Virtual Fitting Solution using 3D Human Modelling and Garments

Sagar Padekar, B.Eng.

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Future Networked Systems)

Supervisor: Prof. Aljosa Smolic & Co. Supervisor Dr Cagri Ozcinar

August 2021

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Sagar Padekar

August 31, 2021

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Sagar Padekar

August 31, 2021

Acknowledgments

I would like to express my gratitude to Prof. Aljosa Smolic, Dr Cagri Ozcinar for their valuable guidance and motivation during this dissertation. I would like to thank my friends and family for their support, motivation, and encouragement throughout the course.

Sagar Padekar

University of Dublin, Trinity College

August 2021

Virtual fitting solution using 3D human modelling and garments

Sagar Padekar, Master of Science in Computer Science

University of Dublin, Trinity College, 2021

Supervisor: Prof. Aljosa Smolic & Co. Supervisor Dr Cagri Ozcinar

Customers are not able to try-on fashion garments because of the current pandemic, moreover a lot of customers are buying garments online and if there was a feature to virtually try-on garments it would ease the decision-making for customers. This project aims to design a system which will use the user's 2D image to project a 3D model of the user and fit garments on the generated 3D model by using deep learning techniques. This project presents an integrated approach for virtual try-on systems by using multiple neural networks for different tasks of pose estimation, body segmentation, garment fitting and 3D modelling of the human body. This system is compared and evaluated against the state-of-the-art implementations of virtual try-on systems by using different quantitative and qualitative metrics. This research also provides critical review for the current virtual try-on systems and the optimisations required for improving their performance and accuracy.

Content

List of Figures

List of Tables

Chapter 1 - Introduction

This chapter provides some background information about the project that is explored, as well as a brief description of its significance and potential impact. This chapter also goes over why this project was developed and what the key goals of the system are. Finally, it explains the structure and organization of this dissertation.

1.1 Motivation -

In recent times there has been a significant boom in online garment shopping because of its several advantages. This has led to a lot of consumers shift away from the traditional in-store shopping methods. Online shopping has its own advantages with respect to choice, price, etc. but it also has a major drawback and that is the customers cannot try on the clothes before buying them. Due to this the customers are hesitant before buying a garment as they cannot try on the clothes and see how they fit and how it looks on their body. Also, because of the recent COVID-19 pandemic, the customers do not have the option to buy clothes from the physical stores because of the lockdown restrictions and safety issues. This has resulted in demand for a feature which could allow the customers to virtually try on the garments before buying them online.

Virtual try on technology has recently piqued curiosity by allowing the customers to try on the garments before-hand and providing them with appropriate information about the product like its colour, fitting, etc. It allows the users to try on different outfits and experience them virtually [14]. This helps the customers to ease the decision-making process and the merchants to boost their sales efficiency.

In the last couple of years, virtual try-on systems are more focused on showing the garment fitting results in 2D format as the 3D solutions are not economical

and require a lot of computational capabilities. The basic fundamentals of such systems is taking the image of the person and cloth and creating a new image which shows the person dressed with the target cloth [1][2][5][14]. Such systems also make use of other neural networks for human pose detection like [8][13] and body part segmentation like [9][10][11][12] for achieving these results. These systems are capable of maintaining the texture of the garments, and also with minimal distortions. Although such systems are good, they do not provide a 3D visualisation for the customers.

There have been several implementations for reconstructing 3D human models from a single 2D image. Although certain implementations are not capable of reconstructing the entire human body or with distorted body parts[18], there are quite a few which can reconstruct the completely clothed human body without causing a lot of distortions [16][17].

Hence there is a need for designing and developing a system that will help to allow virtual try-on of clothes on 2D human images and project them into 3D for robust visualisation.

This dissertation aims to create a system that will provide virtual garment fitting and also showcase the results in 3D format.

1.2 Objectives -

The several techniques outlined in section 1.1: Human pose estimation, body parts segmentation, 2D virtual garment fitting and 3D reconstruction, serve as the foundation for this project. As a result, the fundamental idea of combining the four methodologies can be applied to solve the real-world challenge of 3D virtual garment fitting.

Thus, the objectives of this project are -

- Combining multiple deep learning models to create a system that uses all of the models' inferences and produces a single output.
- Create a pipeline for integrating 2D based virtual garment fitting solutions in conjunction with 3D reconstruction networks, to visualize the virtual try-on results in 3D.
- Analyse different neural networks doing various tasks and determine the best solutions for each task.

1.3 Outline -

The dissertation is arranged into five chapters each dealing with specific areas in detail as described below -

- Chapter 1 : Introduction - This chapter gives a brief introduction to the ideas and concepts of the project. It provides details about the techniques which could be used for implementing the project. It also discusses the motivation and objectives of the dissertation.
- Chapter 2 : State of the Art – This chapter provides a comprehensive overview of the state-of-the-art techniques which are currently in use. It provides a detailed analysis of the all the techniques like pose estimation, body segmentation, and 3D reconstruction which are used to create a virtual try-on neural network. The section 2.3 details the various garment fitting methods which form the basis of this research project.
- Chapter 3 : Design and Implementation – In this chapter, the design of the proposed pipeline for virtual garment fitting is described in detail along with its implementation. This chapter also provides the justification for choosing the models which were used in the proposed pipeline.
- Chapter 4 : Results and Evaluation – This chapter provides the results and details of the output generated from each and every component of the developed pipeline. The section 4.6 provides quantitative and qualitative evaluation of the results against the state-of-the-art methods. This chapter also provides a detailed discussion and analysis of the generated results along with the failure scenarios.
- Chapter 5 : Conclusion – The dissertation report is concluded in chapter 5 which highlights the major contributions from this project and provides directions for improvements for the system and areas to be explored in future research work.

Chapter 2 - State of the Art

This chapter begins with a description of the different techniques used in this research, including human pose estimation, body part segmentation, 2D virtual clothing fitting, and 3D reconstruction. It goes on to discuss several implementations of each of these strategies as well as comparisons between them in relation to this dissertation.

2.1. Human Pose Estimation

Human pose estimation is a method for recognizing and analysing human posture. The modelling of the human body is the most important task in human pose estimation. The procedure normally begins with the extraction of joints from a human body, followed by deep learning algorithms analysing a human position. The three most common human body models are Skeleton-based, Contour-based, and Volume-based [19]. Skeleton-based - The skeletal structure of a human body is represented by a series of joints (key-points) such as ankles, knees, shoulders, elbows, wrists, and limb orientations. Because of its versatility, this model is employed in both 2D and 3D human pose estimation methodologies. Contour-based - The contour and rough width of the body torso and limbs are depicted with borders and rectangles of a person's silhouette in a contour-based model. Volume-based - 3D human body shapes and poses are represented by volume-based models with geometric meshes and shapes, which are often obtained via 3D scans.

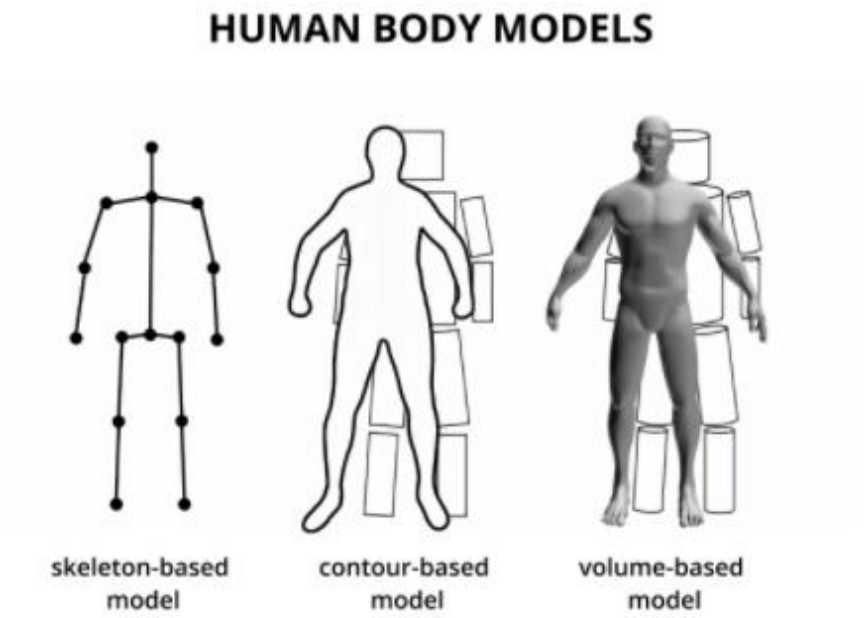

Figure 1: Human Body models [19]

In Skeleton-based modelling, the identification and analysis of X, Y coordinates of human body joints from an RGB image is used to calculate 2D pose estimate, while for 3D posture estimate X, Y, and Z coordinates of human body joints are used. This set of coordinates can be linked together to describe a person's pose. Each skeletal co-ordinate is referred to as a keypoint. The connection between two keypoints is known as a pair [19]. The different state of the art networks which are used for pose estimation are explained in the below subsections.

2.1.1 OpenPose

OpenPose initially finds key-points that correspond to each person in the image, then assigns these key-points to different people. OpenPose neural network is trained on COCO dataset which is split into - 14K annotations for training set and 545 annotations for the validation set. Using the first few layers of VGG-19, the OpenPose network collects features from an image. The features are then sent into two convolutional layer branches that run parallel to each other. The first branch predicts 18 confidence maps, each of which represents a different section of the human posture skeleton. The second branch predicts a collection of 38 Part Affinity Fields (PAFs), which are used to describe the degree of relationship between key-points. Each branch's predictions are refined over successive stages. Bipartite graphs are created between pairs of parts using part confidence maps. Weaker linkages in bipartite graphs are trimmed using PAF values. This process finally estimates the human pose based on the skeleton-based model [13]. The working of this network is explained in more details in section 3.4.

2.1.2 DensePose

The Facebook AI research team created DensePose to estimate 3D postures from a 2D image on a surface-based human model. Densepose is trained on COCO-densepose dataset, which is a collection of 50,000 manually annotated images with image-to-surface correspondences. In the annotation process the image is segmented into twenty-five different regions, where one region is for the background and twenty-four are for representing the human body parts. Each region of the body parts is specified with U,V coordinates for establishing image to surface relation. Densepose is implemented using multiple combinations of neural networks that combine the regression and classification tasks. The performance of these networks is then evaluated and compared with each other and state-of-the-art pose estimation networks. Firstly, the classification process is carried out, in which each pixel is classified into the

background region or one of the body parts, by training the network using cross-entropy loss. In the next step, regression task is carried out which identifies the U,V coordinates of each pixel belonging to the classified semantic region, by training the regressor using L1 loss. Densepose is implemented by using Fully-convolutional networks (FCN), Region-based convolutional networks (RCNN) and Mask-RCNN. These networks are combined with each other by using multi-task cascading. The different architectures are then evaluated using Area under the curve (AUC) and Intersection over Union (IoU) metrics. It is concluded that Densepose implemented with Mask-RCNN using distillation and cascading techniques outperforms the other implementations based on Fully-convolutional networks and Region-based convolutional networks [8].

2.1.3 DeepCut

DeepCut provides an approach for detecting and estimating the human body pose, it can also identify multiple persons and perform segmentation on their body parts. To achieve this 3 problems are defined. 1st problem is to generate a set of body parts which represent the possible organs displayed in the image. 2nd problem is to label each body part where labels are the organ types. 3rd problem is to segment the body parts belonging to the same person. DeepCut models this problems into Integer Linear Programming (ILP) problem and solves it using Faster-region based CNN. DeepCut is able to make inferences on multiple people, their postures and body part segmentation [20].

2.2. Body Part Segmentation

Deep learning techniques are used in image processing for pixel wise segmentation of images. This task is used for semantic segmentation of pixels and predicting them into a set of labels. In body part segmentation, the input image of the human body is taken and then it is classified into different body parts. Deep neural networks are used to achieve this functionality, by using different convolutional layers, pooling layers and batch normalization techniques. For virtual try on applications, it is important to separate the body parts of the human body so as to perfectly align and fit the corresponding garment. Various state of the art segmentation techniques are explained in the following sections.

