VIRTUAL TRIAL ROOM

A Project Report

Submitted by

MANOJKUMAR V K [CB.EN.U4CSE17040]
HASWANTH [CB.EN.U4CSE17020]
RUTHVIK S [CB.EN.U4CSE17060]
RAMMOHAN [CB.EN.U4CSE17063]

Under the guidance of

Dr. (Col.) Kumar P. N., Ph.D
(Chairperson, Department of Computer Sciene & Engineering)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

त्यागेनैके अमृतत्वमानशुः
MATA AMRITANANDAMAYI MATH
श्रद्धावान् लभते ज्ञानम्

Amrita Nagar PO, Coimbatore - 641 112, Tamilnadu

May 2021

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE – 641 112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **VIRTUAL TRIAL ROOM** is submitted by Manojkumar V K (CB.EN.U4CSE17040), G Haswanth (CB.EN.U4CSE17020), Ruthvik (CB.EN.U4CSE17060), Rammohan (CB.EN.U4CSE17063) in partial fulfillment of the requirements for the award of the Degree Bachelor of Technology in Computer Science and Engineering. It is a bonafide record of the work carried out under the guidance and supervision of the Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore.

SIGNATURE	**SIGNATURE**
Dr. (Col.) Kumar P. N., Ph.D PROJECT GUIDE Chairperson Dept. of Computer Sciene & Engineering	Dr. (Col.) P. N. Kumar CHAIRPERSON Dept. of Computer Science & Engineering
Signature of the Internal Examiner	Signature of the External Examiner

DECLARATION

We, the undersigned, solemnly declare that the project report **VIRTUAL TRIAL ROOM** is based on our own work carried out during the course of our study under the supervision of Dr. (Col.) Kumar P. N., Ph.D, Chairperson, Computer Sciene & Engineering, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgement has been made wherever the findings of others have been cited.

Manojkumar V K[CB.EN.U4CSE17040]

Haswanth [CB.EN.U4CSE17020]

Ruthvik S[CB.EN.U4CSE17060]

T. Rammohan.

Rammohan[CB.EN.U4CSE17063]

ABSTRACT

With the advancements in technology, the use of e-commerce has skyrocketed. It is estimated that in 2019, 1.92 billion people purchased goods or services online [(William et al., 2020)]. Suprisingly, Fashion industry is the top-selling industry across the globe. Despite the trickiness and the plethora of shapes and sizes in the industry, fashion still remains the queen of sales [(Smith et al., 2020)]. However, in reference to the study by Reagan et al. (2016), it has been proven that there is a high rate of clothes being returned due to misfit or dissatisfaction with the apparels. According to Reagan et al. (2016), over 60% of the clothes are being returned, resulting in $10billion in extra operational cost. This highlights one of the major drawbacks of the fashion industry. This project provides a crucial feature as an add-on service to the online retailers and e-commerce platforms like Amazon, Myntra, Flipkart, and Ajio, allowing customers to virtually try-on the clothes. It is done so by providing a virtual trial room experience and thereby, improving the online shopping experience and reducing the number of returns. We have developed a Two dimensional (2D) image based try-on network to virtually project the desired clothing item onto the corresponding region of a person using a coarse-to-fine strategy without using 3D information of any form. As a result of our suggested program, we believe that, our targeted market segments, will be greatly benefitted due to the improvement in efficiency and consumer's online shopping experience.

ACKNOWLEDGEMENTS

We would like to express our deep gratitude to our beloved Satguru **Sri Mata Amritanandamayi Devi** for providing the bright academic climate at this university. Their support, has made this entire task appreciable. This acknowledgement is intended to appreciate everyone involved directly or indirectly in our project. We would like to thank our Pro Chancellor **Bramachari Abhayamrita Chaitanya**, Vice Chancellor **Dr. Venkat Rangan. P** and **Dr. Sasangan Ramanathan**, Dean Engineering of Amrita Vishwa Vidyapeetham for providing us the necessary infrastructure required for the completion of the project. We also would like to convey our thanks to **Dr. (Col) P.N. Kumar**, Chairperson of Department of Computer Science Engineering, **Dr. G. Jeyakumar and Dr. C. Shunmuga Velayutham,** Vice Chairpersons of the Department of Computer Science and Engineering for their valuable help and support during our study. We express our gratitude to our guide, **Dr. (Col) P.N. Kumar**, Chairperson of Department of Computer Science Engineering, for his guidance, support and supervision. We feel extremely grateful to **Mr. Sumesh A. K.**, Assistant Professor, Department of Computer Science Engineering, **Dr. Senthilkumar M.**, Assistant Professor, Department of Computer Science Engineering and **Mr. Baskar A.**, Assistant Professor, Department of Computer Science Engineering, for their feedback and encouragement which helped us to complete the project. We would also like to show our appreciation to the entire staff of the Department of Computer Science and Engineering. We would like to extend our sincere thanks to our family and friends for helping and motivating us during the course of the project. Finally, we would like to thank all those who have helped, guided and encouraged us directly or indirectly during the project work. Last but not the least, we thank God for His blessings which made our project a success.

Manojkumar V K	Haswanth	Ruthvik	Rammohan
CB.EN.U4CSE17040	CB.EN.U4CSE17020	CB.EN.U4CSE17060	CB.EN.U4CSE17063

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

UI	User Interface
API	Application Programming Interface
3D	Three dimensional
2D	Two dimensional
CNN	Convolutional Neural Networks
RCNN	Region based Convolutional Neural Networks
ML	Machine Learning
IDE	Integrated Development Environment
I/O	Input/Output
ER	Entity Relationship
VITON	Image based Virtual Try-on Network
MH-Parser	Multi-Human Parser
GAN	Generative Adversarial Networks
CE2P	Context Embedding with Edge Perceiving
AR	Augmented Reality
JSON	JavaScript Object Notation
OS	Operating System
PAF	Part Affinity Field
ASPP	Atrous Spatial Pyramid Pooling
IS	Inception Score

List of Symbols

α	Binary Mask generated in the refinement network
F	Feature Map
ϕ	Convolutional Neural Networks (CNN) in predicting Part Affinity Field (PAF)
ρ	CNN in predicting confidence maps
λ	Hyperparameter
L	PAF
S	Confidence Maps
W	Binary Mask used in Pose Estimation
p	Pixel location
f	Loss function
C_I^G	List of parsed regions
L_p	Pixel-wise loss
z	Input vector
I	Input image or user image
I'	Intermediate image generated
$\hat{I}$	Target image
c	Selected clothing item
G_C	Function approximated by the Encoder-Decoder Generator
L_1	L_1 loss
L_{G_C}	Loss function of the generator

Chapter 1

INTRODUCTION

"Time is more valuable than money" - Jim Rohn. With the convenience of online shopping, it indeed has helped us save a lot of time. Be it groceries, sophisticated electronics or clothes and apparels - with advancement in technology and globalisation, we are able to shop them online at our own comfort. However, there are still some drawbacks to the online shopping experience. One such drawback for online shopping is the slowness of the system, affecting the overall consumer experience. There has yet to be deployment of any potential technologies to improve the user's experience. Hence, users are returning a large amount of purchased clothes owing to misalignment of their expectation of the purchased goods. Our report will be focusing on finding solutions to this major drawback of the online shopping experience, on both consumer and service provider's side.

1.1 Problem Definition

A recent study by Reagan et al. (2016) annotated 60% of the returned clothes from online purchases are due to style issues leading to e-commerce companies spending over $10B extra on operating costs. To tackle this problem, there are existing works based on Three dimensional (3D) models or avatars on which the selected clothes can be trialled upon. However, the extreme difficulty and high costs in 3D model-based approaches, makes it difficult to be implemented in reality. Thus, making 2D approaches popular.
The pipeline for the project can be defined in three stages:

1. Estimate the pose from the user's image
2. Find the corresponding region of the user image where the cloth/apparel has to be fitted.
3. Finally, using a U-net architecture based encoder-decoder network to perform the actual try-on.

This 2D approach could engage a larger audience and potentially reduce expenditure in business.

Chapter 2

LITERATURE SURVEY

2.1 Multi-Human Parsing in the Wild

Li et al. (2017) introduced a novel Multi-Human Parsing model named Multi-Human Parser (MH-Parser). This newly introduced parser makes use of a Graph-based Generative Adversarial Network model to generate global parsing maps and person masks in a bottom-up fashion.

2.2 Towards Accurate Single and Multiple Human Parsing

Ruan et al. (2019) proposes a simple and effective Context Embedding with Edge Perceiving (Context Embedding with Edge Perceiving (CE2P)) framework for human parsing by identifying several useful properties like edge details, feature resolution, etc.

2.3 A CNN Model for Human Parsing Based on Capacity Optimization

Jiang and Chi (2019) introduced a depth-estimation module to improve the robustness of the human parsing model. The authors implemented the model using a stack of sublayers over conventional CNN layers. The results indicates that the integration of these modules improves human parsing without additional labelling.

2.4 Look into Person: Joint Body Parsing & Pose Estimation Network and A New Benchmark

Liang et al. (2018) The novel joint human parsing network includes the iterative refinement of location and multiple feature connections in an end-to-end framework. The Framework investigates the efficient context modelling and, thus, enabling human parsing tasks and pose estimation tasks which are mutually beneficial to each other. This unified framework achieves state-of-the-art performance for both human parsings and poses estimation tasks.

