to reconstruct accurate 3D representations of the human body from 2D input, such as single images or videos. This area is critical in virtual try-ons, AR/VR applications, gaming, and human-computer interaction. The reviewed paper categorizes advancements into two primary approaches: parametric models that rely on predefined templates like SMPL, and non-parametric methods that offer template-free reconstructions, emphasizing detail and flexibility. While recent developments have achieved significant progress in precision and adaptability, challenges persist in handling occlusion, complex motions, and variations in clothing, highlighting the need for further innovation.

Comparison with Previous Work: Early parametric methods like SMPL provided robust reconstructions by fitting 3D templates to 2D images. These methods excelled in computational efficiency but struggled with capturing intricate details such as clothing deformations or occluded body parts. On the other hand, non-parametric approaches like BodyNet and PIFu leverage deep learning to directly generate 3D meshes from input data. These newer methods avoid template constraints, enabling finer detail and more flexibility in pose and texture recovery. Despite their accuracy, non-parametric approaches often face higher computational demands and require large, diverse datasets to generalize effectively.

Dataset: The reviewed paper highlights the significance of high-quality datasets in advancing pose estimation and mesh recovery. Commonly used datasets include Hu-

man3.6M, 3DPW, and LSP. These datasets offer annotated images or videos with corresponding 3D pose information, enabling supervised training of models. However, many datasets have limitations, such as a lack of diverse clothing, environments, and motions, leading to performance gaps in real-world scenarios. Recent datasets such as AMASS and SURREAL incorporate richer details and varied contexts, addressing some of these challenges but leaving room for further improvement.

Model Selection: The review evaluates models based on their ability to balance efficiency, accuracy, and generalization:

- Parametric Models: These models, like SMPL and SCAPE, are ideal for real-time applications due to their lightweight nature but struggle in representing dynamic clothing or complex motions.
- Non-Parametric Models: Approaches like DeepHuman and PIFu utilize neural networks for dense reconstructions and provide higher flexibility but are computationally intensive.
- Hybrid Models: Emerging methods integrate parametric templates with neural networks, combining efficiency with detailed reconstructions.

Methods and Steps:

- Feature Extraction: Extracts key 2D landmarks and textures from images using convolutional neural networks (CNNs).
- Pose Estimation: Detects joint positions using keypoint-based approaches like OpenPose or DensePose.
- Mesh Generation: Converts 2D landmarks and textures into 3D representations, leveraging either parametric or non-parametric methods.
- Fine-tuning with Datasets: Utilizes labeled datasets to enhance model accuracy and generalization.

- Optimization Techniques: Employs perceptual loss and adversarial training for improved realism in mesh recovery.

Results: The reviewed advancements demonstrate significant improvements in the accuracy and versatility of 3D human body pose estimation and mesh recovery. State-of-the-art methods like PIFu and DenseBody achieve detailed reconstructions, accurately handling occlusion and diverse poses. However, challenges remain in real-time performance and generalization across varied environments, emphasizing the importance of continued dataset enrichment and optimization.

Advantages:

- High Accuracy: Modern approaches achieve precise reconstructions, even in complex poses and scenarios.
- Flexibility: Non-parametric methods excel in adapting to diverse environments, body types, and motions.
- Application Versatility: Enables applications in AR/VR, gaming, virtual try-ons, and healthcare.

Disadvantages:

- Computational Demands: High resource requirements limit real-time applications.
- Dataset Dependence: Performance relies on the availability and diversity of annotated datasets.
- Occlusion Challenges: Handling severely occluded body parts remains an ongoing issue.

[8] T. T. Tuan, M. R. Minar, H. Ahn, and J. Wainwright, "Multiple Pose Virtual Try-On Based on 3D Clothing Reconstruction," IEEE Access, vol. 9, pp. 114367–114380, 2021, doi: 10.1109/ACCESS.2021.3104274

The paper proposes a novel system for virtual try-ons, called Multiple Pose Virtual Try-On (MPVTON), that emphasizes accurate 3D clothing reconstruction and multi-pose compatibility. Traditional 2D virtual try-on systems, while capable of generating realistic outputs, often fail when confronted with complex poses or non-front-facing views due to their reliance on simple warping methods like Thin-Plate Splines (TPS). The proposed MPVTON system addresses these limitations by integrating 3D models of clothing and human poses, enabling more dynamic and realistic virtual try-on experiences. This approach provides the flexibility to handle complex body postures, multiple viewing angles, and detailed fabric textures.

Comparison with Previous Work:

- Limitations of 2D Systems: Traditional methods, such as CP-VTON and VITON, utilize 2D warping techniques like Thin-Plate Splines (TPS), which often result in unrealistic deformations, artifacts, and poor adaptability to non-frontal poses or intricate garment patterns.
- Existing 3D Approaches: CloTH-VTON+ incorporated 3D models for better deformation but lacked multi-pose capabilities, restricting its flexibility for varied user interactions.
- Advantages of MPVTON: By integrating 3D garment reconstruction with pose-aligned human models, MPVTON achieves superior realism and multi-view capabilities, outperforming previous works in generating lifelike clothing animations and addressing non-frontal pose scenarios effectively.

Dataset: PVTON utilizes the MPV dataset tailored for multi-pose virtual try-on tasks. The dataset comprises:

- 35,687 Human Images: Capturing varied poses and body shapes for robust training.
- 13,524 Garment Images: High-resolution in-shop clothing images, enabling detailed texture and pattern reconstruction.

- Preprocessing: The dataset underwent meticulous curation, including mask corrections and removal of redundant entries, ensuring consistency and quality for challenging pose variations.
- Training/Test Split: The data is divided into 52,236 training pairs and 10,544 test pairs, optimizing the model's generalization capabilities.

Model Selection: MPVTON relies on advanced components to achieve its objectives:

- SMPL Model: This parametric model accurately represents human pose and shape, aligning garments with users' body postures and ensuring fit accuracy.
- Pix2PixHD Framework: Enhances the visual realism of synthesized try-on images through high-resolution generative adversarial networks (GANs), preserving fine details in textures and patterns.

Methods and Steps:

- 3D Garment Reconstruction: Employs a 3D scanning module to capture the shape and texture of garments. Reconstructs the 3D geometry for precise fitting and deformation.
- Pose Estimation: Detects joint positions using keypoint-based approaches like OpenPose or DensePose.
- Pose Estimation: Uses SMPL-based models to extract pose parameters from input user images. Aligns body shape with reconstructed garments.
- Garment Warping: Implements non-rigid deformation techniques to adjust clothing geometry based on user pose.
- Texture Mapping: Maps high-resolution garment textures onto reconstructed 3D models, preserving intricate details like wrinkles and patterns.
- Image Synthesis: Combines the fitted 3D garment with user images using GAN-based synthesis for seamless integration.

Results: MPVTON demonstrates significant improvements over prior methods in terms of realism, pose adaptability, and garment fit accuracy. Quantitative evaluations reveal lower error rates and enhanced texture retention compared to baseline methods such as VITON and CP-VTON. Qualitative results highlight the model's ability to handle complex poses, dynamic movements, and intricate garment details. The system also showcases robustness in generating multi-view outputs, making it ideal for applications in fashion retail and virtual reality.