2.2.1. Graphonomy

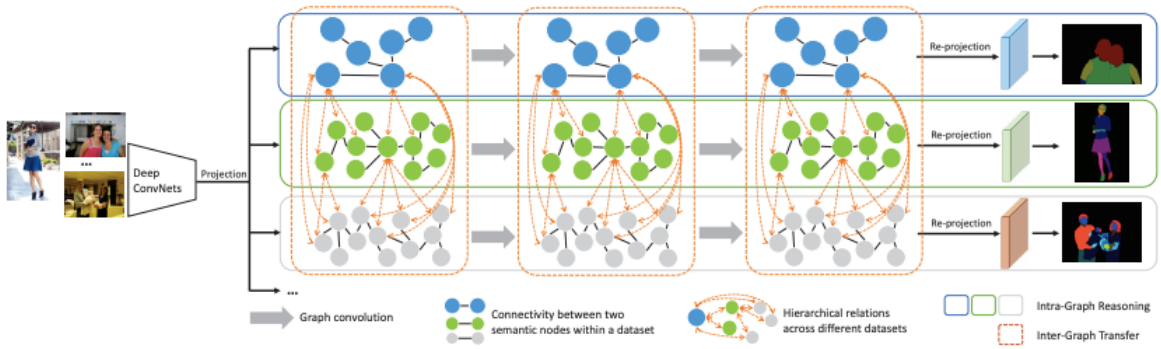

Figure 2: Graphonomy Architecture [12]

As shown in figure 2, Graphonomy uses graph transfer learning to generate universal human parsing for several human parsing tasks and using annotations in a better way. Deep convolutional networks extract visual features that are correlated to a specific semantic part (e.g. neck). These features are then projected into a high-level graph. Semantic edges and nodes are then constructed according to the body structure. By using Intra-Graph Reasoning for transmitting global knowledge, the visual features of the generated body parts are enhanced. The network then transfers and fuses semantic graph representations using Inter-Graph Transfer and hierarchical label correlation to

reduce mislabelling. This network is trained on varying levels of granularity on the CIHP dataset, which is split into - 28,280 images for training, 5,000 for validation and 5,000 for testing. Due to this, the Graphonomy network is able to make predictions in the wild and generate human parsing results with segmented body parts and accurate boundary detection [12].

2.2.2. CIHP_PGN

This neural network provides instance level human parsing by using part grouping network. It generates body segmented maps and edge maps used for creating segmented lines. It achieves this in 4 stages - Semantic part segmentation, Instance-aware edge detection, refinement, and Instance partition process [9]. The working of this network is described in more detail in section 3.3.

2.2.3. LIP_JPPNet

This is deep learning model for body part segmentation and pose detection built using TensorFlow. This network is trained on Look into People (LIP) Dataset. JPPNet network for human parsing and positioning includes multiscale functional connections and iterative location refinement to examine effective feature modelling which provides efficient parsing and posing inferences. For both human parsing and pose estimating tasks, this unified system offers cutting-edge performance [10].

2.3. Virtual clothing fitting

Deep learning techniques in conjunction with image processing are used for virtual garment fitting. Such techniques use deep neural networks and fit the garments on the basis of human pose and body parts. The following sections will describe the various state of the art virtual try on methods.

2.3.1. VOGUE Try-On

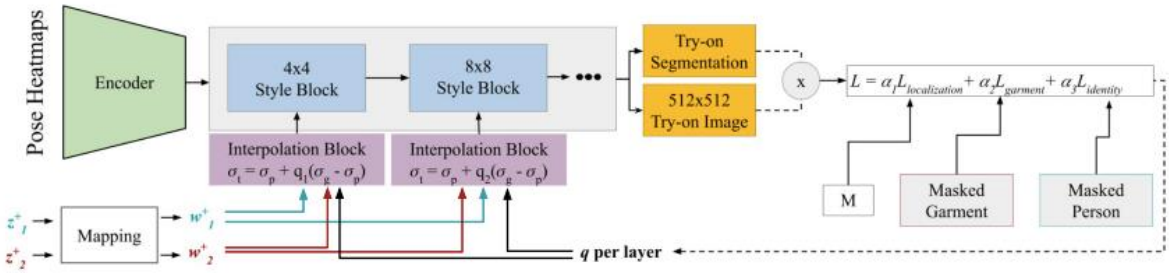

Figure 3: TryOnGAN Architecture [1]

This method uses Generative Adversarial Network as shown in figure 3 for fitting the garment image on the target human body. In each layer of the network, they have trained a pose-conditioned GAN that generates an RGB image and clothing segmentation. Instead of generating a constant input, pose heatmaps are encoded and passed as the input into the first block of the GAN. As shown in figure 3 the GAN takes two images and an encoded heatmap as the input for pose-generator. The generator produces the try on image and respective segmentation by using interpolation techniques. The loss function is minimized over the interpolation coefficients per layer and the network is then able to transfer garment image onto the target human image [1].

2.3.2. Pix2Surf

This research provides an effective way to transfer textures form garments to the 3D clothes on SMPL human body model. In this approach, first the 3D garments are converted into 2D for creating training data. The neural network is then trained on this data to learn the mapping of pixels to 3D surface. The

main objective of this network is to learn the relation from the image pixels with UV parameters of the SMPL surface. This model can only transfer textures of the garments [7].

2.3.3. Multi-Garment Net

This research proposes a method to predict body shape and clothing by using SMPL human body model. This network is trained on DigitalWardrobe dataset which contains 3D scans of the people in different poses and clothes. This model is able to predict garment geometry from the body shape and transfer it another body shapes. The training of this network on such databases which contain 3D scans of the people takes a lot of time and computational power [3].

2.3.4. CP-VTON+

This research proposes a method to fit garments on a 2D human image by learning the body pose, parts, and cloth structure. It achieves this in two stages - The first stage is to learn the cloth structure and warp it on human body according to the pose of the target person. In the second stage it blends the warped cloth by using composition masks [15]. The working of this network is described in more detail in section 3.5.

2.4. 3D reconstruction

Deep neural networks are used to reconstruct the 3D shape of an object from its 2D image. This is achieved by using encoding-decoding techniques. In this the 3D structure of the shape of the object is encoded from the 2D image. After this the decoder is applied on the structure to reconstruct the 3D shape of the input object. Generally, a ResNet is used for 2D encoding and generating the vector from the input image. Decoder is mostly a FCN with a single classification output, which receives the vector and 3D coordinates from the encoder and classifies the coordinate. Each coordinate is classified by the decoder, which then builds a representation of the 3D structure. Several state-of-the-art 3D reconstruction methods are discussed in the below sections.

2.4.1. Deep Human

For 3D human reconstructing from a single RGB image, DeepHuman provides an image-guided volume to volume CNN translation. DeepHuman uses a dense semantic representation created as a supplementary input from the SMPL model to reduce the ambiguity related with surface geometry reconstruction, even for rehabilitating unseen portions [17]. One of the main benefits of this network is that it merges the various scales of image features into 3D space by means of a realistic surface geometry transformation. The visible surface features are further polished by a normal refining network which is linked with the volumetric generating network. This paper also provides a 3D human model dataset called as THuman that contains seventy thousand human models.

2.4.2. Body Net

This network proposes a pipeline for directly inferencing volumetric body shape from a single image. It generates the 3D human models from a single 2D image. The network is trained on MPII Human Pose dataset for 2D pose estimation and SURREAL dataset for body parts segmentation. The evaluations performed on SURREAL and Unite the People datasets, show promising results. This network also allows volumetric body-part segmentation [18].

2.4.3. PiFuHD

This study establishes a multi-level framework for high-resolution 3D reconstructions of clothed persons from a single 2D image. For faster inferences, it employs a multi-level Pixel-Aligned Implicit Function and outperforms other state-of-the-art methods. The working of this approach is described in more detail in section 3.6.1.

2.5. Datasets

Significant progress has been made in recent years for creating datasets which help the neural networks to make accurate predictions on the human body shape and pose. The datasets required in virtual try-on systems must have sufficient annotations of 3D clothing data. The below sections describe state-of-the-art datasets used in virtual garment fitting systems.

2.5.1. DeepFashion 3D

This is a 3D clothing dataset with annotations for image-based garment reconstruction, which provides a method for single-view image reconstruction of garments. This dataset is constructed by using Agisoft tool to generate garment reconstructions in dense point cloud format for garment capturing. The garments are posed on dummy models or real persons for generating deformations and pose augmentations. The dataset is annotated by using feature lines that represent the various parts of the human body like neck, waist, shoulder, etc. These labels are then used for supervised learning.
A hybrid approach is followed for modelling minute geometric details in single view reconstruction of the garments.

The three stages involved in reconstructing the garment images is as follows -

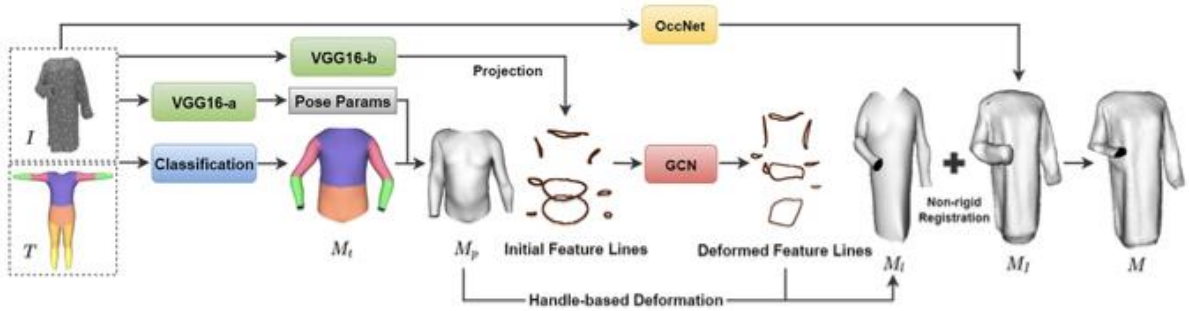

Figure 4: DeepFashion3D Architecture [6]

Template mesh generation - An adaptable template is used for generating the mesh, which built on SMPL model and segmented into 6 regions as shown in the figure. Depending on the cloth topology different regions are activated. Cloth classification network based on VGGNet is used to generate the appropriate template (Mt).

Surface reconstruction - Pose estimation is done first to reduce the searching space and deform Mt to obtain a new mesh Mp. Graph convolution network (GCN) is used for feature line regression. Handle-based deformation technique is then applied to obtain Ml.

Surface refinement - OccNet is applied on the input image directly for reconstructing the surface Mi. The outliers generated because of OccNet are then removed by using adaptive registration techniques [6].

2.5.2. CLOTH3D

CLOTH3D is a large scale dataset designed for clothing on 3D human models. This dataset contains variety of garments and provides comprehensive details with respect to the garment size, shape, pose and texture. This dataset is generated by simulating the garments on various poses and body shapes. The authors have also provided a generative model for synthesizing the garments by using graph convolution techniques. This dataset can be used for fitting the garments on a SMPL model in any position or shape [61].

Chapter 3 - Design and Implementation

This chapter describes the proposed design of the pipeline and explains each of the components in detail, along with justifying the choice of the models which were used. In the next subsection it provides details about the frameworks and technologies that were used for the implementation of this pipeline.

3.1. Pipeline Architecture

The proposed pipeline architecture is as shown in the below figure 5. The pipeline of this system consists of five major components. The first component is cloth-masking component (CMC) which is responsible for generating binary masks from clothing images. The second component is human-segmentation component (HSC) that produces semantic part segmentation of the human body. The next component of the pipeline is pose-detection component (PDC) which estimates the human pose and generates the pose key points of the person. The fourth component of the pipeline is the garment-fitting component (GFC) that fits the garments on the 2D human image. The final component is the 3D-reconstruction component (RC3) which converts the 2D image with garments to generate the 3D visualisation and applies photogrammetry techniques to get the expected result. The pipeline is then completed and a 3D output of the target human with the garments selected is generated. The functionality and working of each component of the pipeline is discussed in detail in the sections below.