2.5 Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

Cao et al. (2017) put-forth an approach to efficiently detect the poses of multiple people in an image. The authors have developed an architecture so as to jointly learn the location of each part and their association via two branches of the same sequential prediction process.

2.6 Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods

Chen et al. (2020) have performed an extensive review of the latest deep learning technologies which are currently employed for human pose detection. The review also summarises existing frameworks, benchmark datasets and so on.

2.7 3D Virtual Trial Room

Ramesh et al. (2018) developed an Augmented Reality (AR)-based 3D virtual trial room using OpenCV. The authors have used image processing concepts like colour bounding analysis and grayscaling over machine learning based approaches. A live video is recorded where the user is capable of changing the colour of clothes and adding additional logos.

2.8 Real-time virtual fitting with body measurement and motion smoothing

Gültepe et al. (2014) put forth a novel virtual trial room framework using the depth sensor using customized motion sensors, physical simulation and size alterations. The proposed framework makes use of the proprietary PhysX engine by Nvidia. A 3D avatar is simulated based on the user's measurements and the selected apparel mesh is overlaid on top of this avatar.

2.9 Tendency to Use the Virtual Fitting Room in Generation Y - Results of Qualitative Study

Moroz et al. (2019) has performed exploratory research on both 2D overlay and 3D mannequin based approaches used in existing Virtual Fitting room applications. The results show that 2D based approaches have greater market opportunities compared to 3D approaches.

2.10 Image based-Virtual Fitting Room

Jie Chen et al. (2019) proposes a novel approach for Virtual Fitting Room. The authors initially identifies the regions of different fashion items using Mask Region based Convolutional Neural Networks (RCNN) and then uses neural style transfer to change the style of the selected fashion items. Out of the 8 models they had trained, the authors claim that their best model outperformed existing solutions.

2.11 Virtual Dressing Room Application

Masri et al. (2019) used separate modules to find the joint locations for rotation, positioning and scaling. These informations are utilized to align the 2D cloth models with the user. Skin colour detection is also applied to handle occlusions. In the final step, the model is overlaid on the user in real-time.

2.12 VITON: An Image-based Virtual Try-on Network

Han et al. (2018) present a 2D image-based virtual try-on network which generates a coarse synthesized image with the target clothing item overlaid on that same person in the same pose. The authors claim that their model seems to perform in comparison to the existing state of the art Generative Models.

2.13 Summary

Virtual Trial Room/virtual fitting room is a relatively new problem domain. Most of the existing solutions use 3D based approaches. However, from Moroz et al. (2019), we can infer that 2D approaches are preferred over 3D approaches mainly because of the cost and complexity in the real-time implementation. However, existing state-of-the start solutions still face issues when it comes to clothes with intricate designs and patterns. The project thus undertaken aims to resolve this by improving the existing solutions.

2.14 Data Set

Source of the datasets: LIP: Overview & VITON dataset

The LIP dataset contains 50,000 images with elaborated pixel-wise annotations with 19 semantic human part labels and 2D human poses with 16 key points.

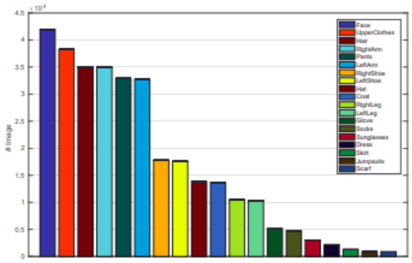


Figure 2.1: LIP Data distribution on 19 part labels

The Image based Virtual Try-on Network (VITON) dataset consists of over 17,000 images of clothes and humans along with their corresponding masks. It also has a pose estimate as JavaScript Object Notation (JSON) file with the key points denoting each region. This was the only dataset that was publicly available for use.

2.15 Software/Tools Requirements

- Python - High-level and general-purpose programming language
- Scikit-learn - For data preprocessing
- Keras - Open source Neural Network library
- Tensorflow - Open source library used for machine learning applications
- PyTorch - Open Source for Deep Learning applications
- OpenCV - Python library consisting of computer vision functions
- Jupyter - Interactive Data Science Environment
- Flask - A micro web framework written in Python
- MySQL - For database

Figure 2.2: Softwares/Tools used in the Project

Chapter 3

PROPOSED SYSTEM

3.1 System Analysis

The **VIRTUAL TRIAL ROOM** project can be broken down into five modules:

S.No	Module
1	Understanding the problem at hand
2	Pose Estimation
3	Human Segmentation
4	Try-on network
5	User Interface & Integration

Table 3.1: Module Details

3.1.1 System requirement analysis

The designed website, when the user makes a request to the Application Programming Interface (API), that is responsible for the *try-on*, invokes shell/bash scripts to run the model. Therefore a linux based system is preferred. Other requirements are:

- **Operating System (OS)** : Any linux based OS
- **Python** : 2.7+
  - ***Flask*** : 1.0.2+ (For User Interface (UI))
  - ***Keras*** : 2.1.6+ (Interface for neural nets)
  - ***Numpy*** : 1.14.3+ (For vectorization)
  - ***Pillow*** : 5.1.0+ (Image manipulation)
  - ***Tensorflow*** : 1.3.0+ (Library for Machine Learning (ML))
- **MySQL** : 5.7+
- Any Integrated Development Environment (IDE) can be used

A high-end system is generally preferred to deploy the model as it heavily loads the computer.

3.1.2 Module details of the system

Analyse existing solutions

The problem at hand was studied profoundly. A detailed literature survey has been performed and various existing approaches were analysed. The state-of-the-art solutions were also investigated with the intent of improving the current available solutions.

Pose Estimation

Given a user's image, a rough estimation of the user's pose would be generated. This module makes use of a modified version of the CVPR 2017 winner's model. It is note-worthy that the model used by Cao et al. (2017) is capable of estimating poses of multiple persons in a single image. However, this feature is not that practical as generally the users do not *try* the clothes in groups.

Human Segmentation

This module aims to find the region in the image where the cloth/apparel selected by the user has to be fit. It makes use of the pose information to parse the corresponding region of interest.

Try-on Network

A neural network is built from scratch to perform the actual try-on by making use of the information from the Human Parsing and Pose Estimation modules in order to get the target image. A UNet based Encoder-Decoder network is used for the same.

Prototype

A basic e-commerce website is built using Flask and MySQL to demonstrate the real-time use of the model.

3.2 System Design

3.2.1 Flow diagram of the system

The top-level data flow along with the module wise details are shown below.