Advantages:

- Pose Adaptability: Dynamically adjusts garments to fit diverse user poses and viewing angles.
- High Realism: Achieves photorealistic outputs with accurate texture and pattern mapping.
- Multi-View Capability: Supports dynamic try-on experiences with multiple perspectives.
- Robustness: Effectively handles challenging scenarios, including complex poses and overlapping garments.

Disadvantages:

- Computational Complexity: High computational cost due to 3D reconstruction and GAN-based synthesis.
- Data Dependency: Requires high-quality datasets for training and performance optimization.
- Limited Real-Time Capability: Processing times may be unsuitable for real-time applications.

[9] Zhou, C., Zhang, W., Liu, X., & Li, Y. (2024). CSD-VTON: A novel method for enhancing virtual try-on image realism and consistency. Image and Vision Computing, 148, 105097. https://doi.org/10.1016/j.imavis.2024.105097

In this paper, Zhou et al. (2024) introduce CSD-VTON, an advanced virtual try-on model aimed at improving the realism and consistency of virtual try-on images, especially in scenarios involving complex poses and higher resolution images. The primary objective of the proposed model is to generate high-quality, realistic try-on images by leveraging a combination of diffusion models, skip connection modules, and a novel feature extraction approach using stable diffusion. This method outperforms existing techniques in preserving garment details and ensuring proper fit during pose changes.

Comparison with Previous Work: The paper compares CSD-VTON with other established virtual try-on methods such as CP-VTON, ACGPN, VITON-HD, and MGD using the VITON-HD and DressCode datasets. It demonstrates that CSD-VTON achieves superior performance across several evaluation metrics (LPIPS, SSIM, FID, KID), particularly in maintaining image realism and consistency compared to GAN-based approaches like CP-VTON and ACGPN, and even diffusion-based methods like MGD.

Dataset: MagicAnimate was evaluated on the TikTok dataset, a collection known for its diverse range of poses, identities, and complex backgrounds. The dataset includes various motion sequences and is used to assess the framework's ability to handle significant pose variations, identity diversity, and complex scenes. This dataset is particularly well-suited for evaluating the robustness of temporal coherence and the preservation of reference appearance throughout video sequences.

Model Selection:

- Warping Module: For garment image warping to align with the human body.
- TCascade Feature Extraction Module: Extracts detailed garment features for conditioning the generation process.
- Skip Connection Supplement Module: Aids in compensating for VAE reconstruction errors, improving image clarity and detail.
- Denoising Network: Focuses on refining the final generated images by reducing noise and improving visual quality.

Methods and Steps:

- Preprocessing: All garment images are resized to 256 × 192 pixels. For the final model, input images are resized to 512 × 384 pixels.
- Training:Trained using the Adam optimizer with a learning rate of 1e-4.Trained for 20 epochs with Adam optimizer. Trained with a learning rate of 1e-5 for 20k steps using AdamW. Trained for 400k steps, using pre-trained weights and stable diffusion as the base generative network. Training was performed on an A40 GPU.
- Algorithm: Cascaded UNet is used for feature extraction. Diffusion Model is the core generative network for virtual try-on image synthesis.

Results: On the VITON-HD dataset, CSD-VTON outperforms competing methods (CP-VTON, ACGPN, VITON-HD, and MGD) across all four metrics: LPIPS (0.106), KID (1.84), FID (10.44), SSIM (0.868). It showed substantial improvements over GAN-based methods and even diffusion-based methods like MGD. On the DressCode dataset, CSD-VTON similarly showed superior performance, especially in preserving garment details and improving image realism. Visual comparisons highlight that CSD-VTON produces more consistent, accurate, and detailed virtual try-on images, particularly in complex scenarios with intricate garment textures and poses.

Advantages:

- Improved Realism and Consistency: CSD-VTON excels in generating high-quality, realistic virtual try-on images with better consistency between the generated images and ground truth.
- Better Garment Fit: The warping module ensures that garments fit more accurately with the target human body, reducing issues of misalignment or poor fitting.
- Handling Complex Poses and Textures: The model preserves intricate garment features like stripes and lace, even under complex poses or for semi-transparent garments.

- Integration with Text-to-Image Models: The system's ability to animate images generated by text-to-image models (e.g., DALL·E3) expands its usability for diverse applications.
- Use of Diffusion Models: The adoption of stable diffusion as a base generative model provides enhanced image realism compared to traditional GAN-based methods.

Disadvantages:

- Training Time and Computational Resources: The training process is computationally intensive, requiring significant GPU resources and a lengthy training period (up to 400k steps).
- Dependency on Pre-trained Weights: The model requires pre-trained weights, particularly for the denoising network and feature extraction, which may limit flexibility in certain scenarios.
- Challenges with Fine-Tuning: While the model performs well across most scenarios, fine-tuning for specific garment types or complex scenarios might require additional training and adjustments.

[10] Saharia, C., et al. "MagicAnimate: High-Fidelity Human Avatar Animation with Temporal Consistency." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

MagicAnimate is a novel framework for generating high-fidelity human avatar animations with an emphasis on temporal consistency. The system leverages diffusion models to create realistic animations by modeling both appearance and temporal coherence effectively. Unlike traditional animation techniques, MagicAnimate introduces a temporal attention mechanism and a refined appearance encoder, aiming to preserve both single-frame fidelity and the smoothness of long-term video sequences. The system's architecture also incorporates image-video joint training and video frame fusion strategies, making it versatile and capable of handling various motion sequences and unseen domains. With

these innovations, MagicAnimate is poised to advance the state of avatar animation technology, offering seamless transitions and high-quality results for diverse animation tasks.

Comparison with Previous Work: MagicAnimate distinguishes itself from previous human animation systems by focusing on both appearance quality and temporal consistency. Previous works often neglected the challenges of maintaining coherence across frames in video sequences, resulting in animations that either lacked smooth transitions or produced inconsistent background details. Additionally, MagicAnimate's ability to handle multi-person animations and integrate with text-to-image generation (such as DALL·E3) sets it apart from prior models, which typically struggled with these tasks. While many models perform well for single-person animations or specific types of motion, MagicAnimate's generalized approach enables better handling of a wider range of styles and movements, including unseen domains.

Dataset: MagicAnimate was evaluated on the TikTok dataset, a collection known for its diverse range of poses, identities, and complex backgrounds. The dataset includes various motion sequences and is used to assess the framework's ability to handle significant pose variations, identity diversity, and complex scenes. This dataset is particularly well-suited for evaluating the robustness of temporal coherence and the preservation of reference appearance throughout video sequences.

Model Selection: The primary model used in MagicAnimate is based on a diffusion model architecture that integrates temporal attention mechanisms and appearance encoders to improve both the spatial and temporal fidelity of generated animations. The framework's design also includes a video fusion technique to improve transition smoothness across long-term animations and a novel approach to image-video joint training. These elements collectively help MagicAnimate achieve high-quality, consistent animations across diverse scenarios.

Methods and Steps:

- Preprocessing: The system begins by preparing the input frames, where the refer-

ence image is paired with a motion sequence. The input images are then processed to extract features related to both appearance and temporal information.

- Training: MagicAnimate uses an end-to-end training pipeline that includes both appearance encoding and temporal modeling. The system is trained using the TikTok dataset, with the model focusing on optimizing both single-frame quality and temporal consistency across video sequences. Key training components include temporal attention layers, appearance encoders, and joint image-video training strategies.
- Algorithm: The core algorithm involves applying a diffusion model with temporal attention and appearance encoding. This allows the system to model both the appearance of the reference image and the temporal dynamics of the motion sequence, ensuring high-quality animation output.