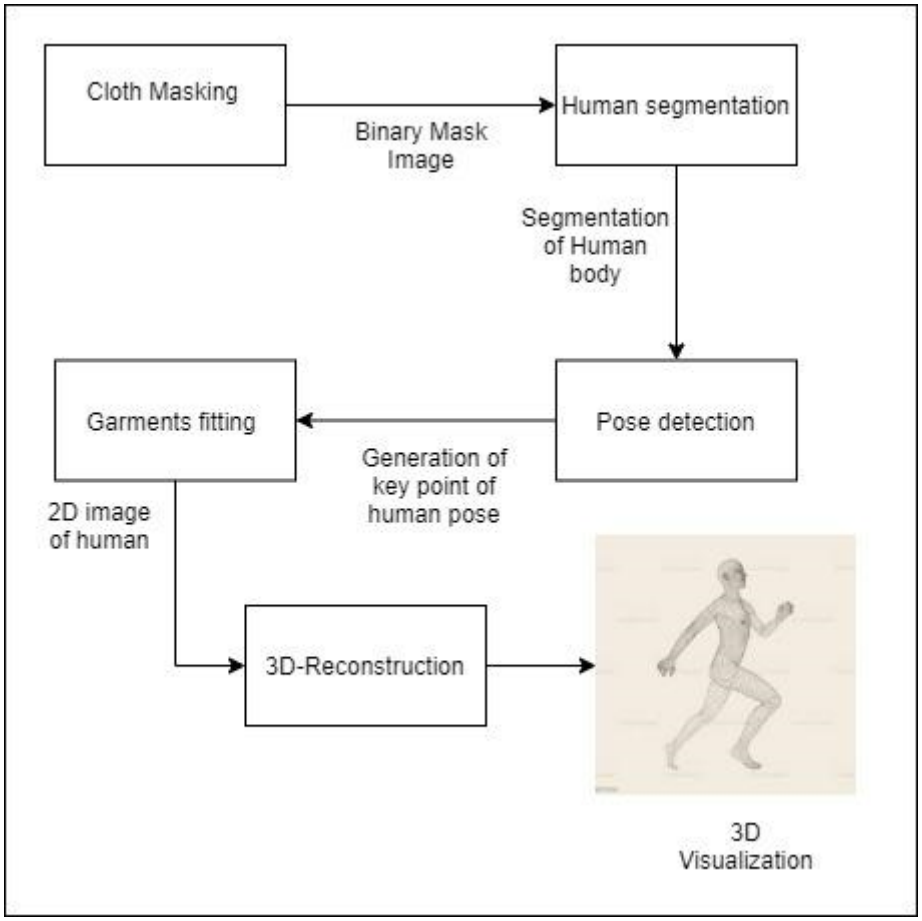

Figure 5: Full Architecture of Proposed Pipeline

3.2. Cloth Masking Component

This component takes the input of a 2D garment image that is to be fitted on the human body. The image is then processed by using the following four techniques to generate the binary mask.

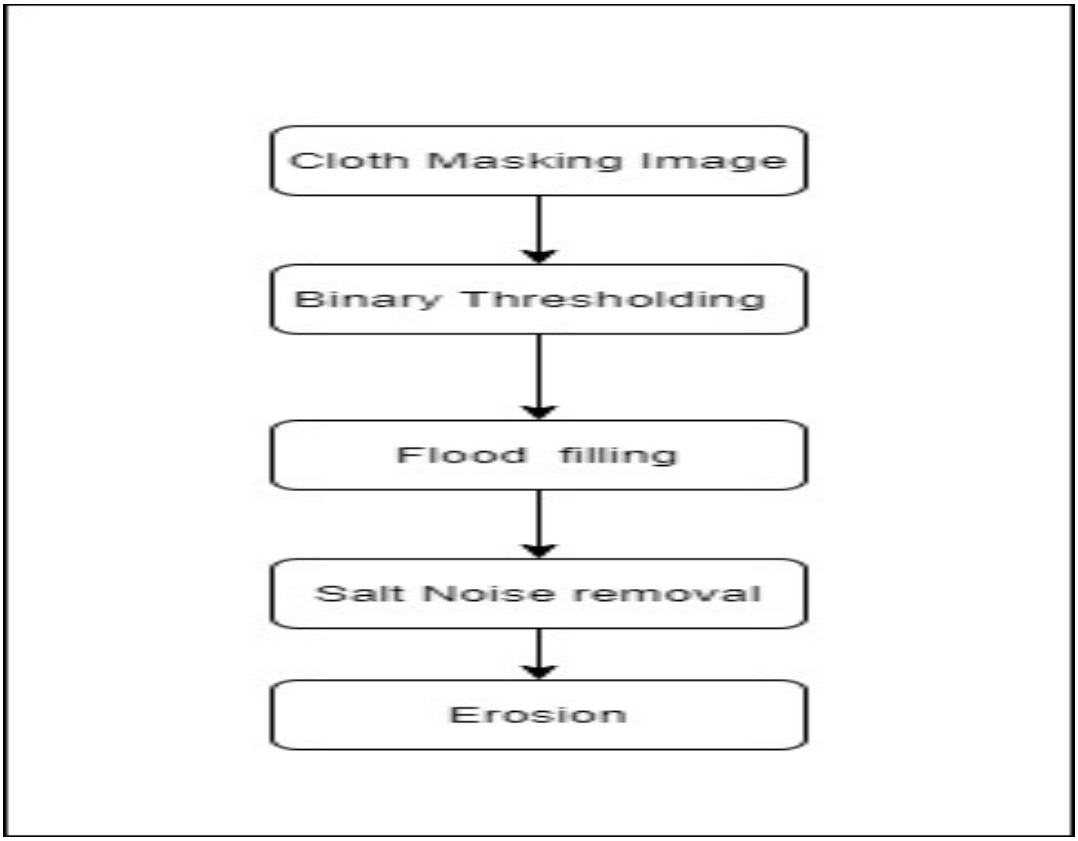

Figure 6: Cloth Masking Techniques

3.2.1. Binary thresholding

Thresholding technique is allocating of pixel values with respect to a specific value, which is the threshold value. Thresholding is used for differentiating the cloth into the foreground from its background. First the image is converted into grayscale format. In binary thresholding each pixel value is compared to the specified threshold and if it is less than the threshold value, then the value is set to 255, else to 0.

3.2.2. Flood Filling

This is an algorithm used to fill the empty spaces in the image. After the binary mask of the cloth is generated then it may contain some void holes. This technique is applied to fill those holes.

3.2.3. Salt noise removal

Salt noise can sometimes be formed on the mask in the form of scattered black and white pixels. After filling the void holes of the cloth-mask, opening operation is carried out on the image. This operation results in removing the salt noise from the image.

3.2.4. Erosion operation

Erosion operation is carried out on the generated mask as the final step of processing. This operation is used to remove pixels from the cloth's boundary. This results in thinning out the edge of the cloth mask from the background.
As a result of the above techniques, the CMC produces a better fitting mask of the clothing image. This mask image is used as an input to the GFC.

3.3 Human Segmentation Component

In this component the 2D image of the human is provided as the input to generate a segmented image on the basis of different body parts. To achieve this result, a state-of-the-art part grouping network is used. The neural network used is CIHP_PGN (Crowd Instance level Human Parsing-Part Grouping Network) . This network is trained and evaluated on the CIHP dataset, and uses different techniques like Semantic part segmentation, Instance-aware edge detection, Refinement, and Instance-partition, and outperforms most of the other models [9]. CIHP_PGN is based on Res-Net 101,Deeplab-v2 for encoding human features. By using different coarse-to-fine techniques in conjunction with this Res-Net, the CIHP_PGN is able to detect semantic data with varying scales and from different sub-regions [9]. The network architecture and its working is

described below.

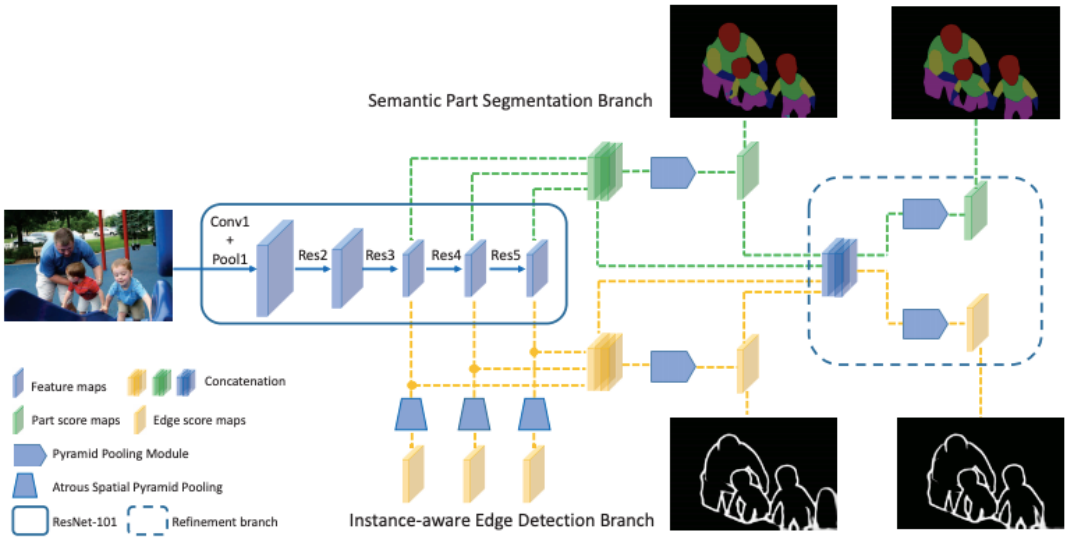

Figure 7: CIHP_PGN Network Architecture

The network is designed to achieve instance level human parsing. PGN generates part segmented maps and edge maps which are used to create segmented lines. The segmented lines are grouped into different regions according to part labels. This network outputs a complete human parse result for the given image on the basis of instance and part segmentation maps. The PGN consists of the following branches -

3.3.1. Semantic Part Segmentation

As shown in the figure 7, the first branch of this network is used for segmentation. This branch is used for identifying and segmenting the different body parts of the provided input image. Body parts are predicted on the basis of shared labels at different scales. To discard the inputs from multiple scales, these predictions are filtered by using a context aggregation pattern. Pixel-wise recognition is performed to allocate each pixel for a labelled body part [9].

3.3.2. Instance-aware Edge Detection

This branch is used for detecting the boundaries or edges of the human body from the given input image. By using deep supervision and atrous spatial pyramid pooling techniques in conjunction with edge score maps this branch is able to perfectly detect the boundaries of the human body from the background of the provided input image [9].

3.3.3 Refinement

This branch is responsible to refine the predictions made by both the segmentation edge detection branches. This uses two pyramid pooling modules in conjunction with the remapped feature maps from both the branches and improves the overall predictions [9].

3.3.4. Instance partition process

The PGN outputs the segmentation and edge maps for the human body. In this process, these maps are scanned to generate horizontal and vertical segmented lines. Breadth-first search algorithm is used to group these lines into different parts. Combining the output of this process with the PGN output, the network finally provides an instance level body segmented human parsing output [9].

By using this network in the pipeline, HSC is able to provide an accurate human parse image with appropriate segmentations of the body parts. This output is then used to feed the GFC.

3.4. Pose Detection Component

The human pose of the provided input 2D image is important with respect to the overall objective of the system. As in the given input of the person's image, the person can be in different positions and the garment must be fitted on the person irrespective of their pose. So, it is important that the pipeline uses a method to detect the pose of the person and then fit the garment accordingly. To estimate the pose, it is necessary to concentrate on human body parts like the neck, elbows, and hands of the provided input image. To achieve this, the pipeline uses Openpose which provides a state-of-the-art neural network for human pose estimation. The overall architecture and methodology of OpenPose is described in detail below

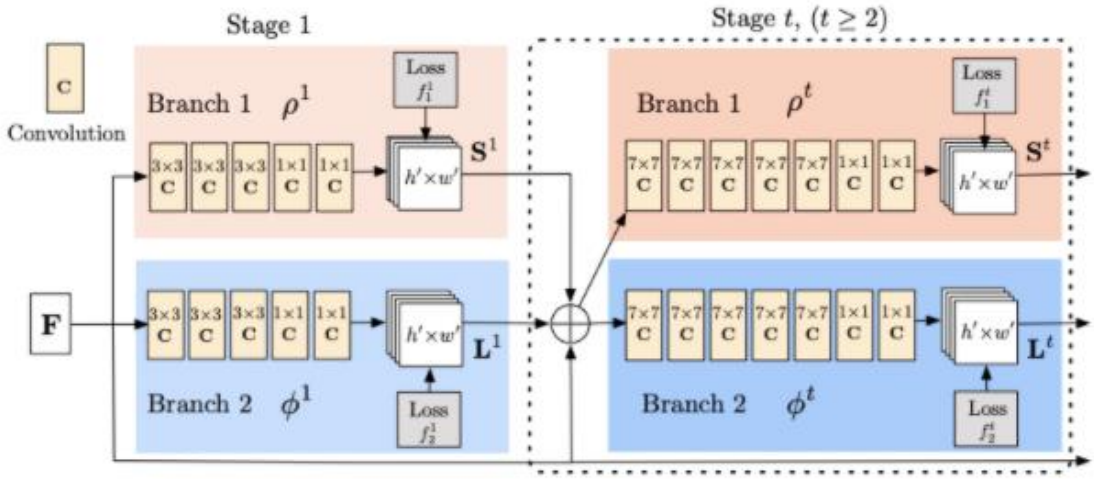

Figure 8: OpenPose Network Architecture [13]

3.4.1 OpenPose Network Architecture

As shown in figure 8, the architecture comprises two branch multistage neural networks [13]. The 1st branch predicts the confidence maps of each part of the human body and the 2nd branch predicts the value of association between different body parts which are called affinity fields. As this network is multi-stage, it first generates a set of confidence maps and affinity fields. In next

stages, it will compare the predictions with the previous stages and the original image and use it to generate refined output of the human pose by using greedy inference [13].