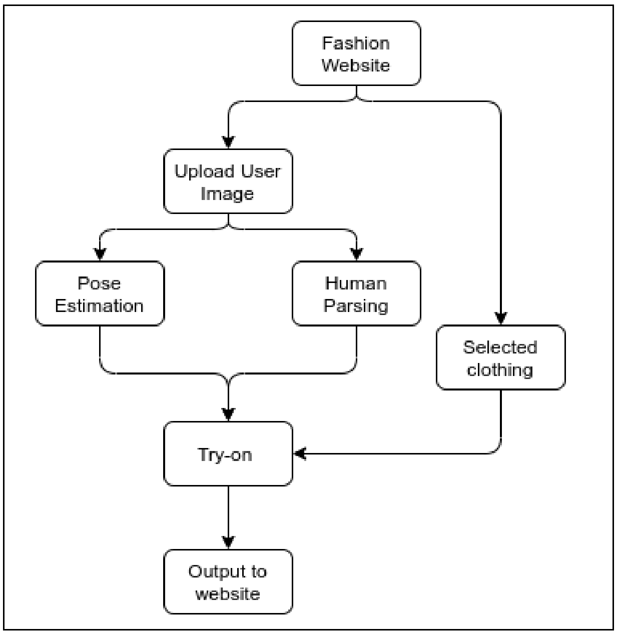


Figure 3.1: Flow Diagram of the system

Pose Estimation

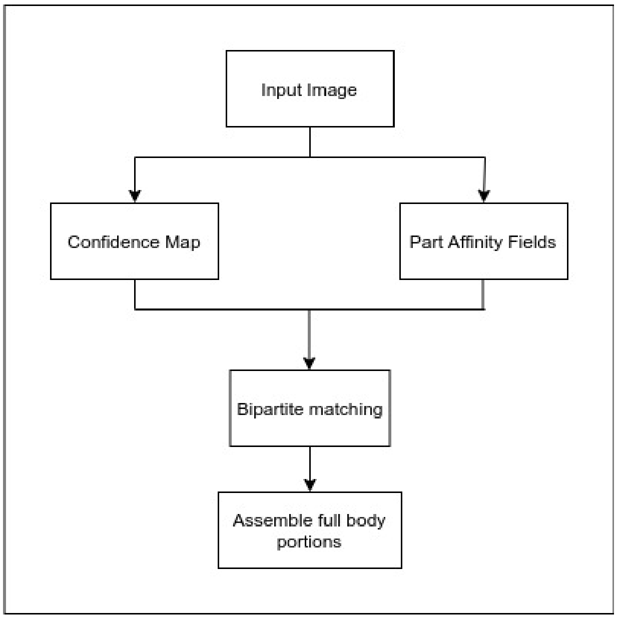


Figure 3.2: Data Flow in the Pose Estimation module

Human Parsing

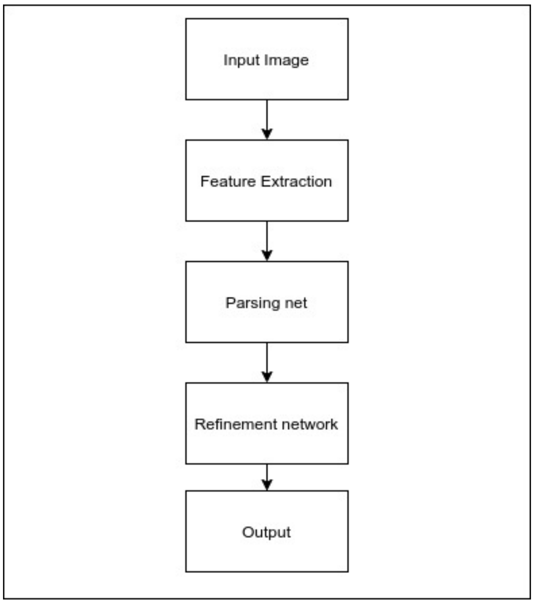


Figure 3.3: Data Flow in the Human Parsing module

TryOn Module

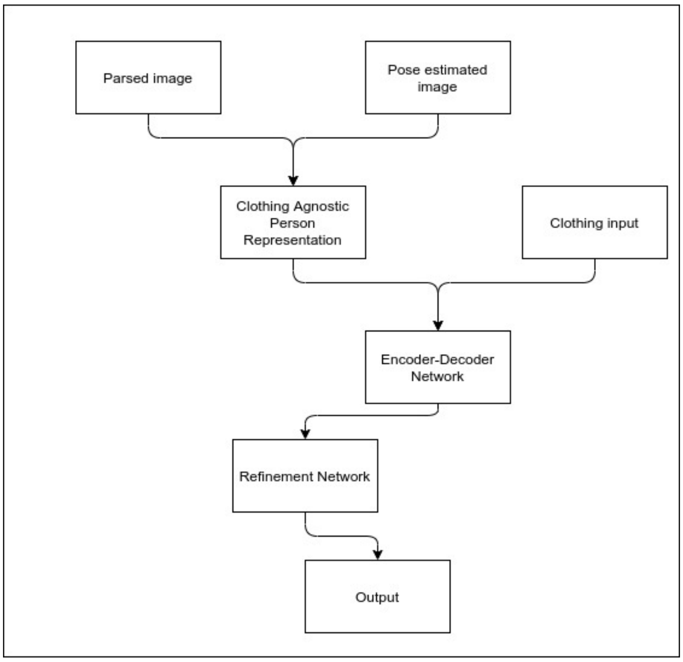


Figure 3.4: Data Flow in the Try-On module

3.2.2 Architecture diagram

This section consists of the module-wise architecture diagrams.

Pose Estimation

The pose-estimation module involves a two-branch multi-stage CNN. Each stage in the first branch predicts confidence maps and each stage in the second branch predicts the PAF. After each stage, the predictions from the two branches, along with the image features, are concatenated for the next stage.

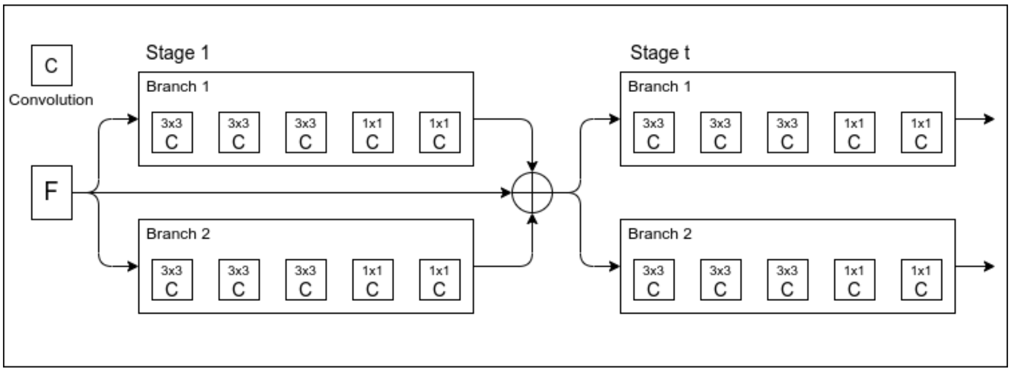


Figure 3.5: Architecture of the Pose Estimation Module

Human Parsing

ResNet-101 is used to extract the feature maps. The part module is then appended to the context of keypoints and parts while simultaneously generating parsing maps. This is then passed to a refinement network to obtain the output of this module.

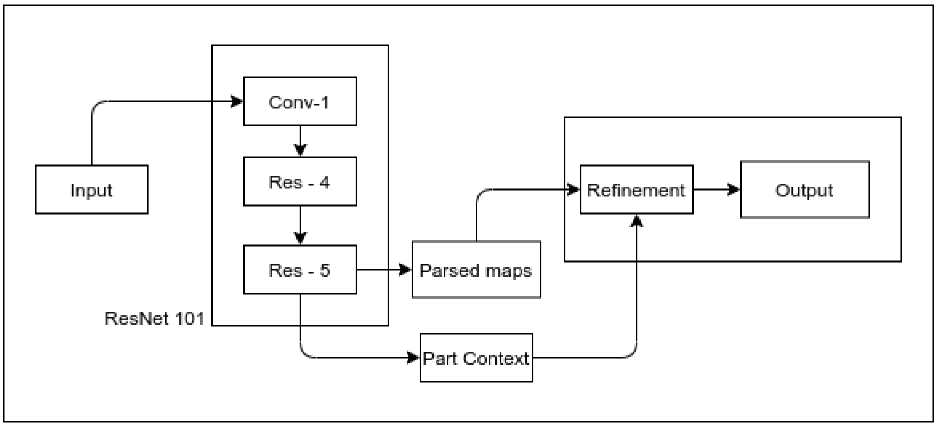


Figure 3.6: Architecture of the Human Parsing Module

TryOn Module

A detailed architecture diagram of the try-on module is shown in 3.7. The very first row depicted in the U-Net based encoder-decoder generator corresponds to the encoding layers and the second row corresponds to the decoding layers. The output obtained from the generator is then passed on to a refinement network so as to enhance the quality of the output image.

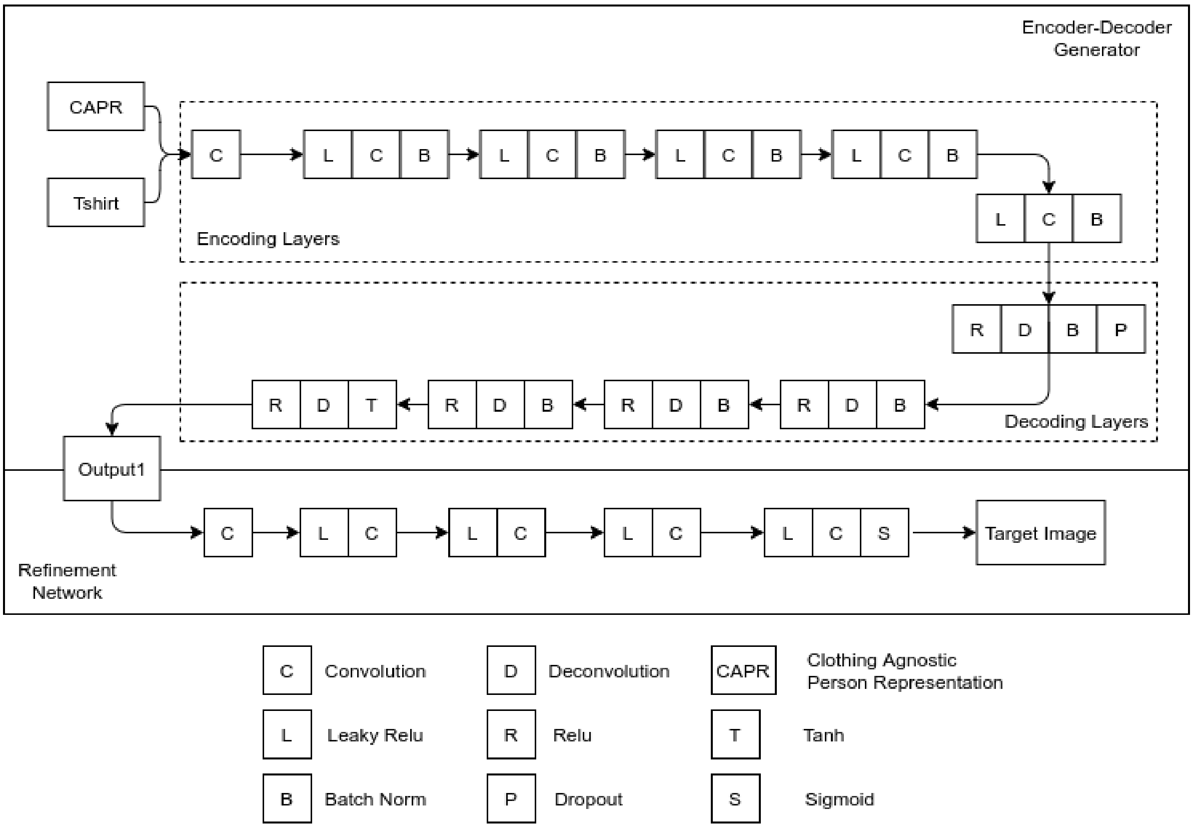


Figure 3.7: Architecture of the TryOn Module

Chapter 4

IMPLEMENTATION

4.1 Pose Estimation

The overall pipeline for this module is clearly illustrated in figure 3.2. The input image of size $w \times h$ is fed to a feed forward neural network that parallely predicts the confidence maps S of body parts and 2D vectors representing the part affinities (degree of association between parts). Finally the confidence maps and part affinity fields are parsed by a greedy inference to output the key points in the input image.
Let F denote the feature maps of the base network and $\phi_1 \& \rho_1$ be CNN at stage 1.