Results: MagicAnimate's performance was evaluated on the TikTok dataset, with results demonstrating clear improvements in both single-frame and video fidelity compared to prior models. The use of temporal attention layers, appearance encoders, and video fusion strategies contributed significantly to the system's success, as shown in the ablation studies. The framework also exhibited strong generalization ability, successfully generating animations for multi-person scenarios and unseen image styles, such as oil paintings and movie stills. Additionally, integrating MagicAnimate with text-to-image generation models (e.g., DALL·E3) demonstrated its flexibility in generating and animating content based on textual descriptions.

Advantages:

- High Temporal Consistency: MagicAnimate excels in preserving temporal coherence, which is essential for generating smooth, realistic animations.
- Improved Appearance Encoding: The appearance encoder enhances both single-frame and long-term video fidelity, outperforming baseline models like CLIP and IP-Adapter.
- Generalization Ability: MagicAnimate demonstrates robust performance in unseen domains and multi-person animation tasks.

- Integration with Text-to-Image Models: The system's ability to animate images generated by text-to-image models (e.g., DALL·E3) expands its usability for diverse applications.
- Seamless Transitions: The video frame fusion technique ensures smooth transitions, preventing abrupt changes or inconsistencies across frames.

Disadvantages:

- High Computational Cost: The use of advanced diffusion models, temporal attention layers, and appearance encoding requires significant computational resources, making the system computationally expensive.
- Limited to Motion Sequences: The model's ability to generate realistic animations is contingent on having a corresponding motion sequence, limiting its use in scenarios where such sequences are unavailable.
- Training Data Dependency: MagicAnimate's performance is heavily reliant on the quality and diversity of the training data (such as the TikTok dataset), which may not generalize well to highly diverse or niche domains without further training or fine-tuning.

Chapter 3

System Architecture

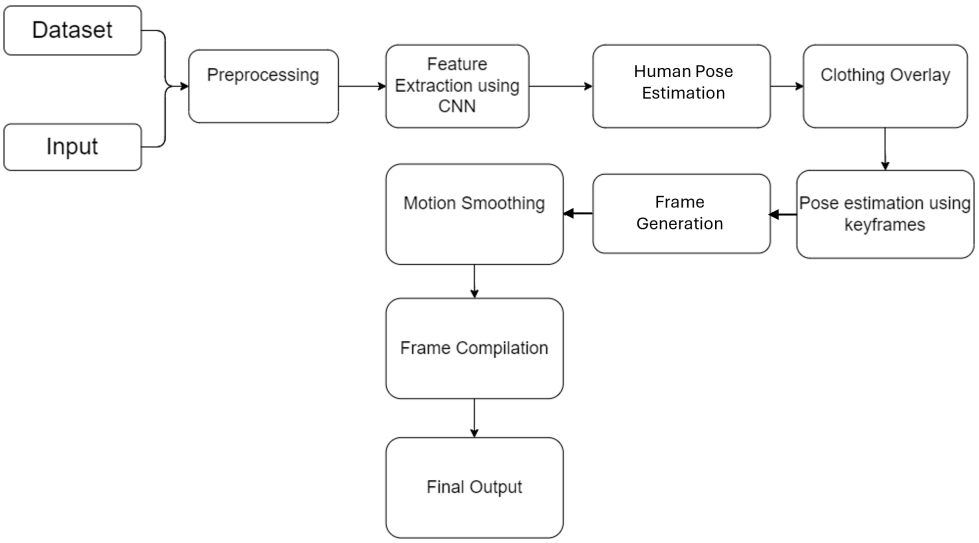


Figure 3.1: System Architecture

The system architecture combines StableVITON and MagicAnimate to create an advanced virtual try-on framework that produces 360-degree dynamic views with realistic garment overlay and seamless movement. It starts with a Preprocessing Module that standardizes input images by resizing, segmenting, and normalizing them. Next, the Feature Extraction Module uses CNNs to capture crucial details of clothing and body features, providing a foundation for accurate pose and clothing alignment.

Pose Estimation then identifies body keypoints, essential for precisely overlaying clothing onto the model. Using these keypoints, the Clothing Overlay Module applies warping techniques, like Thin-Plate Spline (TPS) transformation, to align the garment naturally with the model's pose and shape. To enable smooth animations, the Pose Interpolation Module generates intermediate poses between keyframes, which are used by the Frame Generation Module to create detailed frames representing each step in the movement

sequence.

The Motion Smoothing Module refines the transitions between frames, eliminating jitter to ensure fluid motion. Finally, the Frame Compilation Module arranges the frames in sequence and encodes them into a cohesive animation, creating an interactive, polished virtual try-on experience that can be explored in 360 degrees. This modular design allows for efficient processing and high-quality, lifelike results, ideal for virtual fashion retail applications.

3.1 Dataset

The ST-VTON dataset is designed specifically for training and evaluating virtual try-on systems. It contains high-resolution images of various clothing items, primarily tops and bottoms, along with images of models wearing these items in different poses. Each image pair includes a front-facing image of a model in a neutral pose along with clothing images isolated from a person. Additionally, the dataset provides segmented masks for clothing regions, body parts, and optional background removal, which significantly aids in training segmentation and overlay models. ST-VTON is particularly valuable for virtual try-on applications as it offers diverse clothing styles, textures, and patterns, enabling models to learn how to handle complex clothing features and adapt them to different body shapes and poses. This dataset is also annotated for semantic segmentation, which helps the model align clothing with precise body areas, ensuring realistic garment fitting during the try-on process.

In this project, ST-VTON serves as the primary dataset for training modules that require precise clothing alignment and deformation. It provides rich variations in both clothing and poses, which helps the model generalize better across different body types and styles. The high resolution of ST-VTON images is especially beneficial for preserving clothing details such as patterns, textures, and colors, making the try-on output more realistic and visually appealing.

3.2 Preprocessing Module

- Purpose: This module prepares the input images (either the person or clothing item) to ensure consistent quality and format across all inputs. Preprocessing is crucial

for making sure all images have a standardized appearance and that the model can accurately interpret clothing or body features.

- Tasks:
  - Image Resizing and Cropping: To maintain uniformity, images are resized to a fixed dimension (e.g., 256x256 pixels). This step ensures that all images input into the model have the same dimensions, simplifying further processing steps.

    Formula: For an original image I with height H and width W, resized image I' with new height H' and width W' is calculated by bilinear interpolation or other resizing algorithms.
  - Normalization: Pixel values are scaled to a range (commonly [0,1] or [-1,1]), which stabilizes the training process by ensuring that each pixel intensity falls within a consistent scale.

    Formula: For each pixel value x in the image, normalized pixel value x' is calculated as

    $$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

    for a range of [0,1]
  - Background Removal: Removing the background isolates the person or clothing item, ensuring the model focuses on relevant details. Techniques like Mask R-CNN or U-Net can be used to segment the subject from the background.
  - Dataset Alignment: Each image is categorized and labeled by factors such as pose type or clothing type, streamlining the later stages of feature extraction and overlay.
- Output: A clean, normalized image, free from background distractions, aligned to a consistent dataset format. This image serves as the foundation for accurate feature extraction in subsequent steps.