3.4.2. Key-points

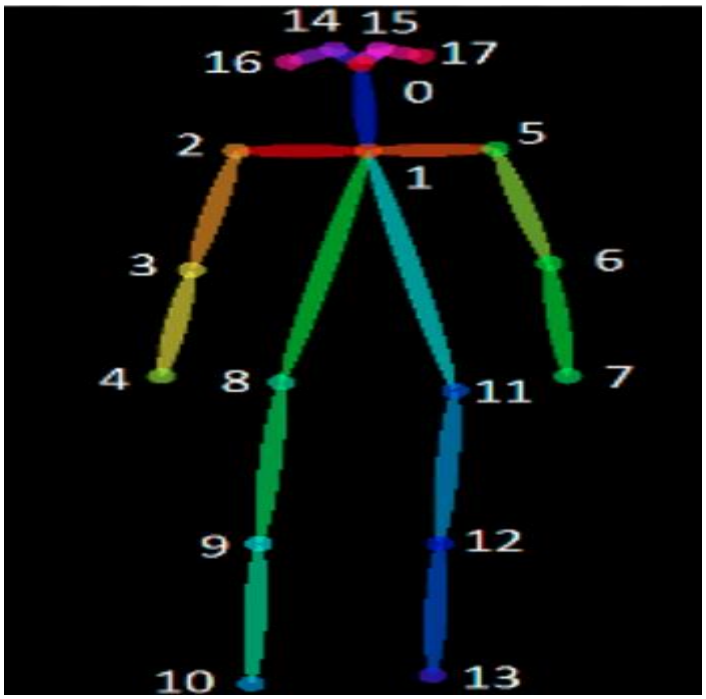

Figure 9: Keypoints in Coco Dataset [13]

The first branch of Openpose network predicts a set of confidence maps, these are mapped to the body parts of the person. These are called key-points which are used to detect different body parts and number of keypoints vary on the basis of the database which is used for training. These key points are mapped with keypoint-id, and each keypoint-id represents a specific body part as shown in figure 9 [13]

This system uses eighteen key-points as shown in figure 9; The total number of key-points in confidence maps is nineteen as one is used to represent the background of the image. This output from the PDC, of the key-points is generated and stored in json format, which is fed to GFC.

3.5. Garment Fitting Component

This component is responsible for fitting the garments on the 2D image provided in the input. To achieve this objective GFC makes use of CP-VTON+, a robust improvement for characteristic-preserving virtual try-on network [15]. The GFC takes input from the previous components and fits the clothes on the person in two stages. The first stage is called the Geometric Matching Module (GMM) which warps the clothes on target human image, and the second stage Try-On Module (TOM) is responsible for blending the garment output according to the human properties. The working of these two stages is explained in the below sections.

3.5.1. Geometric Matching Module

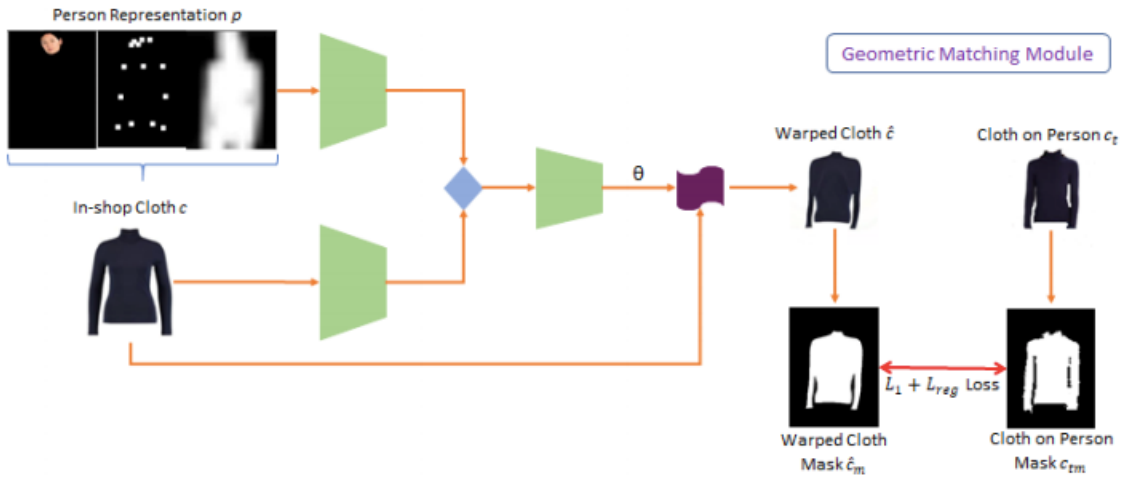

Figure 10: Geometric Matching Module Architecture [15]

The GMM carries out three processes - First is the preparation process, which works on the body pose, shape and cloth mask. For body pose it processes the body pose generated from the PDC, detects the key-points which are hands, neck, shoulder, etc. For body shape it uses the output generated from HSC and carries out the downsampling process to modify the resolution so as it matches the body shape and cloth mask generated from CMC. After this it slices the face

and hair of the person. Next is the synthesis process. In this process it performs down sampling of human body image and cloth mask image to recover vital information for correlation matching. It then carries out thin-plate spline (TPS) transfer to transform the image of the cloth on the front of the human body. The final stage of GMM is the comparison process. In this process it compares the garments that were generated in the synthesis stage to the clothes that were originally worn on the body [14][15].

3.5.2. Try-On Module

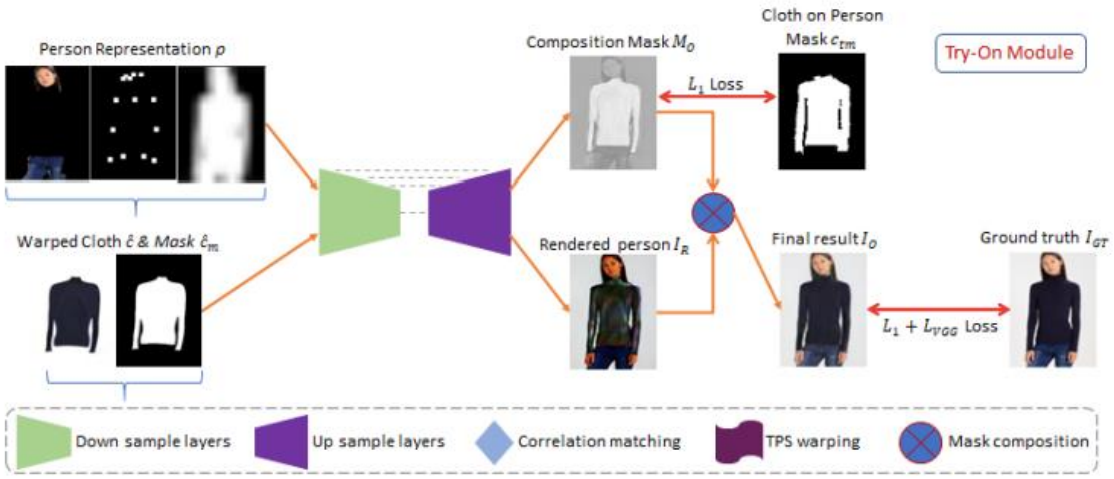

Figure 11: Try-On Module Architecture [15]

The GMM generates a warp cloth image that is roughly aligned with the body contour. TOM is in charge of blending this warp with the target human body in order to create the final try-on output. To achieve this TOM carries out two processes in conjunction with each other. Firstly it directly pastes the warped cloth onto the image of the target human and then it uses Unet to predict a composition mask for rendering smooth try-on results. In the final stage the UNet displays a human image while also predicting a composition mask. To create the final try-on outcome, the produced person picture and the warped garments are fused together using the composition mask and then it compares it with previous try-on output with the actual result by using the following loss function [14][15].

3.6. 3D-reconstruction component

The GFC generates the output in 2D format which needs to be reconstructed into 3D for better visualization. This component takes the 2D image and projects it into 3D by using PIFuHD (Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization) which is a comprehensive framework for inferring 3D geometry of 2D images in a pixel-aligned method while preserving the features of the original image [16]. The architecture and working of PiFuHD is described in the following section

3.6.1. PiFuHD

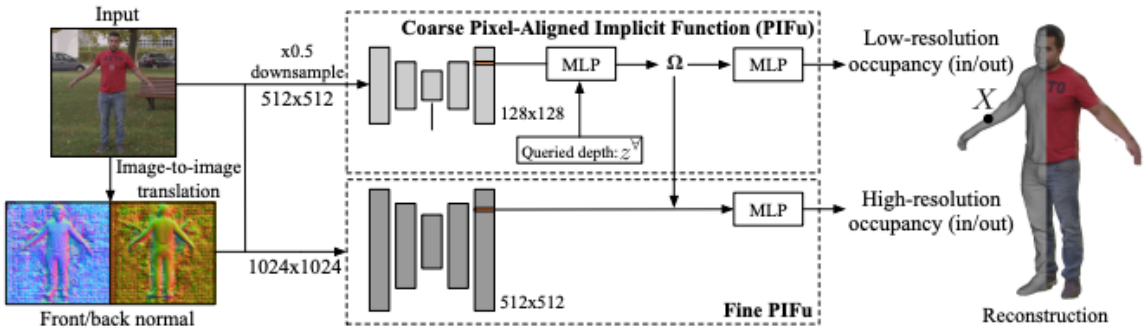

Figure 12: PiFuHD Architecture [16]

As shown in figure 12, the neural network follows a two-tier architecture and considers both the local and global features of the image. The network's architecture is made up of two stages of PIFu modules. First stage focuses on obtaining global features from a picture and the second stage is used for capturing local context information and adds precise detail to the 3D model. To acquire holistic reasoning, the model is trained on lower resolution, downscaled images to cover the greater spatial contexts of the image. The network then predicts the detailed geometry of the human body by examining the image and previous output on a higher-resolution using this contextual knowledge. The global 3D structure is captured at the coarse level by downsampling the image and feeding it into a PIFu model, while the high-resolution features are created by using those first 3D outputs as high-resolution inputs in a comparable

lightweight PIFu network.

3.6.2. Photogrammetry

The 3D mesh output generated from PiFuHD provides a comprehensive projection of the human model with the clothes fitted but does not transfer the texture. To transfer the texture of the cloth onto the generated 3D mesh UV mapping technique is used.

3.6.2.1. UV mapping

UV mapping is a technique used for transferring the texture from a 2D image to a 3D object. In this process the three coordinates of the mesh are unwrapped into two coordinates (U,V) as shown in *Fig 13 - UV Unwrapping [24]*. The texture from the 2D image generated from GFC is then mapped to the unwrapped mesh of the 3D object. This finally generates the 3D model of the person's image with the target garments which were provided in the GFC.