$$L_1 = \phi(F1) \tag{4.1}$$

$$S_1 = \rho(F1) \tag{4.2}$$

In the subsequent stages, the outputs from the previous stages are refined using the feature maps F and the previous PAF $L_{t-\mathit{1}}$ and the previous confidence map $S_{t-\mathit{1}}$. Let ϕ_t be the CNN at stage t.

$$L_t = \phi_t(F, L_{t-1}), \forall t \geq 2 \tag{4.3}$$

$$S_t = \rho(F, S_{t-1}), \forall t \geq 2 \tag{4.4}$$

The loss functions at each branch are calculated pixel wise. The individual loss functions at pixel location p is given below.

$$f_S^t = \sum_{j=1}^{J} \mathbf{W}(\mathbf{p}) \left\| S^t(p) - S^*(p) \right\| \tag{4.5}$$

$$f_L^t = \sum_{j=1}^{J} \mathbf{W}(\mathbf{p}) \left\| L^t(p) - L^*(p) \right\| \tag{4.6}$$

where S^* is the confidence maps' ground truth and L^* is the ground truth of PAF. W is a mask with $W(p) = \mathit{0}$ when p's annotation is missing.The vanishing gradient problem here is addressed by the intermediate supervision that replenishes the gradient periodically. Hence, the objective of this module is:

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t) \tag{4.7}$$

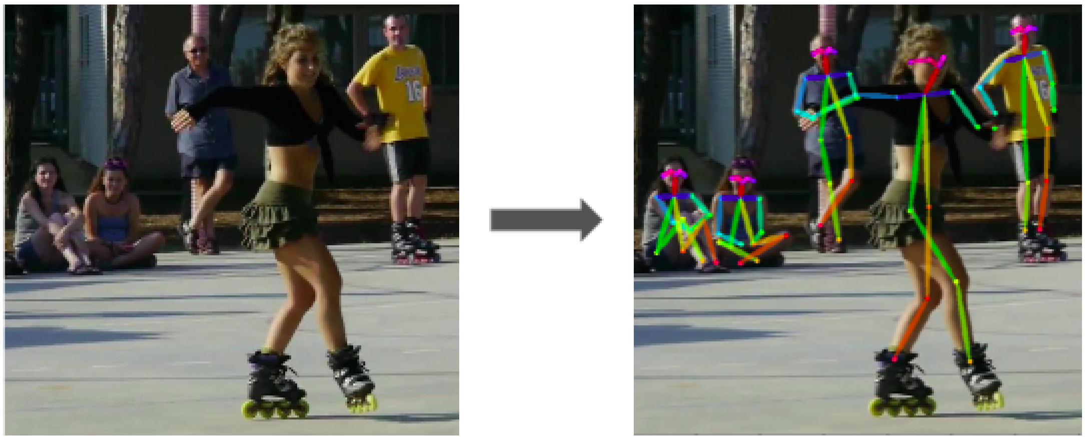


Figure 4.1: Pose estimation - Sample I/O

4.2 Human Parsing

ResNet-101 with atrous convolution is used as the basic parsing subnet contaning the Atrous Spatial Pyramid Pooling (ASPP), effectively segment or parse the input image in multiple scales.

The pretrained JPPNet built on top of existing ResNet-101 [Chen et al. (2017)]. The input image is scaled before feeding to the model with random scaling (from 0.75 to 1.25). Finally to get stable results, inference on multiple scales (0.75 to 1.25) would be performed.

For the given input image I, a list of parsed regions, C_I^G is generated

$$C_I^G = c_i^g | i \in [1, N] \qquad (4.8)$$

where c_i^g denotes the i^{th} region parsed and N is decided by the input human image (the value of N is 9 for a full body image).

A simple but efficient refinement neural network is designed that iteratively refines the parsing results. The intermediate parsing results obtained are reintegrated back into the feature maps with a convolution of 1×1. The entire network is trained end-to-end by feeding the results obtained in the current stage to the next.

A pixel-wise softmax loss L_p is calculated with respect to the ground truth with z as the input vector.

$$L_p = \frac{e^z}{\sum_{j=1}^{K} e^{z_j}} \tag{4.9}$$

The final result is computed as the average probabilities between each scale.

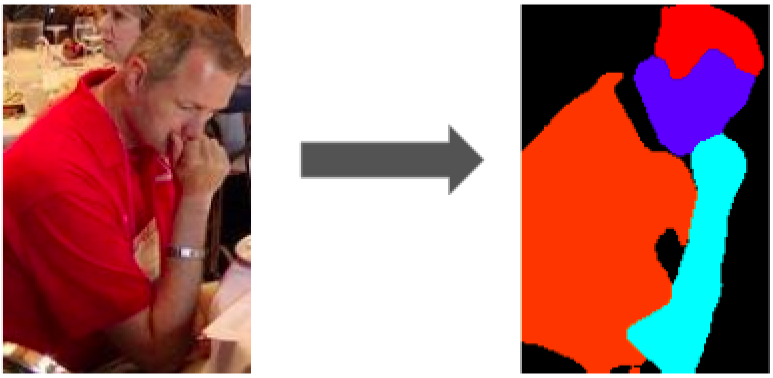


Figure 4.2: Human Parsing - Sample I/O

4.3 Try On

The ultimate objective of the module is, for a given user's image I and a clothing item c, generate a new image $\hat{I}$ wherein, the selected cloth c is naturally transferred to the corresponding region in the user's image while preserving the characteristics of the cloth like design and colour along with the also body part and pose information of the user.

The pose computed from the Pose Estimation module is represented as 19 keypoints. All such heatmaps are further stacked into an 19-channel pose map. Similarly, the output from the Human Parsing module is converted to a 1-channel binary mask. Also, to maintain the identity of the user, the physical attributes such as face, hair, etc. are incorporated to form a clothing-agnostic person representation p. It is notable that this representation is a little more detailed that the previous works.

We utilised an encoder-decoder framework based on UNet architecture with skip connections to directly share data between different layers. Let G_C be the function approximated by the encoder-decoder network which generates a new synthesized image I' and a segmentation mask M. We desire for the synthesized image, I' to be similar to the original reference image I and M to be similar to the mask generated by the Human

Parsing module.

Since the mask is a binary image, L_1 loss is used. However, this doesn't always work well for coloured images. Therefore, the loss function L_{G_C} is formulated from Ledig et al. (2017), Dosovitskiy et al. (2016) and Johnson et al. (2016).

$$L_{G_C} = \sum \lambda \left\| F(I') - F(I) \right\| + \left\| M - M_0 \right\| \tag{4.10}$$

in which λ is the hyperparameter that regulates the contributions of each layer to the loss function L_{G_C}, F is the feature map and M_0 is the result obtained from the human parsing module.

The synthesized image I' is required to composite with the warped clothing item c' which is designed to take place in the refinement network. The final desired output, $\hat{I}$, is the output of the refinement network which is a composition of the warped clothing item, c' and the synthesized image, I'.

$$\hat{I} = \alpha \odot c' + (1 - \alpha) \odot I' \tag{4.11}$$

where α is the binary mask generated by the refinement network and $\odot$ denotes the element-wise matrix multiplication.
In reference to the works of Jetchev et al. (2017) and Radford et al. (2016), the configuration of the Adam optimizer is set with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a constant learning rate of 0.0002. The encoder-decoder network is trained for 500 epochs and the refinement network for 60 epochs.

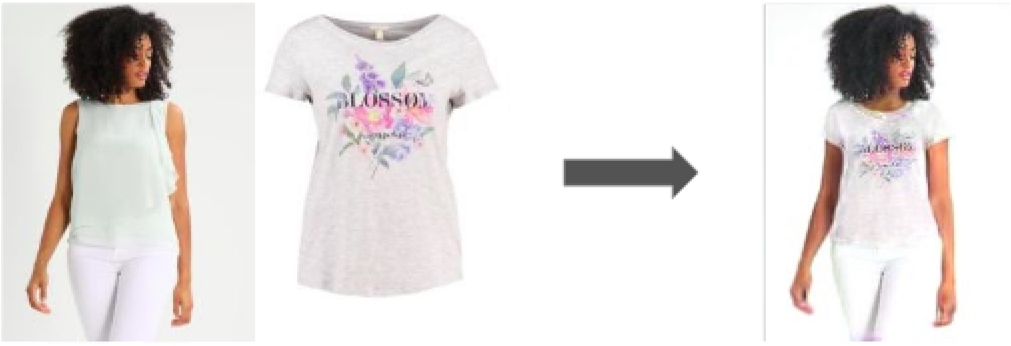


Figure 4.3: TryOn - Sample I/O

4.4 Prototype

A model e-commerce website is designed along with this project to show the real-time use and interaction of the user with this project.

4.4.1 Database

The popular open-sourced MySQL is used as the database for the designed website. The database consists of 5 main tables. The relationship between tables is depicted as an ER diagram in figure 4.4.