3.3 Feature Extraction using Convolutional Neural Network (CNN)

- Purpose: Extracts critical features from input images that represent texture, shape, color, and patterns. The features are essential for preserving clothing details and

enabling accurate body-pose alignment.

- Tasks:
  - Convolutional Layers: Detect edges, textures, and patterns at different image levels, producing a feature map that represents the image's detailed characteristics. Formula: For an input image I and a convolution kernel K, the feature map output F is computed as:

$$F_{i,j} = \sum_{m,n} I_{i+m,j+n} \cdot K_{m,n} \tag{3.2}$$

    where i,j are coordinates on the feature map, and m,n are kernel dimensions. Convolution kernels are learned during training and become specialized in identifying image characteristics.
  - Pooling Layers: Max-pooling or average-pooling is used to downsample the feature maps, reducing dimensionality and emphasizing dominant features.

    Max-Pooling Formula: For a pooling window of size k, the value at (i,j) in the pooled map P is

$$P_{i,j} = \max(F_{i:i+k,\, j:j+k}) \tag{3.3}$$

  - Feature Normalization: To enhance generalization, feature maps are normalized (often using Batch Normalization) to maintain uniform distributions during network operations.

- Output: Feature maps or vectors, which contain a condensed representation of the important visual characteristics, such as the clothing's color, texture, and the person's shape. These feature vectors are used for pose alignment and clothing overlay in later stages.

3.4 Human Pose Estimation

- Purpose: Identify key points representing the joints and structure of the human body in the image. This step provides a skeletal framework necessary for correctly overlaying clothing according to body pose and alignment.

- Tasks:
  - Pose Estimation Algorithms: Algorithms like OpenPose, MediaPipe, or DensePose are applied to detect body joints (keypoints) and infer skeletal structure.

    Keypoint Estimation Process: Pose estimation algorithms use a heatmap for each joint. For joint j, the coordinates (x_j, y_j) are found by locating the peak value in the heatmap, representing the joint's position.
  - Keypoint Representation: Coordinates of joints (e.g., shoulders, elbows, hips) are captured and stored as keypoints, which map out the body's skeleton.

    For a model with N keypoints, pose P can be represented as

$$P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{3.4}$$

- Output: A set of keypoints representing the human pose in the image. These keypoints act as reference points for accurately warping and aligning clothing in later steps.

3.5 Clothing Overlay Module

- Purpose: Overlay selected clothing onto the estimated human pose, ensuring a natural fit that aligns with the body shape and pose.
- Tasks:
  - Warping Using Thin-Plate Spline (TPS): TPS is applied to deform the clothing image to fit the person's body, based on detected keypoints.

    TPS finds a smooth transformation that maps source points p_i on the clothing image to target points q_i on the human model. The transformation T is derived by minimizing:

$$E(T) = \sum_i \|T(p_i) - q_i\|^2 + \lambda R(T) \tag{3.5}$$

    where $R(T)$ is a regularization term ensuring smoothness, and λ controls regularization strength.
  - Texture Adjustment Using GANs: A conditional GAN (cGAN) is used to adjust textures, ensuring a seamless blend with the user's body by matching skin tones, lighting, and fabric folds.

- Scaling and Perspective Handling: Scaling adjustments ensure that the garment fits different body shapes, and perspective transformations align the garment according to body angles.

- Output: An image of the user with virtual clothing overlaid, aligned to their pose and body structure. This static try-on image forms the basis for dynamic animations.

3.6 Pose Estimation using Keyframes

- Purpose: Create smooth transitions between poses for animated try-ons by generating intermediary poses from defined keyframes.

- Tasks:

  - Define Keyframes: Set specific poses as keyframes representing critical moments in the animation.

  - Interpolation for Transition Poses: Linear or spline interpolation fills in the frames between keyframes, creating smooth transitions.

    For two poses P_1 and P_2, intermediate pose P at time t is given by:

    $$P = (1 - t) \cdot P_1 + t \cdot P_2 \tag{3.6}$$

    where t ranges from 0 to 1, providing gradual movement between poses.

- Output: A sequence of intermediate poses that enables smooth animation between keyframes, resulting in a natural, fluid movement of the virtual try-on.

3.7 Frame Generation

- Purpose: Generate individual frames for each interpolated pose, displaying the user wearing the clothing in various poses across the animation sequence.

- Tasks:

  - Integration of Pose and Clothing Overlay: For each pose, use the clothing overlay module to warp the clothing and render it on the person.

– GAN-Based Image Synthesis: Utilize GANs to create realistic renderings of each frame, refining details like clothing textures and shadows.

The GAN loss function for each frame synthesis involves minimizing:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right] \quad (3.7)$$

where G is the generator, D is the discriminator, and z is a noise vector or conditioning input.

- Output: A series of frames showing the user in different poses with the clothing applied, forming the basis for an animated sequence.

3.8 Motion Smoothing Module

- Purpose: Ensure that transitions between frames are smooth and free from abrupt jumps or jitter, creating a polished and realistic animation.
- Tasks:
  - Temporal Smoothing: Apply Gaussian smoothing across frames to reduce inconsistencies. For each frame F_i, a smoothed frame F_i' is given by:

$$F_i' = \frac{1}{\sigma\sqrt{2\pi}} \sum_{j=-k}^{k} F_{i+j} e^{-\frac{j^2}{2\sigma^2}} \quad (3.8)$$

where k is the smoothing window size.

- Output: A smoother, more realistic sequence of frames with minimal jitter, ready for compilation into an animation.

3.9 Frame Compilation

- Purpose: Assemble and organize all frames into a single cohesive animation or video file for the final virtual try-on experience.
- Tasks:
  - Sequential Frame Arrangement: Arrange frames in the correct order with consistent timing.

  - Encoding: Encode frames into a video format (e.g., MP4) using a codec like H.264 for efficient storage and playback.

- Output: A polished video of the user in virtual try-on mode, viewable in 360-degree rotation and suitable for interactive fashion applications.

3.10 Final Output

- Purpose: Provide an interactive, polished, and realistic 360-degree video of the user in the selected outfit.

- Output:

  - User-Ready Animation: A fully-rendered, interactive video of the user wearing the chosen clothing, ideal for use in online fashion platforms.

  - Realistic and Smooth Appearance: High-quality animation with refined motion, ready for engagement and showcasing.

Chapter 4

Implementation & Testing

The virtual try-on process begins with image upload and preprocessing, ensuring that input images are clear, aligned,and prepared for further processing. Images of the person and the target outfit are preprocessed to a resolution of 512 x 512 for consistency.

Body segmentation plays a crucial role in accurate garment alignment and realistic rendering. This work employs Dense-Pose, a deep learning model based on an R-CNN architecture,to produce dense human body segmentation.Identifies the major regions of the body such as the arms, face, and torso, facilitating a smooth fitting of the garment.

To generate the cloth mask, OpenCV and NumPy are used for preprocessing, applying color segmentation, edge detection, and morphological operations to isolatethe clothing foreground and remove background noise. The resulting mask enhances garment warping accuracy.

Pose estimation ensures realistic human posture representation. OpenPose is utilized to detect and extract human body keypoints, including joints and limb locations, from static images or video frames. The extracted keypoints are stored in JSON format, enabling precise alignment of garments over various body poses. This enhances the realism and flexibility of the virtual try-on system.