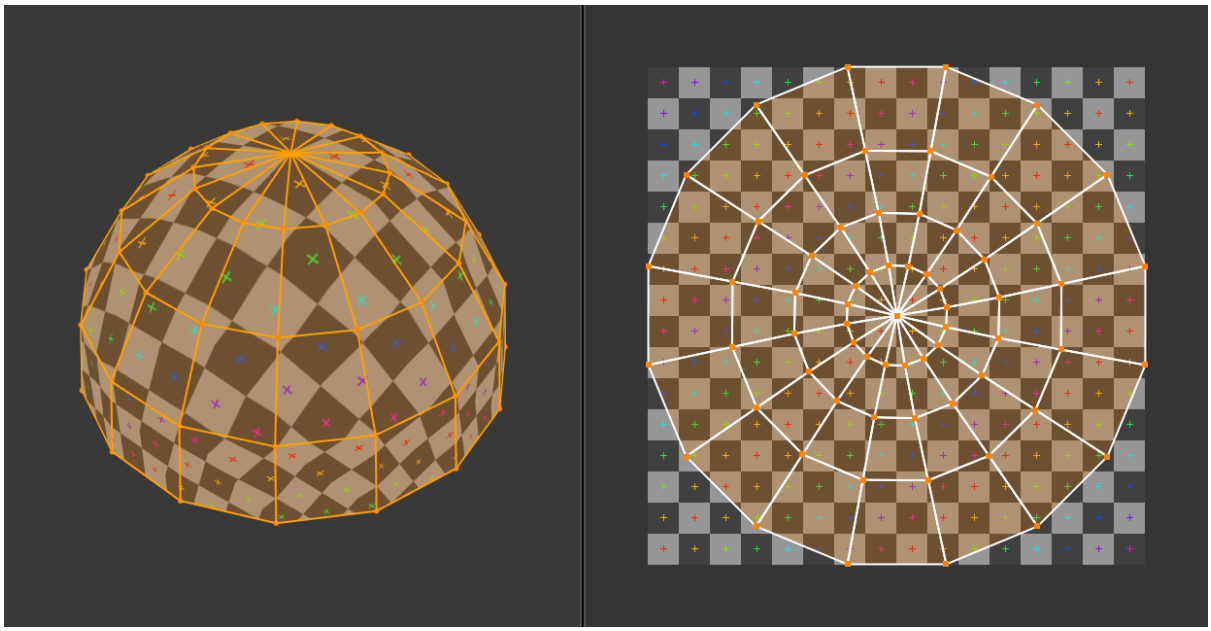

Figure 13: UV Unwrapping [24]

3.7. Justification for models' selection

For designing the pipeline, the most important model was the garment fitting network. As mentioned in section 2.4, there were a number of neural networks to choose which could have achieved the task. The most optimal choices considered were CP-VTON, CP-VTON+ and VOGUE StyleGAN. StyleGAN shows promising results when compared to CP-VTON as it uses interpolation techniques for pose projection [1]. The interpolation technique also has a drawback of generating distorted outputs when the pose projection of the person is unconventional, and hence the results are not good. CP-VTON+ makes improvements on the existing CP-VTON try-on network and outperforms it. Although it showcases better fitting and performance, the only limitation of this network is that it is designed for 2D human models [15]. Since, the pipeline uses 3D reconstruction from 2D images, it is able to overcome this limitation and hence this network was selected.

The body segmentation and pose estimation networks were required for feeding the input for this network. The body segmentation could be done by CIHP_PGN, LIP_JPPNet and Graphonomy. The GFC can work with any of these three networks. There is no major difference between the segmentation results generated from these networks. For LIP_JPPNet and Graphonomy, after the segmented image is generated, there is a need to perform more transformations on the image before feeding it to the GFC. In the case of CIHP_PGN there is no such need of doing further modifications and hence this network was chosen for HSC. For the pose estimation Openpose and DensePose were two options available. Although Densepose is better than OpenPose, it is quite new and there are various bugs in its implementation, it also does not provide any api which could be used directly in the pipeline and hence OpenPose was selected. With respect to the 3D reconstruction networks, PiFuHD was selected as the best choice because it preserves all the texture details of garments and outperforms other techniques as shown in the *Figure 14 - PiFuHD Comparison [16]*.

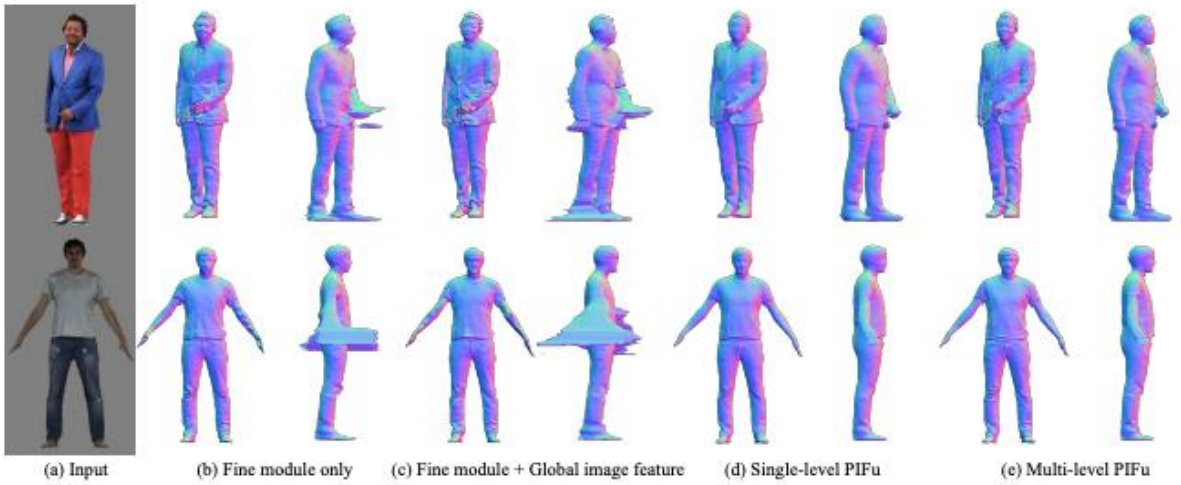

Figure 14: PiFuHD Comparison [16]

3.8. Implementation

This section explains the frameworks and tools used for implementation of the designed pipeline.

3.8.1. Frameworks and Technologies

The pipeline consists of a variety of deep neural networks which perform different tasks. Each neural network and their supporting functionality have its own frameworks and tools for implementation and executing it on a stand-alone personal computer would have been a challenging task considering the computational power required for these deep neural networks.

Each component of the pipeline had dependencies on various frameworks like TensorFlow, Pytorch and Caffe. Therefore, Google colaboratory was chosen for implementing the pipeline. Colab is a Jupyter hosting notebook technically, which provides free access to GPUs along with computing resources [21].

Entire pipeline is developed in python. For implementation of the CMC, libraries like OpenCV and Pillow are used. OpenCV offers in-built functions for carrying out the operations like thresholding, flood-filling, etc which are required for generating the cloth-mask. TensorFlow was used to implement the Human Segmentation Component, because the CIHP_PGN model was trained in that format. The model was imported and used for inference on the input human image. It generated segmentation and edge maps of the human body. The segmentation maps were used, and the other outputs were discarded, as the GFC does not require them.

For implementing PDC, the caffe framework was used as it is compatible with OpenPose and has proven to be faster and more efficient for image processing tasks. The output generated from the CMC, HSC and PDC were then fed to GFC, and the subsequent output generated from GFC was provided to 3RC for

reconstruction of the image.

For the final two components of the pipeline, Pytorch was used for implementation as it was compatible with both state-of-the-art deep neural networks that were used. To perform the photogrammetry operations on the generated 3D mesh, Blender was used. It made it easier to perform the transferring of texture using UV mapping technique.

Chapter 4 - Results and Evaluation

This chapter showcases the outputs generated from each component of the pipeline and provides a brief discussion on them. In the section 4.6, the results and performance of the pipeline is evaluated by using IoU and SSIM metrics, and qualitative evaluation of the generated 3D output is provided. The chapter concludes by discussing the failure scenarios and critical analysis of the results.

4.1. Cloth Masking

4.1.1. Results and Details

As shown in the figure 15, the input image is first converted to grayscale and then the four steps of creating the mask are applied. The RGB format of the image does not produce good results when it is passed to the GMM later, hence these pre-processing and masking steps are necessary.

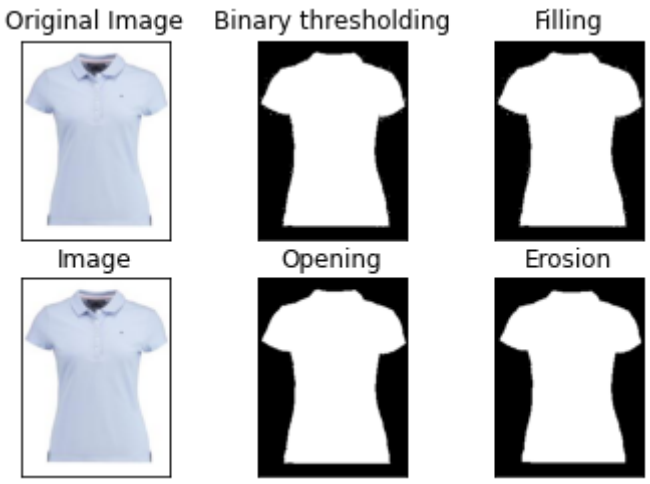

Figure 15: Cloth Masking Results

4.2. Human segmentation

4.2.1. Results and details

Figure 16 is the input image provided for the segmentation component and the results are displayed in figure 17. The first image in figure 17 shows the output produced by instance aware detection, which detects the boundaries of the person and separates them from the background. The second and third image in the output shows the body part segmentation result produced by the network in RGB and grayscale format. The segmented body parts are represented in different shades in the RGB and grayscale image.

Figure 16: Input Image for Try-on

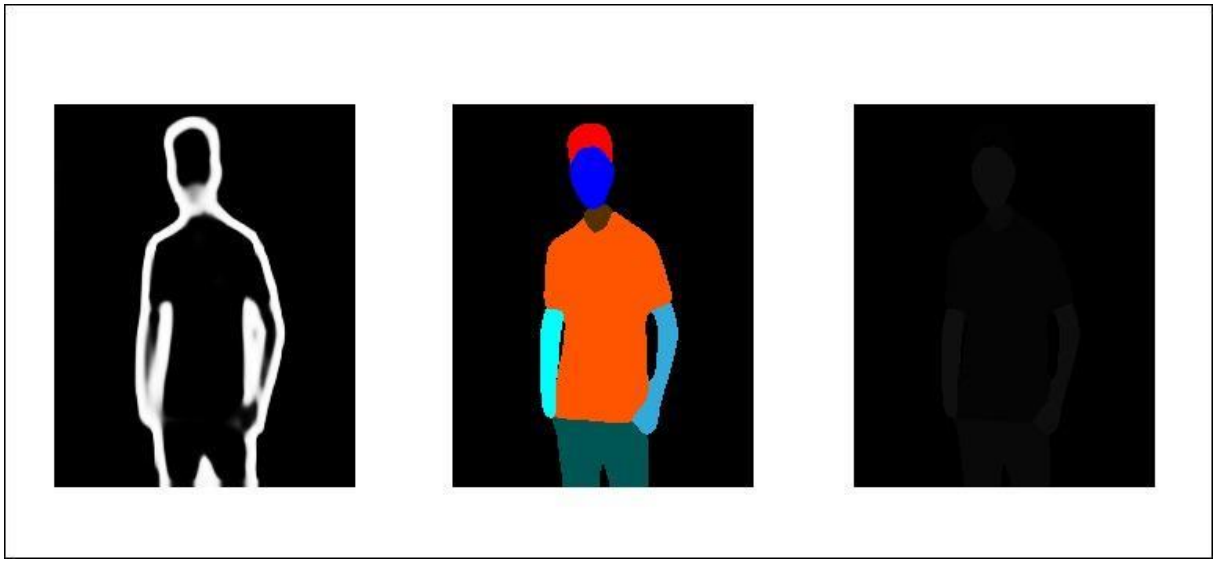

Figure 17: Segmentation Results

4.3. Pose detection

4.3.1. Results and details

The figure 18 shows the skeletal pose estimation generated by PDC. The joints of the human body which are shown in different colors represent the human skeleton.

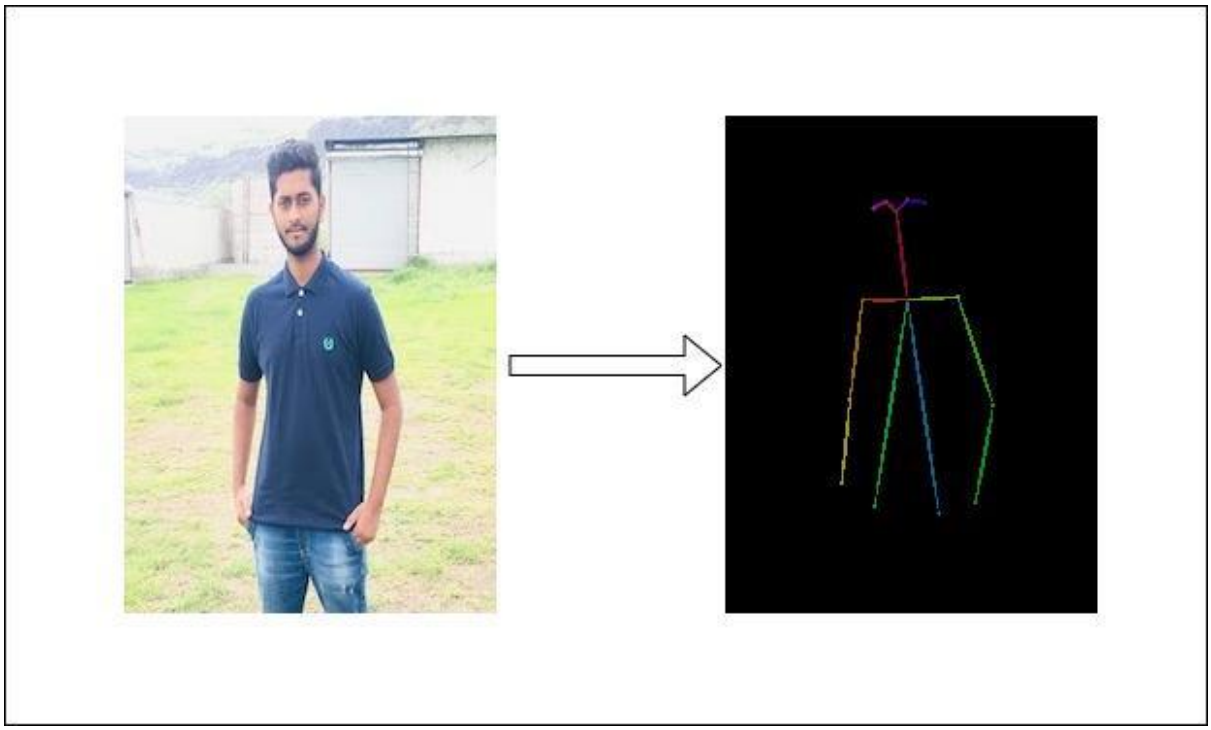

Figure 18: Pose Detection Keypoints

4.4. Virtual try-on

4.4.1. GMM results and details

The figure 19 shows the results of the first stage in GFC. As shown in figure, GMM first checks the pose of the target person and then adjusts the cloth mask accordingly. In the next stage it performs correlation matching between the target body and cloth image by disregarding the face and neck of the person. In the final stage it transforms the image of the cloth onto the body of the person.