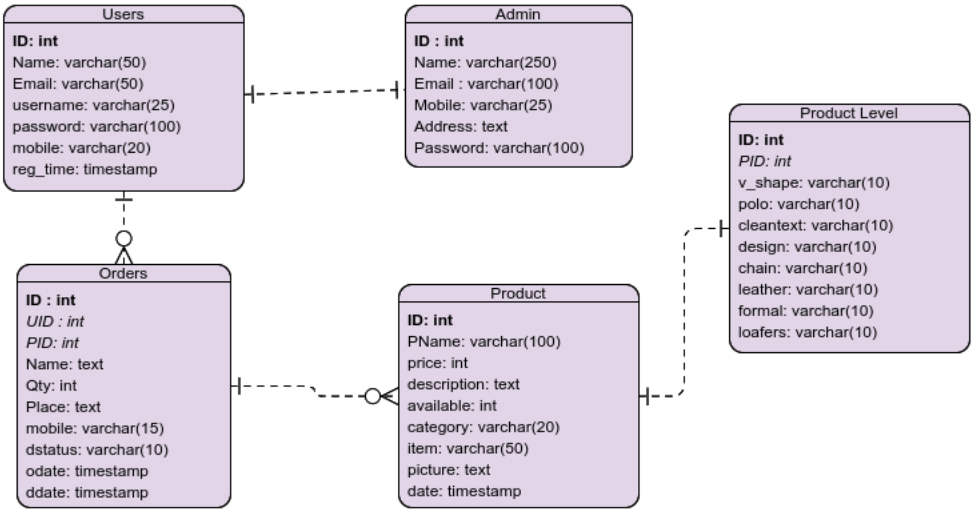


Figure 4.4: ER Diagram

Table Specifications

Table: *Users*

- **SNo** : 1
- **Fields** : *ID*, Name, Email, Username, Password, Mobile, RegistrationTime
- **Type** : The table is in 4th Normal Form
- **Purpose** : Stores the details of all the users registered in the website

Table: ***Products***

- **SNo** : 2
- **Fields** : *ID*, ProductName, Price, Description, Availability, Category, Picture, Date
- **Type** : This table is also in 4-NF
- **Purpose** : The Product table stores the details of all the products that are listed in the website.

Table: ***Admin***

- **SNo** : 3
- **Fields** : *ID*, Name, Email, Mobile, Address, Password
- **Type** : This table is in BC-NF
- **Purpose** : Contains the data of the website administrators/managers.

Table: ***Orders***

- **SNo** : 4
- **Fields** : *ID*, UserID, ProductID, Name, City, Place, Mobile, DeliveryStatus, OrderDate, DeliveryDate
- **Type** : This table is in BC-NF
- **Purpose** : This table contains every transaction/order that took place.

Table: ***ProductLevel***

- **SNo** : 5
- **Fields** : ProductID, VShape, Polo, CleanText, Coloured, Leather, Formal, Loafers
- **Type** : This table is in BC-NF
- **Purpose** : The table stores the characteristics of each product which is used for a naive content-based recommendation.

4.4.2 Server

Flask, a micro web framework that is written in Python is used in the creation of the website. Since, the other modules are also written in Python, Flask makes it easier to integrate and deploy the generated models.

Features

- The website boasts a beautiful UI with carousels in the home page
- Order details can be retrieved anytime, if the user registers before ordering
- Provides Guest Ordering functionality (One need not sign in to place an order)
- For every item that is being viewed, similar items are shown as recommendations using a naive content-based filtering algorithm.
- User can see how he/she will look if he/she wears a particular tshirt.
- A separate dashboard for the admin has been developed for easy maintenance
- Multiple HTML files are integrated by means of inheritance, thus, making it easier for the developers to upgrade/modify the website.

Screenshots

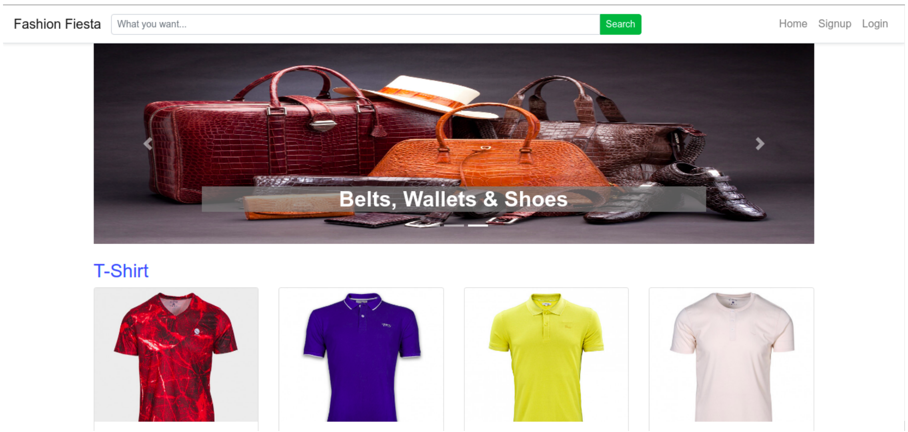


Figure 4.5: Homepage

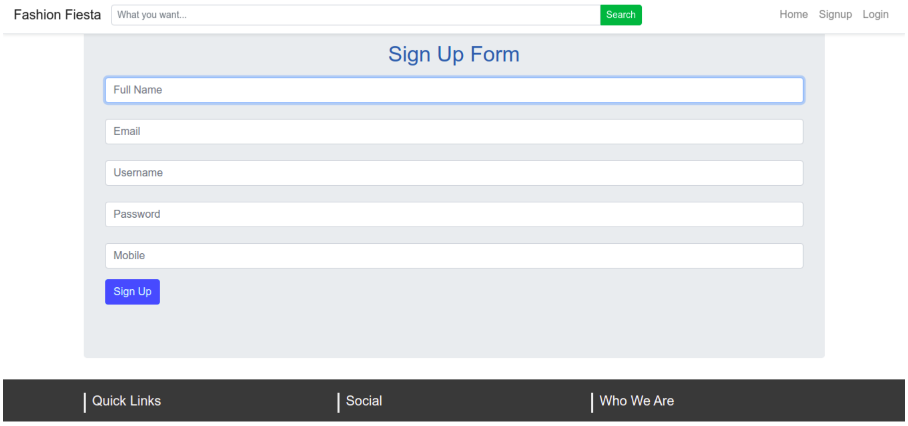


Figure 4.6: SignUp Form

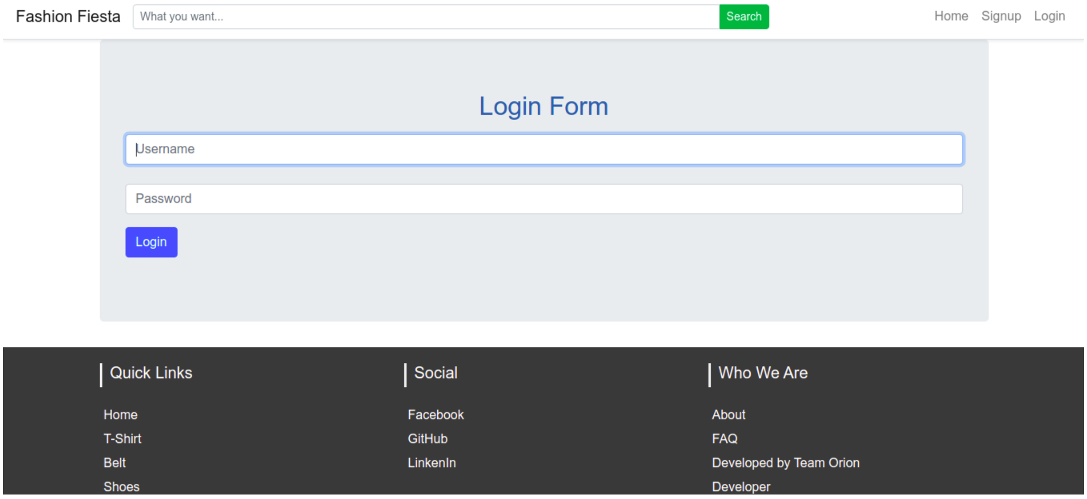
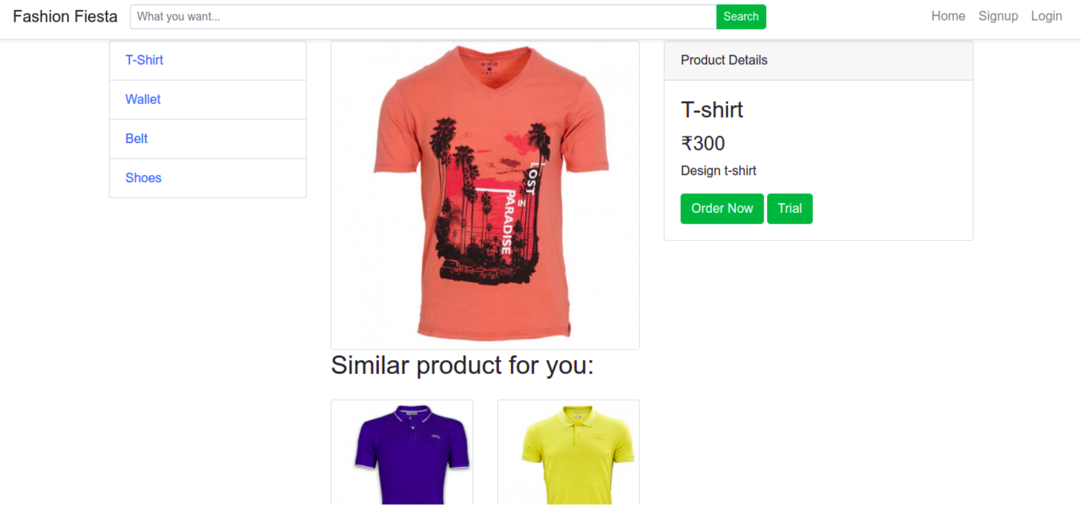