Once body parts and clothing components are segmented,the clothing is warped and aligned to the individual's body shape. Guided by segmentation, pose estimation, and keypoints, the warping process ensures an accurate fit. The GMM aligns the clothing structure, while TPS transformation maintains the fabric's natural bending and stretching properties.

A refined clothing mask is applied to eliminate artificial distortions and preserve realistic garment structure. The output is a static image of the person wearing the target clothing.

To introduce animation, body movements are extracted from a reference video, enhancing interactivity. Pose estimation models, such as DensePose, track limb and joint movements,ensuring that garments move naturally with the wearer. Motion interpolation techniques refine transitions, preventing abrupt movements and improving animation smoothness.

Following movement data extraction, the static try-on image is converted into dynamic animation. The MusePose model generates new frames that mimic natural human movement,incorporating pose-guided synthesis and temporal consistency models to ensure fluid motion. By conditioning frame generation on previous frames, the system maintains a natural flow,reducing choppiness.

The final output is an animated representation of the person wearing the target clothing, with the motion style dictated by the reference video. This approach enhances user engagement by providing a more interactive and realistic virtual try-on experience.

During the testing phase, we evaluated the virtual try-on system using a diverse set of user images and clothing inputs. The test images varied in size, aspect ratio, and quality to ensure the model's adaptability to different input conditions. The system processed these inputs to generate both static and animated outputs, showcasing the applied outfits on the given user images. This helped assess the performance of the model in handling variations in input images while maintaining a visually coherent output.Examples are provided below.

Figure 4.1: Result : Input image of person and target navy blue cloth converted into animated output

Figure 4.2: Result : Input image of person and target maroon cloth converted into animated output

Chapter 5

Results & Discussions

Figure 5.1: Result : Input image of person and target cloth converted into animated output

POSGEN360 creates an animated output of the person wearing the garment by using an image of the person and an image of the clothing as inputs.By maintaining fabric details, natural folds, and a realistic fit across a range of body shapes, the model guarantees precise garment alignment.

Table 5.1: Static Image VTON Metrics Comparison

MODEL NAME	**512 x 512**		
	SSIM[a]	***LPIPS***[b]	***FID***[b]
Our GAN Model	0.870	0.052	14.05
CP-VTON	0.791	0.141	31.96
ACGPN	0.863	0.067	15.22

[a]Higher is better.

[b]Lower is better.

The table compares virtual try-on (VTON) models based on SSIM, LPIPS, and FID for 512 × 512 resolution images, evaluating their ability to generate realistic outputs. Higher SSIM indicates better structural similarity, while lower LPIPS and FID suggest improved perceptual quality and realism. Among the models, Our GAN Model performs best, achieving the highest SSIM (0.870), lowest LPIPS (0.052), and lowest FID (14.05), ensuring superior image fidelity.ACGPN follows closely, while CP-VTON performs the worst with the lowest accuracy and highest perceptual differences.

Table 5.2: Image to Animation Metrics Comparison

MODEL NAME	**512 x 512**			
	SSIM[a]	***PSNR***[a]	***LPIPS***[b]	***FVD***[b]
Our Diffusion Model	0.718	29.56	0.285	171.9
FOMM	0.648	29.01	0.335	405.2
MRAA	0.672	29.39	0.296	284.8
TPSMM	0.673	29.18	0.299	306.1
Disco	0.668	29.03	0.292	292.8
SD-12V	0.670	29.11	0.295	225.5

[a]Higher is better.

[b]Lower is better.

The table evaluates various image-to-animation models based on SSIM, PSNR, LPIPS, and FVD for 512 × 512 resolution outputs, assessing quality, realism, and temporal consistency.Our Diffusion Model outperforms all others, achieving the highest SSIM (0.718) and PSNR (29.56), lowest LPIPS (0.285), and best FVD (171.9),indicating superior image quality and smoother animations.FOMM performs the worst, with the lowest SSIM (0.648) and highest LPIPS (0.335) and FVD (405.2), suggesting poor realism. Other models (MRAA, TPSMM, Disco, and SD-12V) show moderate performance. Overall,Our Diffusion Model generates the most visually accurate and temporally coherent animations.

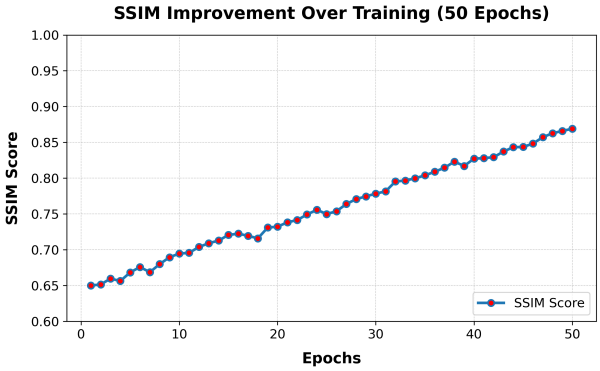


Figure 5.2: Structural similarity measure across epochs for static VTON model

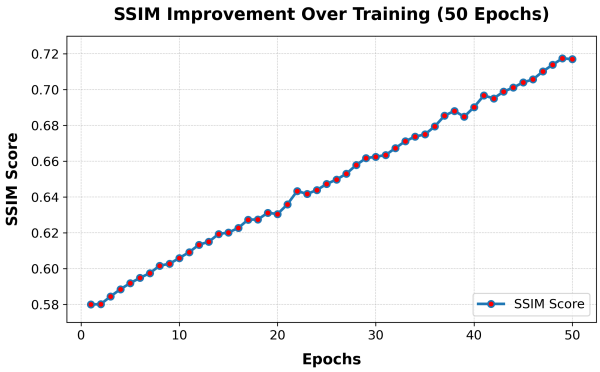


Figure 5.3: Structural similarity measure across epochs for animation model

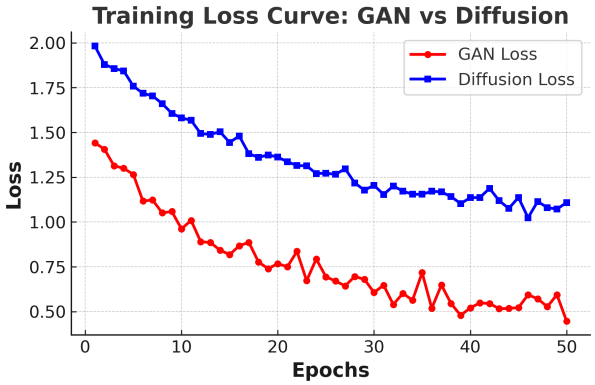


Figure 5.4: Result : Input image of person and target cloth converted into animated output

Figure 5.5: Result of different outfit on the same person

Chapter 6

Conclusions & Future Scope

In conclusion, our project on enhancing 3D-based virtual try-ons through advanced 3D modeling, real-time video analysis, and deep learning techniques marks a significant advancement in improving online shopping experiences. By integrating Convolutional Neural Networks (CNN) for feature extraction and Generative Adversarial Networks (GAN) for realistic garment overlay, the project effectively addresses the challenge of poor fit and high return rates in e-commerce. The system's ability to provide accurate and dynamic garment fitting in real-time not only improves the user experience but also contributes to sustainability by reducing returns.

The future development of this project presents several exciting opportunities to enhance its capabilities.