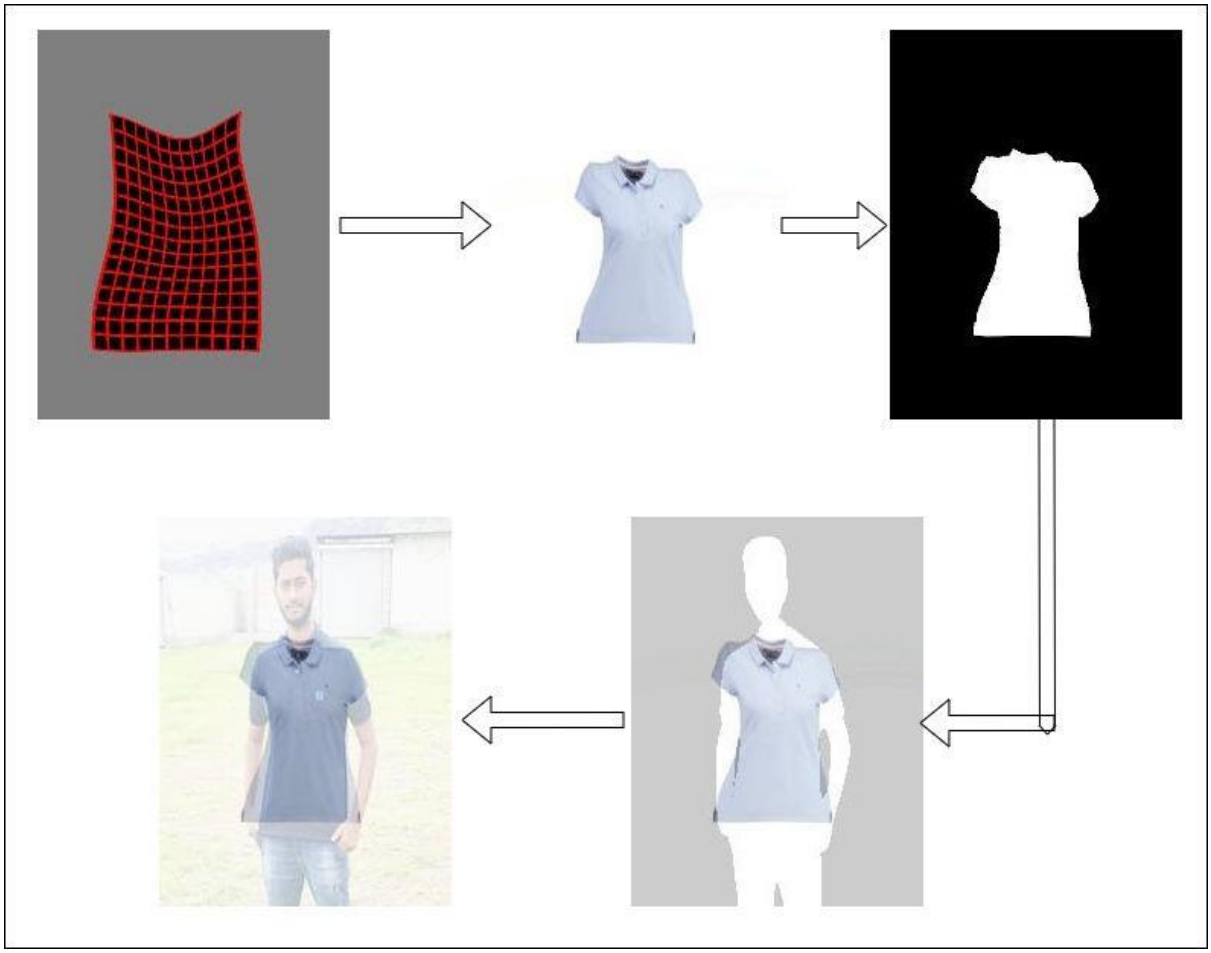

Figure 19: GMM Results

4.4.2. TOM results and details

In the figure 20, TOM output in the various stages is displayed. TOM processes the warped cloth image output provided by GMM, followed by generating a composition mask to produce smooth results.

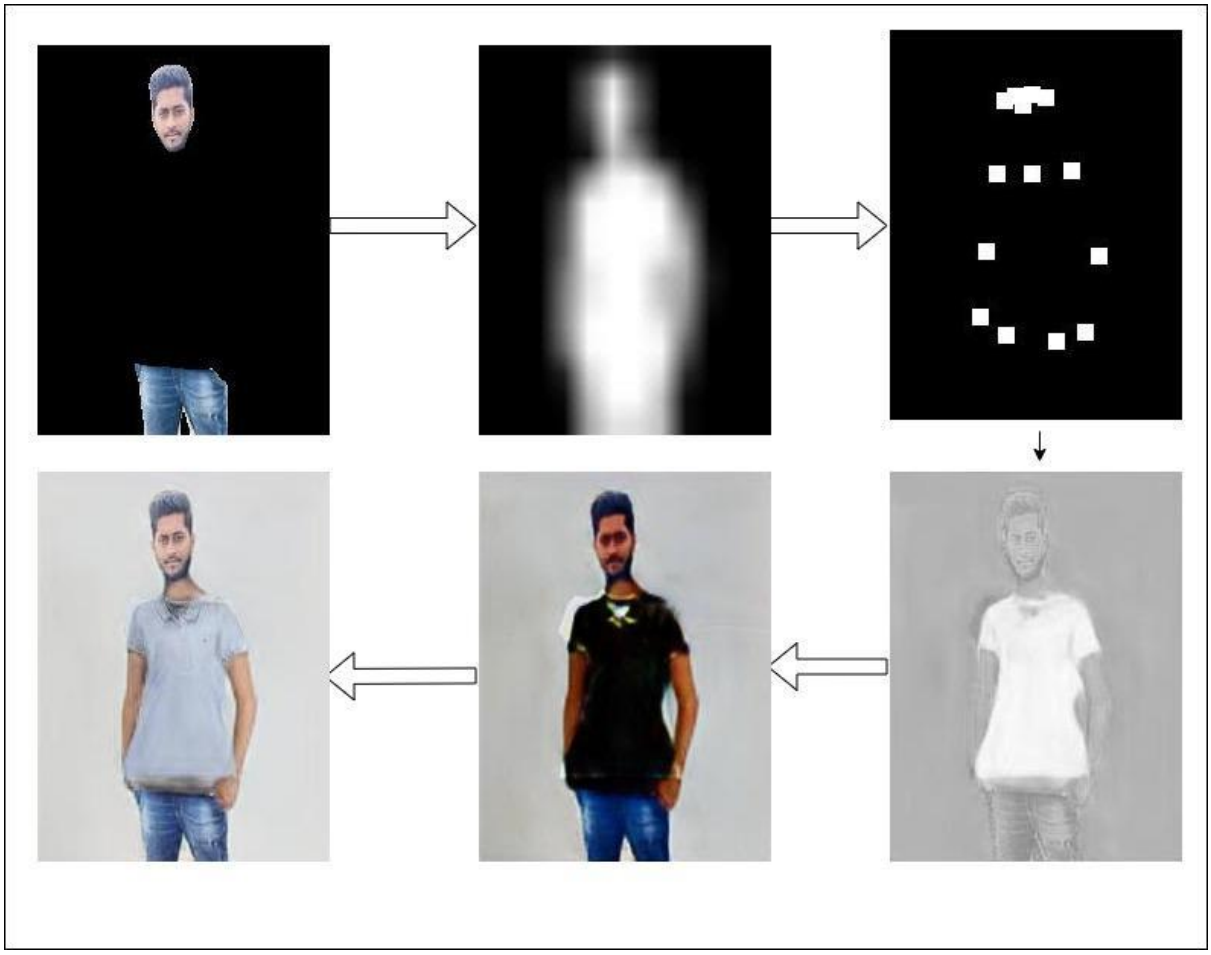

Figure 20: TOM Results

4.5. 3D reconstruction & photogrammetry

4.5.1. 3RC Results and details

The figure 21 shows the 3D reconstruction output generated by the PiFuHD network. This output is then used in blender for UV unwrapping as shown in figure 22. After performing the texture transfer the final try on result is presented in figure 23.

Figure 21: 3RC Results

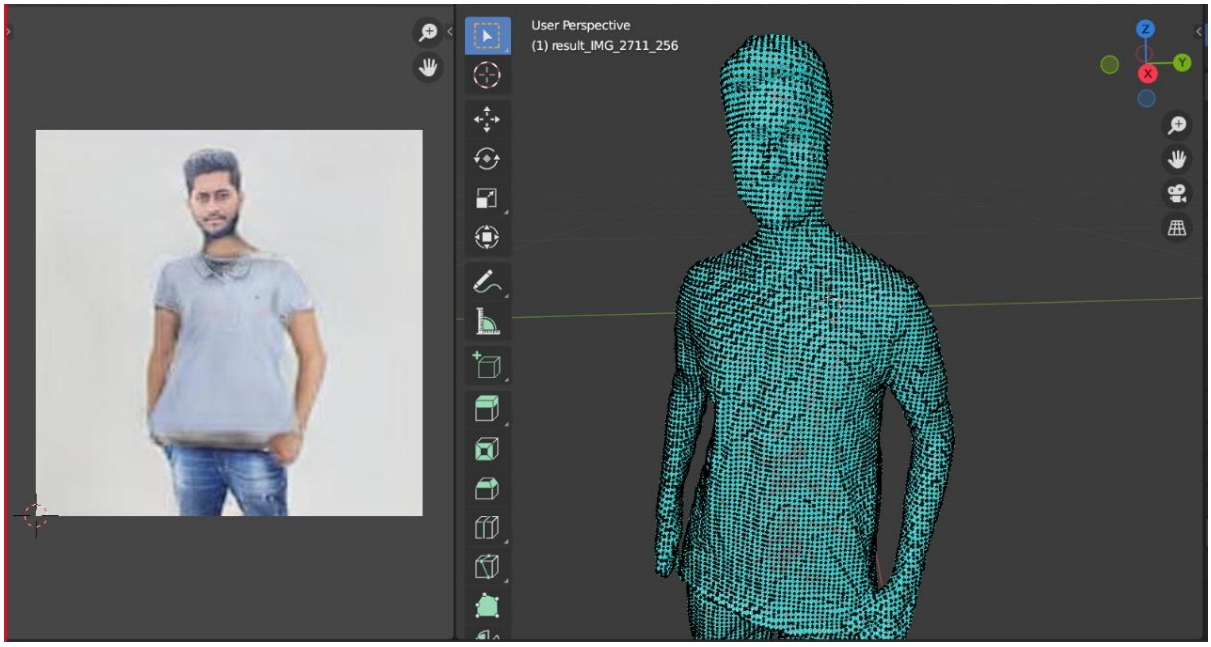

Figure 22: Photogrammetry UV Unwrapping


Figure 23: UV Texture Transfer

4.6. Evaluation

4.6.1. Quantitative evaluation

The most important metrics for deep learning with respect to image processing are Intersection over Union (IoU) and Structural similarity (SSIM).
IoU is used for calculating mean average precision that is used for determining the accuracy of object detection in comparison to the ground truth of the dataset annotations. The value of IoU ranges from 0 to 1, where 0 indicates that there is no overlap between the predictions and the ground truth while 1 indicates that the predictions and ground truth annotations are completely overlapped. IoU is calculated by the below equation -

$$IoU = \frac{(target \cap prediction)}{(target \cup prediction)}$$

SSIM is an evaluation metric which is used for measuring structural similarity on the basis of contrast, structure and luminance. The formula for SSIM is shown below, in which luminance, contrast and structure comparison functions respectively represent the weight of alpha, beta, and gamma [24].

$$\mathrm{SSIM}(I, \hat{I}) = [\mathcal{C}_l(I, \hat{I})]^{\alpha} [\mathcal{C}_c(I, \hat{I})]^{\beta} [\mathcal{C}_s(I, \hat{I})]^{\gamma}$$

The table 4.1 shows comparison of IoU and SSIM scores between CP-VTON+ and CP-VTON.

