National College of Ireland

Virtual Garment Imposition using ACGPN

MSc Research Project
Data Analytics

Saylee Vijay More
Student ID: X18180167

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Saylee Vijay More
Student ID:	X18180167
Programme:	Master's in Data Analytics **Year:** 2019-2020
Module:	MSc Research Project
Supervisor:	Christian Horn
Submission Due Date:	17 August 2020
Project Title:	Virtual Garment Imposition using ACGPN
Word Count:	8485 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Virtual Garment Imposition using ACGPN

Saylee Vijay More
X18180167

Abstract

The demand for online shopping is increasing day by day. Despite having several advantages, online shopping industry fails to enables the customers to virtually try-on garment on themselves before buying it. Developing a method as such would be a more significant advantage for online shopping industry as well as the customers. The goal of this research is to check whether the Adaptive Content Generative and Preserving Network (ACGPN) model can be used to impose virtual garments on the user's images or not? All the experiments of this research are done using the python programming language. ACGPN model is tested on three types of images depending on a different level of poses such as Easy, Medium and Hard. After the inferencing ACGPN model, the results turned out that the model works accurate on Easy pose images, good on Medium pose images and fails miserably on Hard pose images. In future, if the weights of ACGPN model is adjusted, it would yield good result on all types of pose images. This successful model then can be further converted into the Torch script and can be imported into the Android or IOS application using Pytorch Mobile.

1 Introduction

In these days, the online shopping market is booming (Kom,2019). Manufacturer and distributors make every effort to enhance their customers ' experiences of this online shopping. Online shopping demand is on the rise each day due to the many advantages of state-of-the-art technology (Kanti,2019). The "Always Open" policy follows online shopping, which further assists customers in shopping and reduces the extra work of going to the shop and shopping. It provides more variety, though, and the buyer may compare the other identical commodity to pick the cheapest and the cheaper one. Online shopping provides certain services, such as a return scheme and rewards schemes that will arrange the products according to their expense, reputation and importance. This interface allows the user to be mobile, meaning that he may order the product to be shipped at the location requested. While all of these benefits are obtained, the online retail industry does not allow customer access to wearing the garment even before putting the order (Kanti, 2019). All facilities were given to the user through the online shopping site, but before purchasing, the user can not try the product. This leads to the user returning or replacing the product sometimes.

In recent years, there are many Virtual garment try-on application has been published but did not get much accepted by the users. An application named "Zyler" which says that the user can virtually try on clothes, but in reality, the working process is not what a user may

like. Zyler application takes images from the user, segmentation layer emphasis only on the face of the user's image. This face image is imposed on a different body type (not on the user's body). Why will a user want to see their face on the different body (different shape of the body)? User will be more acceptable if the garment is imposed on their whole body.

Fig1. Zyler Application output

The solution to this problem can be derived from ACGPN as ACGPN segments the whole-body parts and impose a virtual garment on the user's input image. ACGPN consists of three features. 1. It is a semantic generation module which uses segmentation to map the human body with target clothes. 2. Clothing wrapped module which adjusts the garment images to deformed garment mask. 3.Content Fusion Model which adds the data to previous product to quickly discover the generation of the human body structure in the resulting combination layer.

2 Research Question

Can ACGPN be used to impose Virtual Garments on Custom images?

3 Related Work

3.1 Open Pose

For previous years, a mixture of contextual measurements of body parts and spatial dependency were used to determine an approximation of human posture. Tree base models (Andriluka et al.; 2010) and non-tree base models can be categorized using these space dependencies. The paper addresses one of the tree-based models (Ramanan et al.; 2005). Our software believes that even when doing unconventional things, people take such classic actions, such as cycling and ball kicking. A discriminative appearance model is developed with detection-estimated limbs. The paper also discriminates against features which distinguish a human figure in a frame in other frames. It can effectively monitor many individuals in a video. The multi-view body estimation algorithm is designed for low-

resolution unregulated environments (Germann et al.; 2011). Two steps are taken to do this. Extract the body's posture using the temporal, spatial outline that suits the triangular 3D posture for each frame. This approximate pose can have some symmetrical pieces flip ambiguities. For this purpose, a technology of optical flow is used to detect a sequence. The resultant 3D skeleton suits both figures.

For pose estimation on paper, a contrasting network with two discriminators and a multi-position generator is used (Chen et al. 2017). Fair poses are the two discriminators' defining irrational poses. The multitasking generator uses this discriminant as an expert who identifies true and false positions and trains them to construct a position that tricks the expert as actual. This method provides a more robust posture appraisal that can cross, inhibit and twist human bodies. This approach is also ideal for other problems with the measurement of shapes such as facial mark identification with DCNNs. The paper explores a particular approach to measuring poses through the use of a compositional model (Tang et al.; 2018). Compositional patterns define a significant part and sub-part hierarchies. They also offer high order relationships between body components which help to overcome ambiguities at low levels. The implementation of a thoroughly trained composition model addresses the problem of previous versions in complex situations.