- Personalized Fit Adjustments: Integrating detailed user measurements to further customize garment fitting.
- Expansion to More Clothing Types: Extending the technology to support a wider range of products, including accessories and footwear, could offer a more comprehensive virtual shopping experience.
- Augmented Reality (AR) Integration: Implementing AR could enhance real-time fitting, enabling users to visualize garments on their own avatars in an immersive environment.
- Motion Smoothing Enhancements: Improving the behavior of garments during movement could create even more lifelike simulations.
- User Feedback Systems: Incorporating real-time user feedback into the system would allow for continuous refinement of fitting algorithms, improving accuracy and user satisfaction.

References

[1] K. Sun, P. Zhang, J. Zhang, and J. Tao, "Dgm-flow: Appearance flow estimation for virtual try-on via dynamic graph matching," *Knowledge-Based Systems*, vol. 302, p. 112377, 2024.

[2] R. Velastegui, M. Tatarchenko, S. Karaoglu, and T. Gevers, "Image semantic segmentation of indoor scenes: A survey," *Computer Vision and Image Understanding*, vol. 248, p. 104102, 2024.

[3] T. Islam, A. Miron, X. Liu, and Y. Li, "Stylevton: A multi-pose virtual try-on with identity and clothing detail preservation," *Neurocomputing*, vol. 594, p. 127887, 2024.

[4] T. Wang, X. Gu, and J. Zhu, "A flow-based generative network for photo-realistic virtual try-on," *IEEE Access*, vol. 10, pp. 40 899–40 909, 2022.

[5] H.-J. Lee, B. Koo, H.-E. Ahn, M. Kang, R. Lee, and G. Park, "Full body virtual try-on with semi-self-supervised learning," *Electronics Letters*, vol. 57, no. 24, pp. 915–917, 2021.

[6] J. Lee, M. Lee, and Y. Kim, "Mt-vton: Multilevel transformation-based virtual try-on for enhancing realism of clothing," *Applied Sciences*, vol. 13, no. 21, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/21/11724

[7] Z. Huang, R. Khan *et al.*, "A review of 3d human body pose estimation and mesh recovery," *Digital Signal Processing*, vol. 128, p. 103628, 2022.

[8] T. T. Tuan, M. R. Minar, H. Ahn, and J. Wainwright, "Multiple pose virtual try-on based on 3d clothing reconstruction," *IEEE Access*, vol. 9, pp. 114 367–114 380, 2021.

[9] C. Zhou, W. Zhang, and Z. Lian, "Enhancing consistency in virtual try-on: A novel diffusion-based approach," *Image and Vision Computing*, vol. 148, p. 105097, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885624002014

[10] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *arXiv*, 2023, available at https://arxiv.org/abs/2311.03383.

Appendix A: Presentation

POSGEN360-VIRTUAL TRY-ON

Project Group 01

Aadarsh Suresh	(U2109001)
Amal Thomas	(U2109008)
Celina Elizabeth Jacob	(U2109021)
Hathik H	(U2109028)

Under the Guidance of Ms.Ancy C A,
Assistant Professor,Dept of CU

RSET
RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)

Dept. of Computer Science and Business Systems,
RSET

Introduction

Overview

This project aims to create an immersive virtual try-on platform that transforms static images of clothing into dynamic, 360-degree video representations. By enabling realistic try-ons that show garment movement and fit, the platform addresses key challenges in online fashion retail, including sizing uncertainty, high return rates, and customer engagement.

Project Statement

- Traditional online shopping struggles with providing an immersive experience, leaving customers unable to accurately assess garment fit and natural movement.
- Elevated return rates, customer dissatisfaction, and a less engaging overall virtual fashion experience

Objectives

- Fit Accuracy
- Realism Enhancement
- Return Rate Reduction
- Consumer Confidence
- Immersive Experience

Introduction

Importance and relevance of the project

- **Enhances Online Shopping Experience** : Provides a more interactive and engaging shopping experience, improving customer satisfaction.
- **Reduces Return Rates** : Addresses sizing uncertainties, minimizing product returns and associated costs for retailers.
- **Increases Consumer Confidence** : Allows customers to visualize clothing more accurately, leading to informed purchase decisions.
- **Aligns with E-commerce Growth**: With the rapid expansion of online shopping, there is a high demand for innovative virtual try-on solutions.
- **Meets Consumer Expectations** : Today's consumers expect personalized, tech-driven experiences, making this platform highly relevant.
- **Supports Digital Fashion Trends** : The rise of digital fashion shows and virtual fitting rooms highlights the need for interactive platforms.

Project Scope

Project Scope

- This project aims to develop a virtual try-on platform that transforms static clothing images into dynamic, 360-degree video representations.Using machine learning for realistic garment visualization, it enhances customer engagement and reduces return rates in online fashion retail.
- The platform will leverage AI to improve garment visualization, reducing sizing uncertainties and boosting consumer confidence.It bridges the gap between online and in-store shopping while supporting the digitalization of fashion.
- This project will not involve physical garment production, in-store fitting technologies, or hardware development like smart mirrors or body scanners.It is solely focused on fashion and excludes applications in other industries like eyewear or cosmetics. Additionally, it will not provide manual garment customization or alterations.

Methodology

1. Image Upload & Preprocessing

- Upload person and target clothing images.

2. Segmentation & Mask Generation

- Generate person mask (segment body & clothing).
- Generate cloth mask (remove background).

3. Garment & Body Segmentation

- Segment body parts (face, arms, torso).
- Refine clothing mask (remove noise).

4. Warping & Alignment (VITON-HD)

- Extract body keypoints (OpenPose/DensePose).
- Warp clothing to fit using GMM.
- Generate refined clothing mask for overlay.

5. Clothing Transfer & Try-On

- Blend warped clothing with person image.
- Post-process to remove artifacts.

Methodology

6. Animation

- Extract pose keypoints from reference video.
- Ensure smooth pose transitions.

7. Image-to-Video Animation

- Animate try-on image using pose sequences.
- Generate frame-by-frame motion (MusePose).
- Refine & smooth movement.

8. Final Video Output

- Compile frames into final animated video.

Methodology

Tools & Libraries

- Deep Learning Frameworks: PyTorch, TensorFlow (optional).
- Image Processing: OpenCV, PIL, NumPy.
- Segmentation Models: SCHP, U^2-Net, Mask R-CNN.
- Pose Estimation: OpenPose, DensePose, MusePose.
- Warping & Alignment: Geometric Matching Module (GMM).
- Web & Deployment (Optional): Flask, FastAPI, ONNX.
- Video Processing: FFmpeg, GAN-based frame interpolation.
- Hardware Acceleration: CUDA, cuDNN, NVIDIA GPUs.

Techniques Used

- Human Parsing & Segmentation – Extracting body parts and clothing regions.
- Geometric Warping – Aligning the target clothing with the person's body.
- Pose Transfer & Keypoint-Based Animation – Animating the person using motion keypoints.
- GAN-Based Image Synthesis – Enhancing realism in try-on and animation.
- Temporal Consistency & Frame Interpolation – Smoothing animation transitions.

Methodology

Algorithms and Models implemented

Virtual Try-On (VITON-HD)

- GMM (Geometric Matching Module) – Warps the clothing to align with the person.
- Refinement Network – Blends the warped clothing with the person's image.
- Human Parsing Models (SCHP, LIP) – Segments person and clothing areas.