Table 1: Evaluation of GFC - Values taken from [15]

Network	IoU	SSIM
CP-VTON+	0.8425	0.8163
CP-VTON	0.7898	0.7798

The above table shows that CP-VTON+ outperforms the predecessor and makes accurate predictions for the class labels as compared to CP-VTON [15].

4.6.2. Qualitative Evaluation of the Generated Final Output

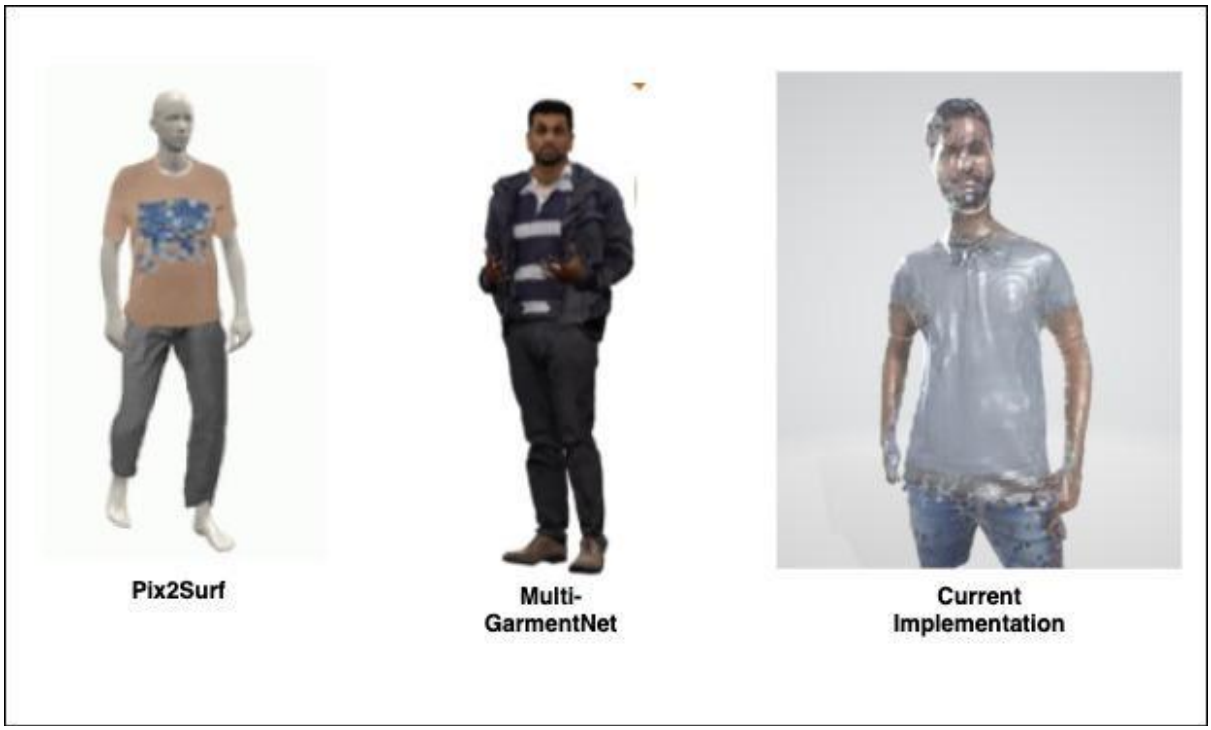

Figure 24: Qualitative Evaluation

As shown in figure 24, Pix2Surf and Multi-Garment net are the closest implementation of fitting garments on the 3D human models [7][3]. Pix2Surf does not retain the original properties of the human body like skin colour, hair, etc. It just simply fits the garments on the SMPL body mesh, while the approach followed in this project is able to inherit all the human body characteristics. Multi-Garment net shows promising results, but the network requires 3D scanned images of the clothes and target humans for pose detection. It also uses SMPL for generating the 3D mesh. The pipeline described in this project uses PiFuHD which is way better than the SMPL model and provides better results by retaining all the characteristics of the human body along with the clothes. The output generated in this implementation provides a simpler, better and efficient method for 3D virtual garment fitting as compared to Pix2Surf and Multi-garment net.

4.7. Failure Scenarios

Figure 25: GMM Failure {test image from VTON DB}

The garment fitting component fails to produce the expected output in some cases as shown in the figure 25. This is caused because the garments or poses in the training dataset were not well represented. Due to this the GMM is not able to adjust the cloth mask according to the pose of the person and thus results in producing an unexpected output.

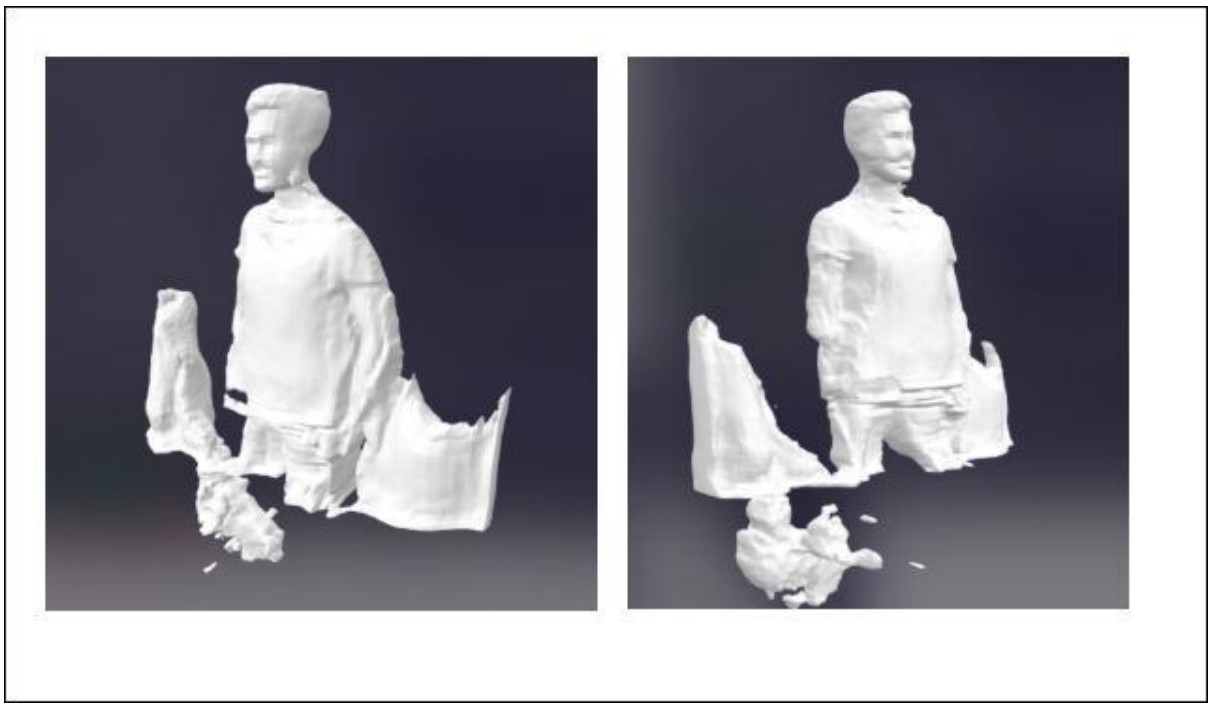

Figure 26: 3RC Failure

When using low quality images as an input, the 3D reconstruction of such images generates a lot of noise and 3RC is not able to separate the background from the object as shown in the figure 26. The input image goes through various neural networks for 2D garment fitting, due to which the quality of the image generated and fed to the 3RC is sometimes not up to the mark. This results in generating noisy data in the 3D mesh.

4.8. Critical Analysis

The output images produced by the individual components and the overall pipeline displayed in section 4.1 show promising results. With respect to cloth masking, body segmentation and pose estimation components the results are very good.

For the garment fitting component, the final output of the try-on result shows some distortions after the cloth is put on the target body. The reason for this distortion is because the network is trained on the images of the size 192 x 256 (width x height) pixels from the VTON dataset. Since the network is trained on low resolution of images, the input which is to be fed to this network must also be of the same size. On the other hand, the 3D reconstruction component uses PiFuHD which is supposed to handle inputs of higher resolution to give optimal results. There was no significant improvement in the result of the 3D reconstruction even after transforming the try-on output to a higher resolution before feeding it to PiFuHD. Since the VTON dataset could not provide high resolution images, the network could not be modified to train on higher quality of images.

In the output of the garment fitting component as seen in figure 20, the cloth image is overlapping on the left-hand of the person. The network is not able to differentiate between the clothing area and the human body skin colour. This could be solved by using better models in the body segmentation and pose estimation components of the pipeline.
The texture transfer using UV unwrapping technique works smoothly and has produced amazing results as seen in figure 23.

Chapter 5 - Conclusion

The research provides a unique approach to solve the challenge of 3D virtual garment fitting by using 3D reconstruction mechanisms for generating body shape with the garments while retaining the original characteristics of the person and texture of the clothes. On comparing the results with state-of-the-art methods, this system proves to produce qualitatively better results.

The major contributions of this project are –

- A novel pipeline is developed that comprises four neural networks performing different functions to work together and perform virtual garment fitting.
- The developed system can take a 2D image input of the person and the garment and show the result in 3D, retaining all the textures and body characteristics.
- The results showcased of this study prove to be more closer to the real world of trying clothes physically. This study provides a beginning step in a potential research direction.

The dissertation also shows some failure scenarios in the final try-on output generated by the pipeline. These scenarios can be handled by doing certain modifications in the pipeline.

Analysing the results of the 3D try-on module, it can be argued that although the system shows promising results based on the 2D approach it will have its limitations when the target humans are in varying poses.

5.1. Future Works

The developed pipeline can be improved in many ways to achieve a better quality result. The garment fitting component of the pipeline can be trained on a higher resolution of images, with people in various postures and more different types of apparels. This could improve the performance of the entire pipeline and provide better results generated from the 3D reconstruction component.
More testing is required to assess the performance of this pipeline, the testing must be carried out on people with different colours, hairstyles and in different poses with different types of garments.

The pipeline has different models for pose detection and garment fitting, which should not be the optimal method for clothing transfer. As both these networks are trained on different datasets, there would be failure scenarios where the garments are not aligned according to the pose of the person. To avoid such failure scenarios, ablation of pose estimation components should be considered, and that mechanism should be incorporated in the garment fitting component itself.

After the generation of output from the 3D reconstruction component, the UV unwrapping process is used for texture transfer. This process is done manually by using Blender. In the future, this process should be done automatically and the ways to do it must be explored.

In all the current virtual try-on systems for clothes, only visual output is shown to the user. The system should also use a metric for informing the user about the fit of the garment. This metric can be calculated by considering the model's accuracy and accordingly the user could see the visual output as well as get an idea about the fit of the apparel on his body.

Since the pipeline generates a 3D mesh, this work could also be extended in

Augmented Reality applications. Such applications can display the customer, how they look in different garments at different locations.

5.2 Summary

The dissertation begins by giving a brief idea about the concepts and methods required for virtual garment fitting. It then gives a general overview of the state-of-the-art approaches used for 2D and 3D virtual try-on. It provides a detail analysis of the methodologies used in such systems like pose estimation, body segmentation, 3D projection and the datasets that are used for garment fitting. It showcases the proposed pipeline and its individual components along with the details that justify the choice of the models used and an in-depth analysis of their working and implementation. It shows the results of the pipeline with a brief discussion and provides a quantitative and qualitative evaluation and comparison against the state-of-the-art models. It also discusses the failure scenarios and provides a critical analysis of the entire flow of the system. In the final chapter it concludes by showcasing the major contributions of this research, the possible modifications for getting better results and the future scope of such systems.

Bibliography

[1] Lewis, K., Varadharajan, S., and Kemelmacher-Shlizerman, I. 2021. VOGUE: Try-On by StyleGAN Interpolation Optimization. *arXiv preprint arXiv:2101.02285*.