Figure 4.7: Login Form

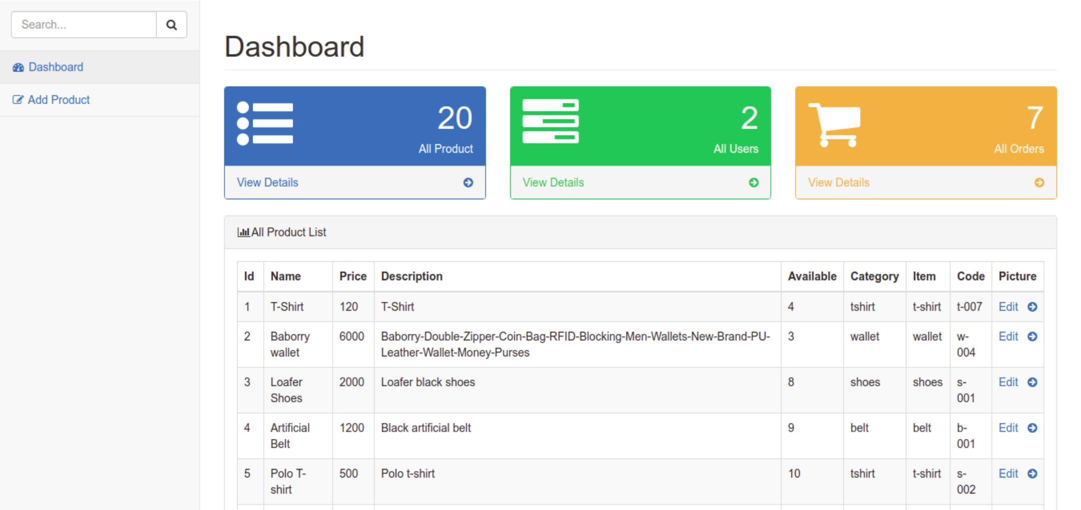


T-shirt

₹300

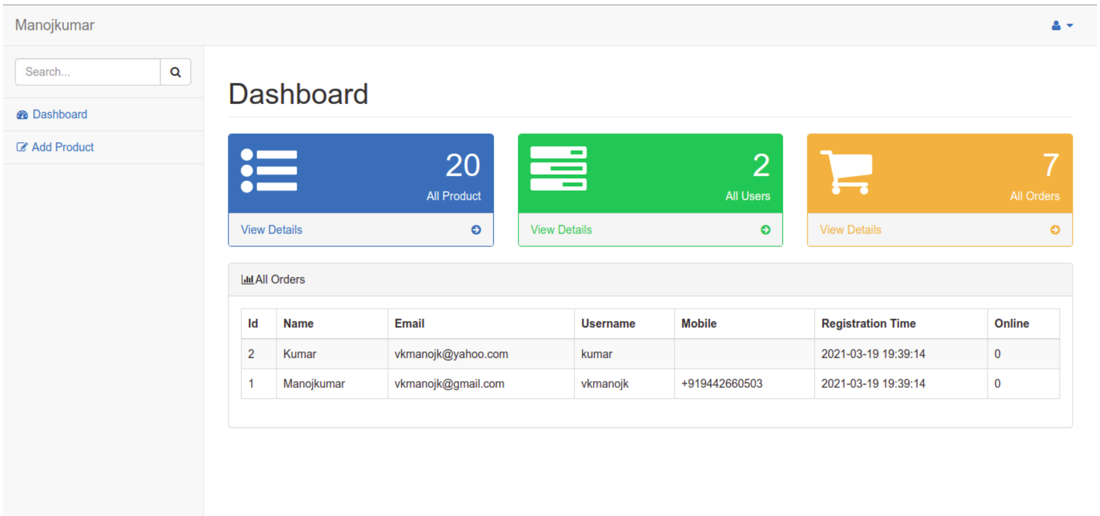
Design t-shirt

Figure 4.8: Product Recommendations



Id	Name	Price	Description	Available	Category	Item	Code	Picture
1	T-Shirt	120	T-Shirt	4	tshirt	t-shirt	t-007	Edit
2	Baborry wallet	6000	Baborry-Double-Zipper-Coin-Bag-RFID-Blocking-Men-Wallets-New-Brand-PU-Leather-Wallet-Money-Purses	3	wallet	wallet	w-004	Edit
3	Loafer Shoes	2000	Loafer black shoes	8	shoes	shoes	s-001	Edit
4	Artificial Belt	1200	Black artificial belt	9	belt	belt	b-001	Edit
5	Polo T-shirt	500	Polo t-shirt	10	tshirt	t-shirt	s-002	Edit

Figure 4.9: Admin Dashboard - Product Details

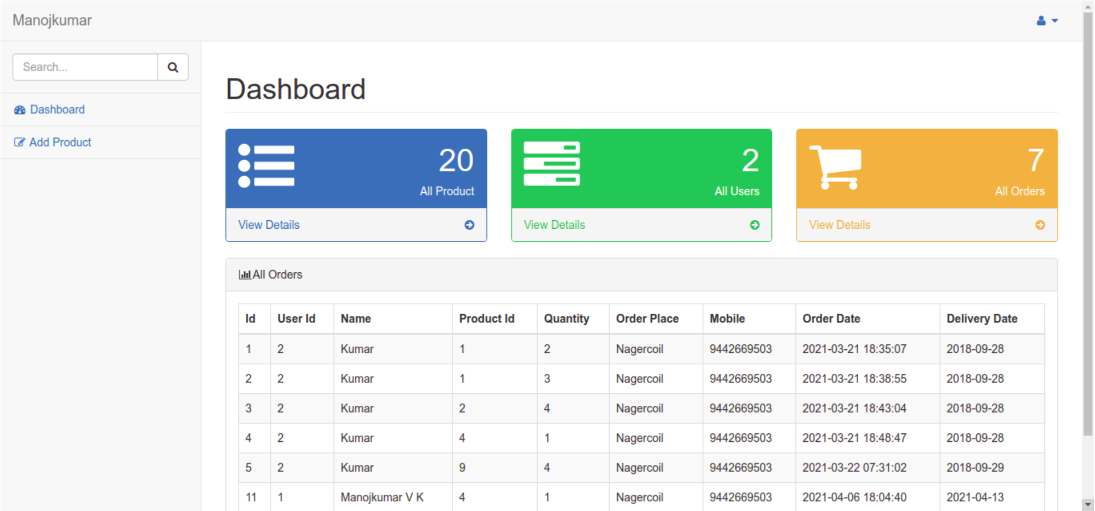


Dashboard

All Orders

Id	Name	Email	Username	Mobile	Registration Time	Online
2	Kumar	vkmanojk@yahoo.com	kumar		2021-03-19 19:39:14	0
1	Manojkumar	vkmanojk@gmail.com	vkmanojk	+919442660503	2021-03-19 19:39:14	0

Figure 4.10: Admin Dashboard - User Details

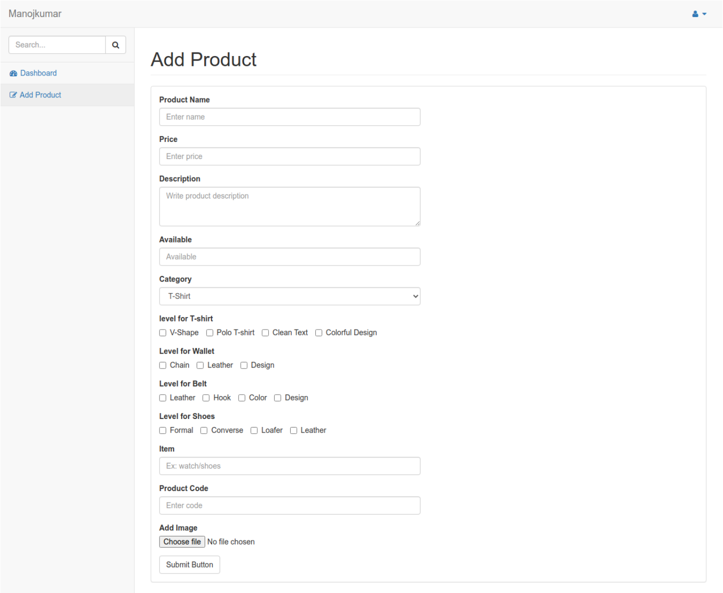


Dashboard

All Orders

Id	User Id	Name	Product Id	Quantity	Order Place	Mobile	Order Date	Delivery Date
1	2	Kumar	1	2	Nagercoil	9442669503	2021-03-21 18:35:07	2018-09-28
2	2	Kumar	1	3	Nagercoil	9442669503	2021-03-21 18:38:55	2018-09-28
3	2	Kumar	2	4	Nagercoil	9442669503	2021-03-21 18:43:04	2018-09-28
4	2	Kumar	4	1	Nagercoil	9442669503	2021-03-21 18:48:47	2018-09-28
5	2	Kumar	9	4	Nagercoil	9442669503	2021-03-22 07:31:02	2018-09-29
11	1	Manojkumar V K	4	1	Nagercoil	9442669503	2021-04-06 18:04:40	2021-04-13