Pose Estimation & Animation (MusePose)

- OpenPose / DensePose – Extracts body keypoints for pose recognition.
- MusePose Generator – Creates motion frames for animation.
- Temporal Consistency Techniques – Ensures smooth movement transitions.

Image Processing & Enhancement

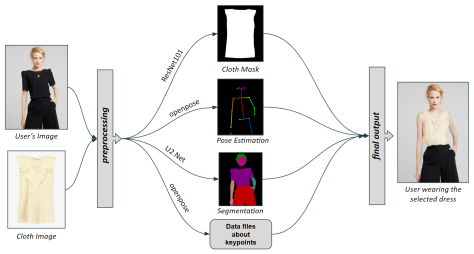
- Mask R-CNN / U^2-Net – Removes background and refines segmentation.
- Super-Resolution Networks – Enhances output image/video quality.

Methodology

Dataset

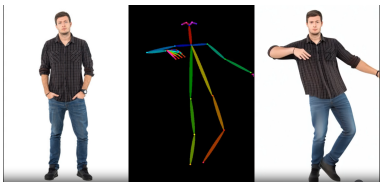
- VITON Dataset – Standard dataset for training virtual try-on models.
- COCO Dataset – Contains human pose keypoints for training OpenPose.

Methodology



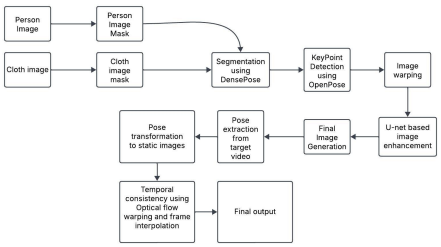
Methodology

Pose Estimation – Extracts human pose keypoints from a reference video.
Pose Transfer Learning – Maps new poses onto a static person image.
GAN-Based Image Synthesis – Generates realistic human motion.
Temporal Smoothing – Ensures smooth animation transitions.



System Design

Architecture/System flow chart



System Design

1. Input Data Preparation

- Person Image: The original image of the person.
- Person Image Mask: A mask generated to segment the person from the background.
- Cloth Image: The target clothing item to be transferred onto the person.
- Cloth Image Mask: A segmented version of the clothing image to remove its background.

2. Segmentation & Keypoint Detection

- Segmentation using DensePose:
  - Extracts body parts and structure from the person image.
  - Helps in aligning the clothing with the person's shape.
- Keypoint Detection using OpenPose:
  - Identifies joints and keypoints of the body (head, arms, torso, legs).
  - Used to align the warped clothing and animate the person later.

3. Virtual Try-On Process

- Image Warping (Geometric Matching Module - GMM):
  - The target clothing is deformed and aligned to fit the person's shape.
- Pose Transformation to Static Images:
  - The person's static image is modified based on detected keypoints.
  - Ensures that the new outfit looks natural on the person.

System Design

4. Animation & Pose Transfer

- Pose Extraction from Target Video:
  - Extract motion sequences from a reference video using pose estimation.
- Final Image Generation:
  - A deep learning model synthesizes realistic frames of the person wearing the new clothing.

5. Enhancement & Temporal Consistency

- U-Net Based Image Enhancement:
  - Refines and removes artifacts in generated images.
- Temporal Consistency using Optical Flow Warping & Frame Interpolation:
  - Ensures smooth transitions between frames for realistic motion.

6. Final Output

- Compiling frames into a final video of the person moving in new clothing.

Implementation

1. Input Data Preparation

- Person Image & Mask: Extracted using a segmentation model to remove the background.
- Cloth Image & Mask: Segmented to isolate the clothing item.
- Preprocessing: Resizing, normalization, and background removal for consistency.

2. Human Segmentation & Keypoint Detection

- Body Segmentation: Extracts different body parts using DensePose.
- Keypoint Detection: Identifies body joints and skeleton structure using OpenPose.

3. Clothing Warping & Overlay

- Image Warping: Transforms the clothing shape to fit the person's body.
- Clothing Overlay: The warped clothing is blended seamlessly with the segmented person.

4. Animation Process

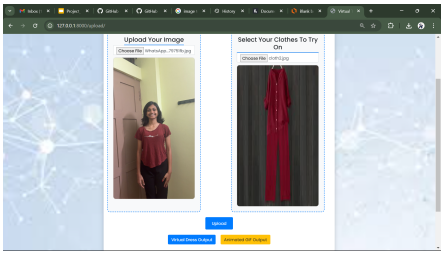
- Pose Extraction from a Dancing Video: Keypoints are extracted from each frame of the selected video.
- Pose Transfer: The person's image is modified to match the extracted poses.
- Frame Generation: Synthesizes new images for each frame, ensuring realistic movements.

5. Video Generation & Refinement

- Image Enhancement: Removes artifacts and improves image quality.
- Smooth Motion Transition: Optical flow warping ensures frame-to-frame consistency.
- Final Video Output: The animated sequence is compiled into a smooth dancing video.

Implementation

Screenshots

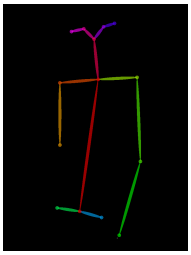


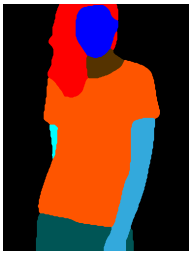
Implementation

Screenshots

Implementation

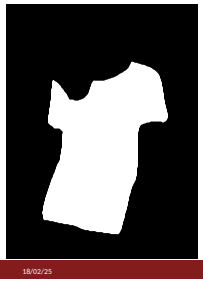
Screenshots



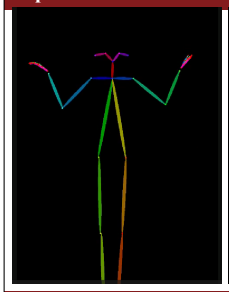


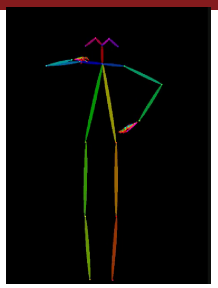
Implementation

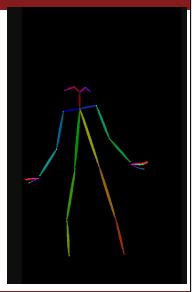
Screenshots



Implementation







Implementation

Challenges

Animating a Static Image

- Lack of depth and multi-view information makes it hard to create realistic movement.
- Artifacts and distortions appear when generating new body positions.

Accuracy Issues

- OpenPose-based animation relies on estimated pose keypoints, which may not align perfectly.
- Clothing deformation during animation might not look natural.

Occlusion Handling

- Some body parts might be hidden in the input image, making it difficult to infer missing details in different poses.

Texture Preservation

- The texture of the swapped clothing might not stay consistent when generating frames.

Temporal Consistency

- Consecutive frames may not transition smoothly, causing flickering or unnatural motion.