[2] T. Liu, H. Xu and X. Zhang, "3D Clothing Transfer in Virtual Fitting Based on UV Mapping," 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018, pp. 1-6, doi: 10.1109/CISP-BMEI.2018.8633225.

[3]Bhatnagar, B., Tiwari, G., Theobalt, C., and Pons-Moll, G. 2019. Multi-Garment Net: Learning to Dress 3D People from Images. In *IEEE International Conference on Computer Vision (ICCV)*.

[4] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. (2019). Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation.

[5] Wei-Lin Hsiao, and Kristen Grauman. (2020). ViBE: Dressing for Diverse Body Shapes.

[6] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. (2020). Deep Fashion3D: A Dataset and Benchmark for 3D Garment Reconstruction from Single Images.

[7] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. (2020). Learning to Transfer Texture from Clothing Images to 3D Humans.

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. (2018). DensePose: Dense Human Pose Estimation In The Wild.

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. (2018). Instance-level Human Parsing via Part Grouping Network.

[10] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. (2018). Look into Person: Joint Body Parsing & Pose Estimation Network and A New Benchmark.

[11] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. (2017). Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing.

[12] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. (2019). Graphonomy: Universal Human Parsing via Graph Transfer Learning.

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.

[14] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. (2018). Toward Characteristic-Preserving Image-based Virtual Try-On Network.

[15] Minar, M., Thai Thanh Tuan, Ahn, H., Rosin, P., and Lai, Y.K. 2020. CP-VTON+: Clothing Shape and Texture Preserving Image-Based Virtual Try-On. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[16] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. (2020). PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization.

[17] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. (2019). DeepHuman: 3D Human Reconstruction from a Single Image.

[18] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. (2018). BodyNet: Volumetric Inference of 3D Human Body Shapes.

[19] Chen, Y., Tian, Y., and He, M. 2020. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding, 192*, p.102897.

[20] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. (2016). DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation.

[21] colab.research.google.com. (n.d.). *Google Colaboratory*. [online] Available at: https://colab.research.google.com/?utm_source=scs-index.

[22] Foundation, B. (n.d.). *Blender Institute*. [online] blender.org. Available at: https://www.blender.org/institute/.

[23] docs.blender.org. (n.d.). *Introduction — Blender Manual*. [online] Available at: https://docs.blender.org/manual/en/latest/editors/uv/introduction.html

[24]Zhihao Wang, Jian Chen, and Steven C. H. Hoi. (2020). Deep Learning for Image Super-resolution: A Survey.

****

[25] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. (2021). Parser-Free Virtual Try-on via Distilling Appearance Flows.

[26]Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. (2021). VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization.

[27] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. (2021). Disentangled Cycle Consistency for Highly-realistic Virtual Try-On.

[28] Kedan Li, Min jin Chong, Jeffrey Zhang, and Jingen Liu. (2021). Toward Accurate and Realistic Outfits Visualization with Attention to Details.

[29] Minar, M., Tuan, T., and Ahn, H. 2021. CloTH-VTON+: Clothing Three-dimensional reconstruction for Hybrid image-based Virtual Try-ON. *IEEE Access, PP*, p.1-1.

[30] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. (2020). Do Not Mask What You Do Not Need to Mask: a Parser-Free Virtual Try-On.

[31] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. (2020). Towards Photo-Realistic Virtual Try-On by Adaptively Generating Preserving Image Content.

[32] Neuberger, A., Borenstein, E., Hilleli, B., Oks, E., and Alpert, S. 2020. Image Based Virtual Try-On Network From Unpaired Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[33] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. (2020). Semantically Multi-modal Image Synthesis.

[34] Minar, M., Thai Thanh Tuan, Ahn, H., Rosin, P., and Lai, Y.K. 2020. CP-VTON+: Clothing Shape and Texture Preserving Image-Based Virtual Try-On. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[35] Minar, M., Thai Thanh Tuan, Ahn, H., Rosin, P., and Lai, Y.K. 2020. 3D Reconstruction of Clothes using a Human Body Model and its Application to Image-based Virtual Try-On. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[36] Yu, R., Wang, X., and Xie, X. 2019. VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[37] Han, X., Hu, X., Huang, W., and Scott, M. 2019. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[38] Kubo, S., Iwasawa, Y., Suzuki, M., and Matsuo, Y. 2019. UVTON: UV Mapping to Consider the 3D Structure of a Human in Image-Based Virtual Try-On Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[39] Lee, H., Lee, R., Kang, M., Cho, M., and Park, G. 2019. LA-VITON: A Network for Looking-Attractive Virtual Try-On. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[40] Ayush, K., Jandial, S., Chopra, A., Hemani, M., and Krishnamurthy, B. 2019. Robust Cloth Warping via Multi-Scale Patch Adversarial Loss for Virtual Try-On Framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[41] Ayush, K., Jandial, S., Chopra, A., and Krishnamurthy, B. 2019. Powering Virtual Try-On via Auxiliary Human Segmentation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[42] Yildirim, G., Jetchev, N., Vollgraf, R., and Bergmann, U. 2019. Generating High-Resolution Fashion Model Images Wearing Custom Outfits.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[43] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. (2018). Toward Characteristic-Preserving Image-based Virtual Try-On Network.

[44] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. (2018). VITON: An Image-based Virtual Try-on Network.

[45] Raj, A., Sangkloy, P., Chang, H., Hays, J., Ceylan, D., and Lu, J. 2018. SwapNet: Image based Garment transfer. *European Conference on Computer Vision, ECCV*.

[46] Zhenyu Xie, Xujie Zhang, Fuwei Zhao, Haoye Dong, Michael C. Kampffmeyer, Haonan Yan, and Xiaodan Liang. (2021). WAS-VTON: Warping Architecture Search for Virtual Try-on Network.

[47] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. (2021). Dressing in

Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-on and Outfit Editing.

[48] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip H. S. Torr, and Nicu Sebe. (2021). Cloth Interactive Transformer for Virtual Try-On.

[49] Minar, M., Tuan, T., and Ahn, H. 2021. CloTH-VTON+: Clothing Three-Dimensional Reconstruction for Hybrid Image-Based Virtual Try-ON. *IEEE Access, 9*, p.30960-30978.

[50]Fincato, M., Landi, F., Cornia, M., Cesari, F., and Cucchiara, R. 2021. VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 7669-7676).

[51] Lewis, K., Varadharajan, S., and Kemelmacher-Shlizerman, I. 2021. VOGUE: Try-On by StyleGAN Interpolation Optimization. *arXiv preprint arXiv:2101.02285*.

[52] Debapriya Roy, Sanchayan Santra, and Bhabatosh Chanda. (2020). LGVTON: A Landmark Guided Approach to Virtual Try-On.

[53] Sun, F., Guo, J., Su, Z., and Gao, C. 2019. Image-Based Virtual Try-on Network with Structural Coherence. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 519-523).

[54] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. (2019). M2E-Try On Net: Fashion from Model to Everyone.

[55] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. (2019). End-to-End Learning of Geometric Deformations of Feature Maps for Virtual Try-On.

[56] X. Gao, Z. Liu, Z. Feng, C. Shen, K. Ou, H. Tang, and M. Song 2021. Shape Controllable Virtual Try-on for Underwear Models. *arXiv:2107.13156*.

[57] Kedan Li, Min Jin Chong, Jingen Liu, and David Forsyth. (2020). Toward Accurate and Realistic Virtual Try-on Through Shape Matching and Multiple Warps.

[58] Amir Hossein Raffiee, and Michael Sollami. (2020). GarmentGAN: Photo-realistic Adversarial Fashion Transfer.

[59]Jandial, S., Chopra, A., Ayush, K., Hemani, M., Krishnamurthy, B., and Halwai, A. 2020. SieveNet: A Unified Framework for Robust Image-Based Virtual Try-On. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[60]Yu, L., Zhong, Y., and Wang, X. 2019. Inpainting-Based Virtual Try-on Network for Selective Garment Transfer. *IEEE Access, 7*, p.134125-134136.

[61] Zhou, Zhenglong, Bo Shu, Shaojie Zhuo, Xiaoming Deng, Ping Tan, and Stephen Lin. "Image-based clothes animation for virtual fitting." In SIGGRAPH Asia 2012 Technical Briefs, p. 33. ACM, 2012.

[62] Pachoulakis, Ioannis, and Kostas Kapetanakis. "Augmented reality platforms for virtual fitting rooms." The International Journal of Multimedia & Its Applications 4, no. 4 (2012): 35.

[63] Ting L, Lingzhi L, Xiwen Z. "Real-time 3D virtual dressing based on users' skeletons." In 2017 International Conference on Systems and Informatics, pp. 1378-1382. IEEE, 2017.

[64] Pereira, Francisco, Catarina Silva, and Mário Alves. "Virtual fitting room augmented reality techniques for e-commerce." In International conference on ENTERprise information systems, pp. 62-71. Springer, Berlin, Heidelberg, 2011.

[65] Mo Y I, Lee S G, Chang W. Method for converting 2d image into pseudo 3d image and user-adapted total coordination method in use artificial intelligence, and service business method[P]. U.S. Patent Application:10/583,160,2005-12-5.

[66] Isıkdogan, Furkan, and Gokcehan Kara. "A real time virtual dressing room application using kinect." CMPE537 Computer Vision Course Project (2012).

[67] Izadi, Shahram, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton et al. "KinectFusion: realtime 3D reconstruction and interaction using a moving depth camera." In Proceedings of the 24th annual ACM symposium on User interface software and technology, pp. 559-568. ACM, 2011.

[68] Li, Lu, Zhenjiang Miao, and Mangui Liang. "3D reconstruction based on Kinect." In Signal Processing (ICSP), 2014 12th International Conference on, pp. 1247-1250. IEEE, 2014.

[69] Yang, Bo, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trigoni. "3d object reconstruction from a single depth view with adversarial learning." arXiv preprint arXiv:1708.07969 (2017).

[70] Lin, Chen-Hsuan, Chen Kong, and Simon Lucey. "Learning efficient point cloud generation for dense 3D object reconstruction." arXiv preprint arXiv:1706.07036 (2017).

[71] Kurenkov, Andrey, Jingwei Ji, Animesh Garg, Viraj Mehta, JunYoung Gwak, Christopher Choy, and Silvio Savarese. "DeformNet: Free-form deformation network for 3d shape reconstruction from a single image." In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 858-866. IEEE, 2018.

[72] Sun, Xingyuan, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. "Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2974-2983. 2018.

[73] Chang, Angel X., Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese et al. "Shapenet: An informationrich 3d model repository." arXiv preprint arXiv:1512.03012 (2015).

[74] Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild." arXiv preprint arXiv:1802.00434 (2018).

[75] Kaur, Parneet, Hang Zhang, and Kristin J. Dana. "Photo-realistic Facial Texture Transfer." arXiv preprint arXiv:1706.04306 (2017).

[76] Mertens, Tom, Jan Kautz, Jiawen Chen, Philippe Bekaert, and Frédo Durand. "Texture Transfer Using Geometry Correlation." Rendering Techniques 273 (2006).

[77] Freeman, William T., and Alexei Efros. "Texture synthesis and transfer for pixel images." U.S. Patent 6,919,903, issued July 19, 2005.

[78] Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A skinned multi-person linear model." ACM Transactions on Graphics (TOG) 34, no. 6 (2015): 248.

[79] Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[80] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, pp. 91-99. 2015.

[81] Lin, Tsung-Yi, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. "Feature Pyramid Networks for Object Detection." In CVPR, vol. 1, no. 2, p. 3. 2017.

[82] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Computer Vision (ICCV), 2017 IEEE International Conference on, pp. 2980-2988. IEEE, 2017.

[83] Varol, Gül, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. "Learning from synthetic humans." In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 4627-4635. IEEE, 2017.

[84] Tilley, Alvin R. The measure of man and woman: human factors in design. Vol. 1. John Wiley & Sons, 2002.

[85] WrapR - a UV mapping tool: https://wrap-r.com/

Appendices

Abbreviations	Expansion
CMC	Cloth Masking Component
HSC	Human Segmentation Component
PDC	Pose Detection Component
GFC	Garment fitting Component
RC3	3D - Reconstruction Component
CIHP	Crowd Instance-level Human Parsing
PGN	Part Grouping Network
CIHP_PGN	Crowd Instance level Human Parsing-Part Grouping Network
CP-VTON	Clothing Shape and Texture Preserving Image-Based Virtual Try-On
TPS	Thin plate spline
PIFuHD	Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization
IoU	Intersection over Union
SSIM	Structural similarity