Figure 4.11: Admin Dashboard - Order Details



Add Product

Product Name
Enter name

Price
Enter price

Description
Write product description

Available
Available

Category
T-Shirt

level for T-shirt
□ V-Shape □ Polo T-shirt □ Clean Text □ Colorful Design

Level for Wallet
□ Chain □ Leather □ Design

Level for Belt
□ Leather □ Hook □ Color □ Design

Level for Shoes
□ Formal □ Converse □ Loafer □ Leather

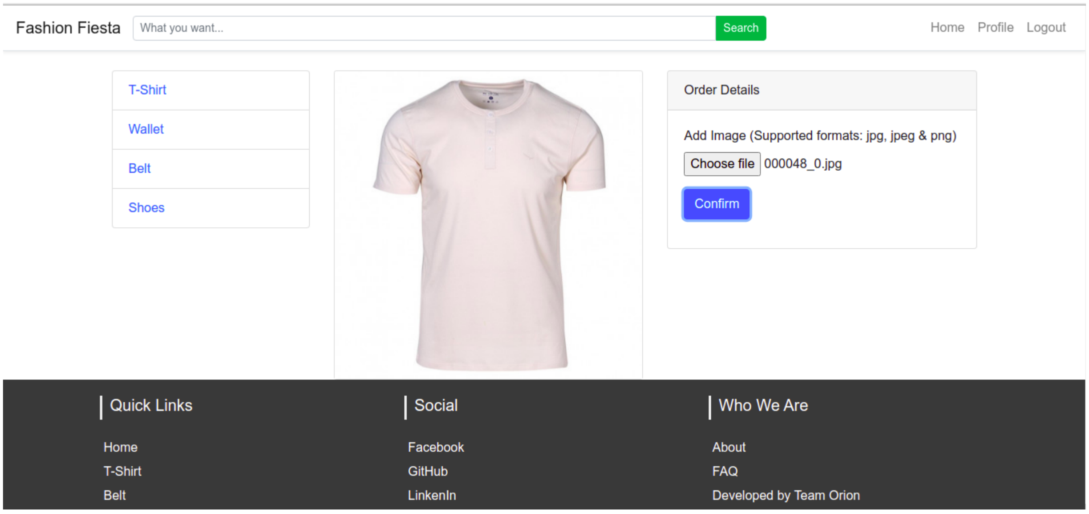
Item
Ex: watch/shoes

Product Code
Enter code

Add Image
Choose file No file chosen

Submit Button

Figure 4.12: Admin Dashboard - Add Product



Add Image (Supported formats: jpg, jpeg & png)

Choose file 000048_0.jpg

Figure 4.13: TryOn - Select TShirt

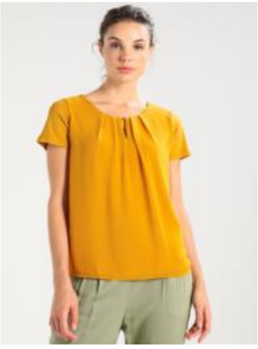
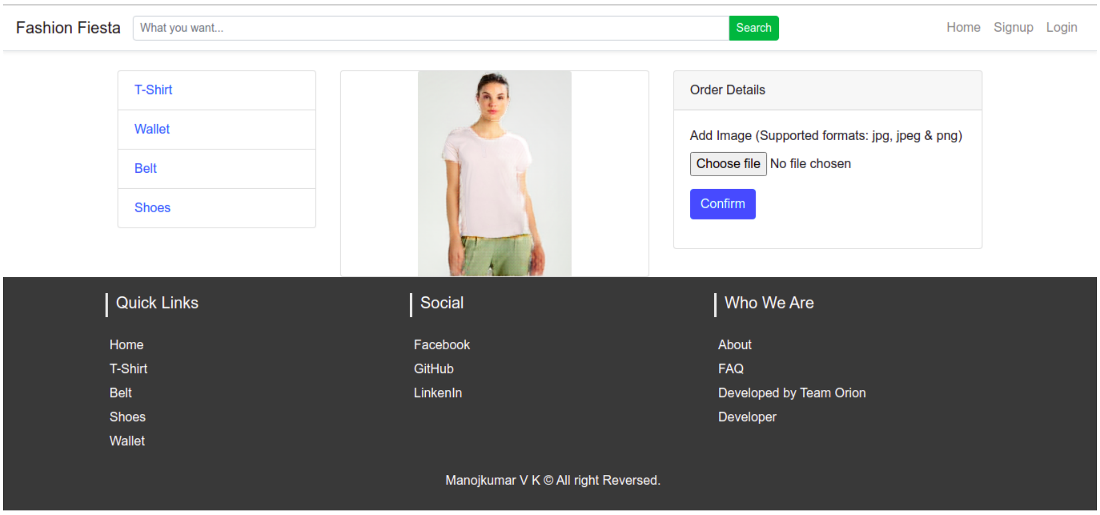


Figure 4.14: TryOn - User Image



Add Image (Supported formats: jpg, jpeg & png)

Choose file No file chosen

Manojkumar V K © All right Reversed.

Figure 4.15: TryOn - Result

Chapter 5

RESULTS AND DISCUSSION

5.1 Qualitative Results

A comparison of images generated by VITON[Han et al. (2018)] and our work can be seen in figure 5.1. Based on the comparison,it highlights the struggle faced by the VITON model to capture complex designs in the tshirt. Our model is able to represent it better with greater detail and accuracy, thus, confirming the effectiveness of the framework.

Nevertheless, attributing to the lack of availability of dataset beneath the waist region, the generated images in rows four and five of figure 5.1 is slightly distorted below the waist.

5.2 Quantitative Results

Inception Score (IS)

Inception scores are typically used to evaluate the synthesis quality of image generation models (mostly Generative Adversarial Networks (GAN)). Theoretically, the lowest possible score is *zero*, whilst the highest possible score is *infinity*. They are inversely correlated, meaning, the lower the score, the lower the quality of the image.

The inception scores of images synthesized by VITON and our model was calculated and is listed in the table 5.1. The IS of our model is greater than the IS of the VITON model approximately by 0.4.

User Study

Random sample images of persons and clothes were fed as input to the existing work [Han et al. (2018)] as well as our own model. The input images and the corresponding output from both the models were presented before random users as a form without disclosing which work is ours. It is done so to prevent bias in the results.

The user study scores evaluate the virtual try-on results, the images generated by the model, to be realistic. The obtained results are shown below in table 5.1 and figures 5.2, 5.3 and 5.4. ***Work A*** denotes the VITON model and ***Work B*** denotes our model.

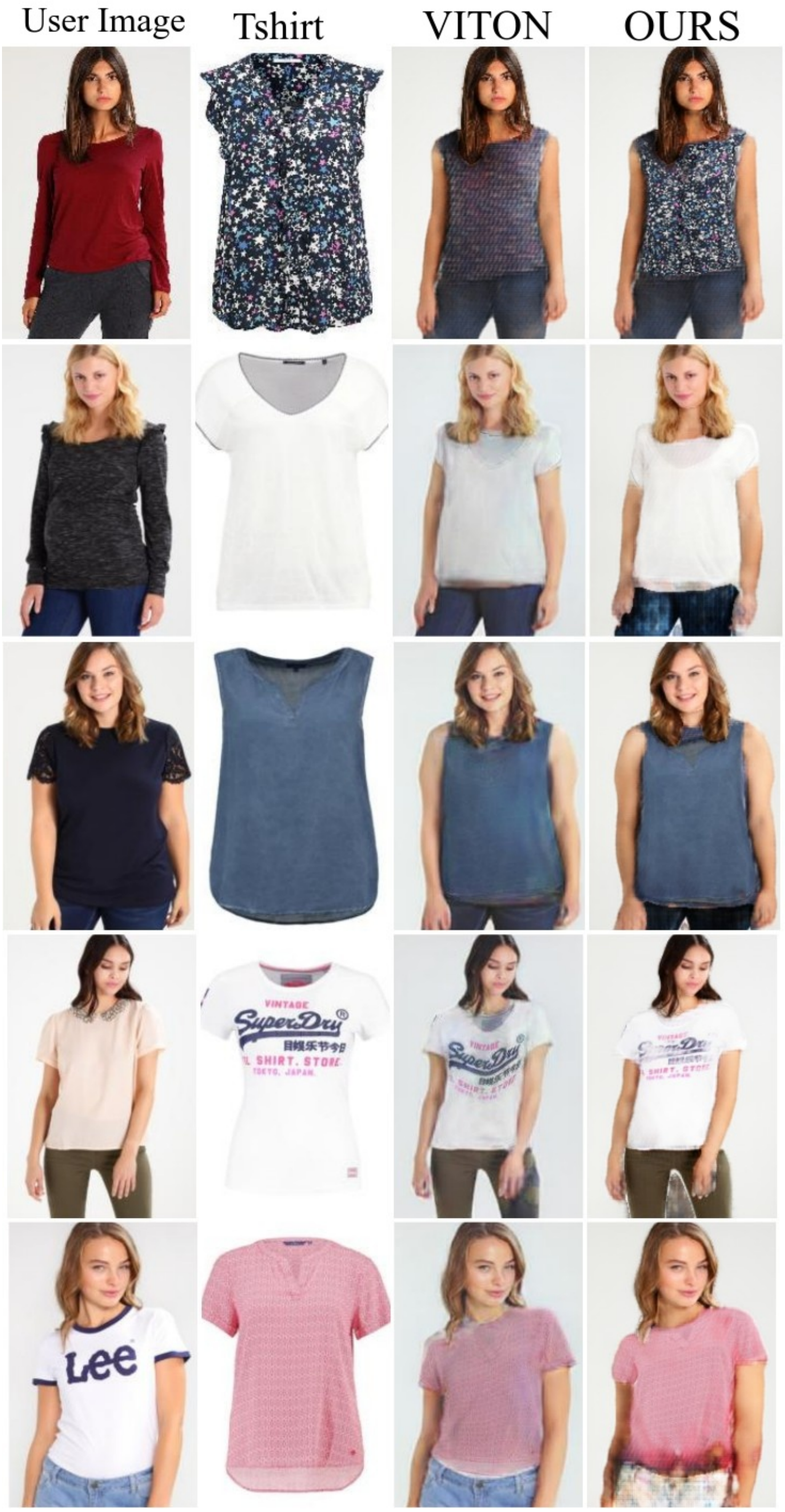


Figure 5.1: Qualitative Results

Work	Inception Score (IS)	User Study	Vote %
VITON	2.514 ± 0.130	25%	36.63
Our Model	2.975 ± 0.150	75%	63.37