System Limitations

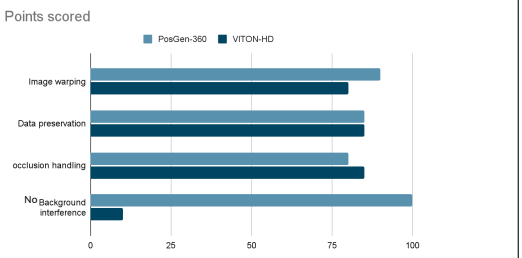
- PC with required specification to train the model

Implementation

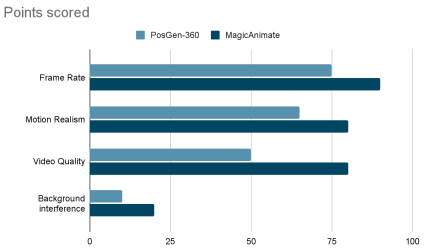
Solution for challenges

- Trained the model on a larger dancing dataset for better motion generation.
- Used a more accurate keypoint-based model to improve upon OpenPose..
- Improved static image accuracy using the VITON-HD dataset for better body part estimation.
- Enhanced warping techniques to maintain texture consistency during animation.
- Trained on sequential video data to improve frame-to-frame stability.

Results and Evaluation



Results and Evaluation



Results and Evaluation

Metric	Value	Interpretation
SSIM (Image Fidelity)	0.70 - 0.80	Moderate similarity, some artifacts present
PSNR (Image Quality)	18 - 22 dB	Noticeable noise and blending issues
Pose Transfer Accuracy	70% - 85%	Some misalignment in complex poses
Optical Flow Consistency	0.6 - 0.75	Motion is not always smooth, minor flickering
Frame Interpolation Quality	10-15% artifacts	Some visible inconsistencies in transitions
Processing Time per Frame	30 sec - 1 min	High computation time per frame
GPU Utilization	60% - 80%	Computationally demanding, but not fully optimized

Results and Evaluation

PosGen-360 vs. MagicAnimate:

- MagicAnimate achieves smoother motion and higher video quality due to better pose transfer.
- PosGen-360 has higher background interference, affecting final output clarity.
- Processing time is higher in PosGen-360, making it less efficient.

PosGen-360 vs. VITON-HD:

- VITON-HD produces only static try-on images, while PosGen-360 generates animated videos.
- VITON-HD has better clothing blending, resulting in higher image quality.
- PosGen-360 doesn't struggles with background interference, whereas VITON-HD struggles with background images..

Conclusion

Summary of the work done

- Developed a virtual try-on system that swaps a person's clothing with a target outfit and animates the result.
- Utilized the VITON-HD dataset to refine static image generation and body part estimation
- Used OpenPose-based animation for movement generation.
- Trained the model on a larger dancing dataset to enhance motion generation.
- Integrated the model with front end (HTML,CSS,JS)

Conclusion

Achievements and key findings

- Successfully generated realistic virtual try-on results for static images.
- Improved pose-based animation quality by using an advanced keypoint model.
- Enhanced clothing texture consistency during animation using improved warping techniques.
- Achieved better frame-to-frame stability through training on sequential video data.

Conclusion

Limitations of the project

- Animation accuracy still needs improvement, as artifacts appear in complex movements.
- Lack of 3D depth information results in unrealistic deformations in some poses.
- Clothing draping and physics are not fully realistic, especially for loose garments.
- Occluded body parts are sometimes incorrectly estimated, affecting final output quality
- System Limitations

Conclusion

Suggestions for future work improvements

- Use 3D Human Reconstruction (e.g., SMPL, PIFuHD) to improve depth perception and animation realism.
- Implement advanced motion transfer models like Liquid Warping GAN or AnimateDiff for smoother animations.
- Incorporate neural texture mapping techniques to enhance clothing detail preservation.
- Leverage diffusion-based video generation models for more natural frame transitions.
- Integrate physics-based cloth simulation to improve garment draping and realism.

References

[1] K. Sun, P. Zhang, J. Zhang, and J. Tao, "Dgm-flow: Appearance flow estimation for virtual try-on via dynamic graph matching," Knowledge-Based Systems, vol. 302, p. 112377, 2024.

[2] R. Velastegui, M. Tatarchenko, S. Karaoglu, and T. Gevers, "Image semantic seg mentation of indoor scenes: A survey," Computer Vision and Image Understanding, vol. 248, p. 104102, 2024.

[3] T. Islam, A. Miron, X. Liu, and Y. Li, "Stylevton: A multi-pose virtual try-on with identity and clothing detail preservation," Neurocomputing, vol. 594, p. 127887, 2024

[4] T. Wang, X. Gu, and J. Zhu, "A flow-based generative network for photo-realistic virtual try-on," IEEE Access, vol. 10, pp. 40899–40909, 2022.

[5] H.-J. Lee, B. Koo, H.-E. Ahn, M. Kang, R. Lee, and G. Park, "Full body virtual try-on with semi-self-supervised learning," Electronics Letters, vol. 57, no. 24, pp. 915–917, 2021.

[6] J. Lee, M. Lee, and Y. Kim, "Mt-vton: Multilevel transformation-based virtual try-on for enhancing realism of clothing," Applied Sciences, vol. 13, no. 21, 2023. [Online].

[7] Z. Huang, R. Khan et al., "A review of 3d human body pose estimation and mesh recovery," Digital Signal Processing, vol. 128, p. 103628, 2022

[8] T. T. Tuan, M. R. Minar, H. Ahn, and J. Wainwright, "Multiple pose virtual try-on based on 3d clothing reconstruction," IEEE Access, vol. 9, pp. 114367–114380, 2021.

[9] C. Zhou, W. Zhang, and Z. Lian, "Enhancing consistency in virtual try-on: A novel diffusion-based approach," Image and Vision Computing, vol. 148, p. 105097, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/ pii/S0262885624002014

[10] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in arXiv, 2023, available at https://arxiv.org/abs/2311.03383

Appendix B: Vision, Mission, PO, PSO, and CO

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To evolve into a department of excellence in information technology by the creation and exchange of knowledge through leading-edge research, innovation and services, which will in turn contribute towards solving complex societal problems and thus building a peaceful and prosperous mankind.

Department Mission

To impart high-quality technical education, research training, professionalism and strong ethical values in the young minds for ensuring their productive careers in industry and academia so as to work with a commitment to the betterment of mankind.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze

complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and abil-

ity to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Business Systems Programme will:

PSO 1: Programming and Software Development Skills
Demonstrate ability to analyse, design, and implement software solutions incorporating various programming concepts.

PSO 2: Engineering Management and Collaboration
Comprehend professional, managerial, and financial aspects of business and collaborate on the design, implementation, and integration of engineering solutions.

PSO 3: Decision-Making and Analytical Techniques in Engineering and Business
Create, select, and apply appropriate techniques and business tools, including prediction and data analytics, for complex engineering activities and business solutions.

Course Outcomes (CO)

After successful completion of the course, the students will be able to:

Course Outcome 1: Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

Course Outcome 2: Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

Course Outcome 3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

Course Outcome 4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

Course Outcome 5: Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

Course Outcome 6: Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

Appendix C: CO-PO-PSO Mapping

Mapping of Course Outcomes with Programme Outcomes

CO - PO Mapping

CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
1	2	2	2	1	2	2	2	1	1	1	1	2
2	2	2	2		1	3	3	1	1		1	1
3									3	2	2	1
4					2			3	2	2	3	2
5	2	3	3	1	2							1
6					2			2	2	3	1	1

Mapping of Course Outcomes with Programme Specific Outcomes

CO - PSO Mapping

CO	PSO 1	PSO 2	PSO 3
1	3	1	2
2	3	3	2
3		3	
4		1	1
5	1	1	1
6		2	