Table 5.1: Quantitative Results

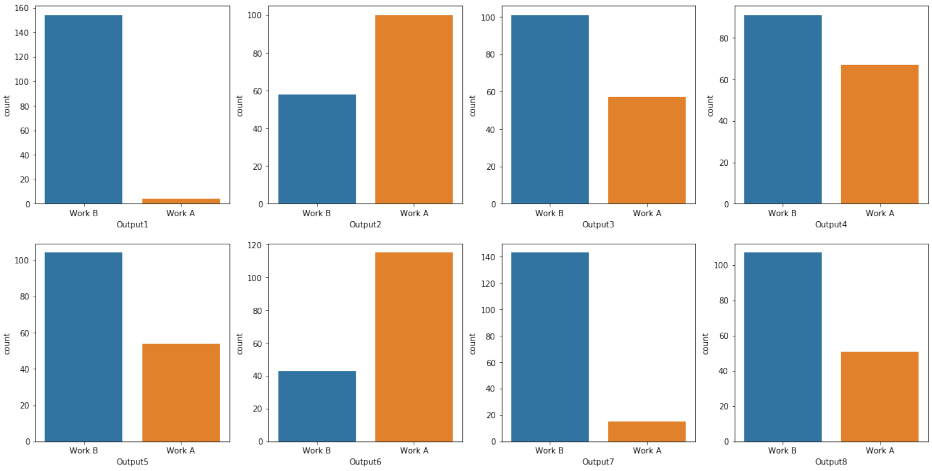


Figure 5.2: Quantitative Results represented as Bar Chart

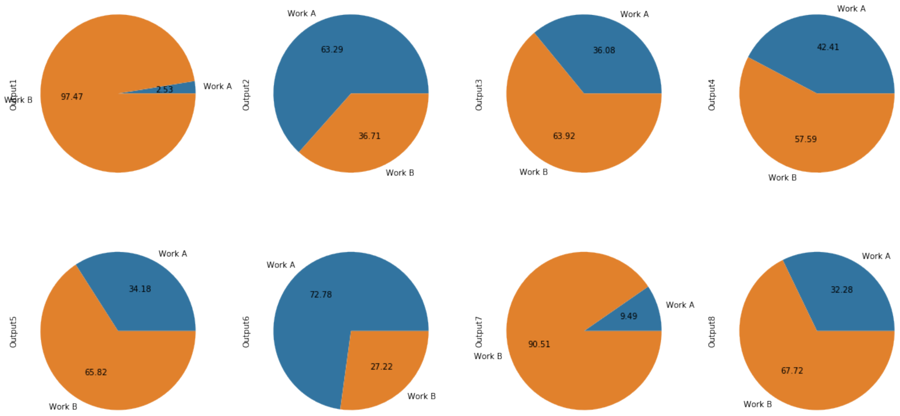


Figure 5.3: Quantitative Results represented as Pie Chart

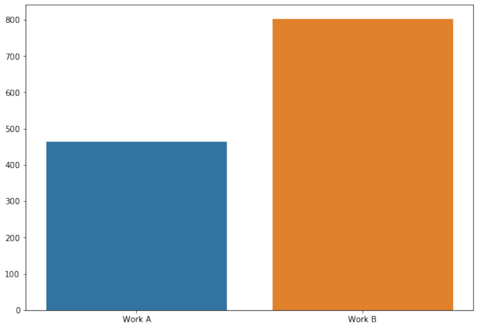


Figure 5.4: Quantitative Results - Vote share

Chapter 6

CONCLUSION

An e-commerce website using Flask and MySQL was created to demonstrate the uses of the model in real-time. The website is specifically designed to be user friendly with separate dashboards for the admins. By doing so, it results in higher customer satisfactions and eases the maintenance procedures resulting in reduced cost of operation for the e-commerce industry.
It is evident from the above results that the model created in this project has upper hand at synthesizing images with greater accuracy and detail in contrast to the existing VITON [Han et al. (2018)] model. The addition of refinement neural networks and a few tweaks in all the three models(Pose Estimation, Human Parsing, TryOn) has helped in successfully capturing the complex designs better than the previous research works.

The conducted experiments prove that the developed model has achieved promising results, both quantitatively and qualitatively. This also indicates that 2D methods are better alternatives to expensive 3D methods.

However, since the dataset contains only images of female models and tops or shirts, certain output images produced from our model are slightly distorted below the waist regions. The developed model also has difficulty in cases such as:

- When the input user image is sleeved while the selected clothing is sleeveless. Here, even though the model performs better than the existing works, there is a little distortion in the hand regions.
- When images from the wild are fed to the model. It can also be attributed to lack of diverse datasets or homogeneity in the chosen dataset.

Chapter 7

FUTURE ENHANCEMENT

As discussed in the conclusions, even though the model gives promising results, it is not yet ready to be actually deployed in real time.

The current level of the project focuses only on trying the selected shirt on the user. This could be further extended to pants/skirts, watches, chudidhars and even sarees.

When synthesizing a new image, each module takes around 2-3 mins making the total time required for the actual try-on aroung 4-7 mins. This time is not ideal and can be further reduced through deploying the model in a high-end system or cloud-based service providers.

Involving the model in real-time also raises concern for the user's privacy since his/her image is being uploaded to the server every time he/she trials an apparel. The current level of the project is only deployed in the local system and the image uploaded is saved to the local folder/file system itself. These issues can be tackled by incorporating WSGI servers with encryptions. Federated Learning can also be used, where the model is trained across multiple decentralized edge devices or servers holding local data samples, without exchanging them. However, this is still a relatively new domain in which tech giants like SAIL, Google, Owkin, etc are currently invested [Gossett et al. (Gossett et al.)].

REFERENCES

1. Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). “Realtime multi-person 2d pose estimation using part affinity fields.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

2. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., et al. (2017). “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.” *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

3. 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, 102897.

4. Dosovitskiy, A., Brox, T., et al. (2016). “Generating images with perceptual similarity metrics based on deep networks.

5. Gossett, S. et al. “These 11 startups are working on data privacy in machine learning.

6. Gültepe, U., Güdükbay, U., et al. (2014). “Real-time virtual fitting with body measurement and motion smoothing.” *Computers & Graphics*, 43, 31–43.

7. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L. S., et al. (2018). “Viton: An image-based virtual try-on network.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7543–7552.

8. Jetchev, N., Bergmann, U., et al. (2017). “The conditional analogy gan: Swapping fashion articles on people images.

9. Jiang, Y. and Chi, Z. (2019). “A cnn model for human parsing based on capacity optimization.” *Applied Sciences*, 9(7), 1330.

10. Jie Chen, Junwen Bu, Z. H. et al. (2019). “Image-based virtual fitting room.” *GitHub repository*.

11. Johnson, J., Alahi, A., Fei-Fei, L., et al. (2016). “Perceptual losses for real-time style transfer and super-resolution.

12. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., et al. (2017). “Photo-realistic single image super-resolution using a generative adversarial network.

13. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J., et al. (2017). “Multiple-human parsing in the wild.” *arXiv preprint arXiv:1705.07206*.

14. Liang, X., Gong, K., Shen, X., and Lin, L. (2018). “Look into person: Joint body parsing & pose estimation network and a new benchmark.” *IEEE transactions on pattern analysis and machine intelligence*, 41(4), 871–885.

15. Masri, A., Al-Jabi, M., et al. (2019). “Virtual dressing room application.” *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, IEEE. 694–698.

16. Moroz, M. et al. (2019). “Tendency to use the virtual fitting room in generation y-results of qualitative study.” *Foundations of Management*, 11(1), 239–254.

17. Radford, A., Metz, L., Chintala, S., et al. (2016). “Unsupervised representation learning with deep convolutional generative adversarial networks.

18. Ramesh, Ankit, B. . V. S. et al. (2018). “3d virtual trial room.” *IJERT*.

19. Reagan, C. et al. (2016). “A $260 billion ’ticking time bomb’: The costly business of retail returns.

20. Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., and Zhao, Y. (2019). “Devil in the details: Towards accurate single and multiple human parsing.” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4814–4821.

21. Smith, S. et al. (2020). “Ideas and market analysis.

22. William, J. et al. (2020). “Council post: Outsourcing in the 2020s: Where is this bandwagon heading?