Convolution Neural Networks produces image-dependent space models and photographic features in paper intending to build a computer architecture (WEI et al. 2016). This paper is based on the relations between the variables for articulated pose estimation. Credentials from previous frames are fed into a sequential neural network to create a detailed approximation of the position of the component. The complexity of the gradient disappearing was also discussed via a standard objective learning method. The picture heat map is used as a paper reference to CNN (Bulat and Tzimiropoulos; 2016). It has a two-part architecture where a set of N part heatmaps is obtained from the first subnetwork in which individual body parts are detected using per pixel sigmoid loss. Heatmaps obtained are sent to regressor subnetwork where heatmaps are staked along with the image to confidence maps body part representation. Cascade proposed in this paper is flexible enough to integrate with other CNN architectures.

3.2 Virtual Mirror-Based Garment Try-on

In 2013, shen developed a mixed reality system to try 3D virtual garment. This mixed reality system allows the user to choose a garment of their choice and see themselves virtually trying on the selected garment through the Virtual Mirror. This system enables the user to choose a different variety of garment to try on without using a fitting room. The contribution of this research is that the system customizes a partially visible or invisible embodiment depending on the person’s body shape such as the size of the body, skin colour which assists the user to have a proper garment fitting. Many of the 3D real human models from the CAESAR database were used to validate the methodology and customization of the body to match the body size was achieved. The Kinect camera was used for pose identification, body shape, user segmentation and facial skin colour identification. The final version matched the reference model precisely and looked natural. For skin colour matching a process to transfer the skin tone of the user to the embodiment consistency consisting of three steps was proposed. (I)Facial features were discovered using the Active Shape Model (ASM) approach (1]. (ii) To find cheek areas and patches, a piece by piece linear curve was applied. iii) A global conversion method was used in the last step to convert the colours of facial patches to the embodiment model.

In the real-time example, the use of performance estimation algorithms matched the 3D representation with the user's body picture precisely. Thanks to its precision and speed, the simulation method for virtual clothes has been used to animate. Customization of body parts such as height, bush size, hip size, hook, knee height, top arm, and forearm length, shoulder width has been arranged with an RGB-D Sensor for measuring body part. This helped to customize the embodiment that could be used virtually. (wager,2020). This paper proposed three scenarios a) virtual clothes on the embodiment, ii) Virtual clothes on the user's embodiment combines with user's faces image, iii) Virtual clothes on the user's image.

The processing time is rather censorious in many real-time applications. The average time for each frame was 110 milliseconds. It was concluded that among the three scenarios, users preferred virtual clothing to the user's body the most. The virtual clothes of the embodiment of the user, however, combine with the pictures of the user's face to give a more realistic view. For this project, the future work was a large body rotation. In general, consumers want to move to see the test clothes or real garments from various angles. To detect large body rotations, use of only RGB-D camera was not sufficient. (IMIGIZE, 2017)

The user can also select clothes by hand movement on the screen, with the help of Microsoft Kinect SDK(Filkov,2020). The machine takes various photographs of the outfit in conjunction with specific human postures and movement for a more realistic effect. A human skeleton was selected from the Microsoft Kinect skeleton collection to align clothing objects with the correct portion of the body.

This results in the movement of the clothing (Kenhub,2020) when the user moves quickly in front of the mirror. The device will use Kinect (kexugit,2020), for tracking and synchronizing the right-hand movement of the user in order to pick the garment image. It is placed on the human body following the choice of the garment made by gesture. In the hardware, the human body can be recognized and monitored in the runtime. Besides, during execution, the software takes the data and transforms it into human body skeleton joints to help even track down two people in front of the mirror at the same time. Following the positioning of the user's foot, the top clothing is placed on the human body. This paper was therefore produced in analysis of the movements of the arm and leg before the camera. To better the future, the author needed to explore and see, among other movement possibilities, the back and side turn of the human body.

3.3 Accessories Try-on

A virtual trial of garments has been carried out (Hu,2020) in the 2D/3D domain in the target body. Many virtual attempts have only been made to concentrate on the improvement of the clothes and no attention on hair, shoes and accessories. This article represents a new way to make hairs, shoes, clothing, watches, necklaces and hats completely automatic and acceptable. This paper is primarily aimed at transferring wearing items without intervention from the reference body to another target body. The process used in this paper is classified into two categories for clothing redressing. The introduction of the 3D fabric mesh and the target body mesh is taken, and the desired target body model is then given as the output after the implementation process. Since this is not an automated process, the intended body mesh in the different pose cannot be controlled.

The second category implies the ability to fit into other target body mesh (easy to transfer to your next target body model) with the 3D garment mesh fitted on the target body model.

The implementation of the 3D-CODED method proposed by Groueix et al. was modified for the identification of shape or dense correspondence. The 3D-CODED input is taken by two body mesh's, and then the body mesh is distorted. To make both inputs fit. The complex correspondence can, therefore, be acquired.

Human models have been obtained from the data set Fine Alignment by Scan Texture (FAUST), and data from Skinned Multi-Person Linear Model(SMPL). The conclusion for this paper is that the approach uses the composite location of the target body model and reference body model as well as the multi-layered garment concept. Limitations: - this method fails on dresses, as the fair portrayal cannot assign a specified point of the garment to two legs; this results in to tear off the garment.

3.4 Content-Based Image Retrieval

In the clothing store, it is sometimes tough for a user to find suitable wear. It would be beneficial if an app was developed that enables the user to give an example of cloth in the form of pictures they are looking. In this literature review, an image retrieval system is proposed, which allows the cloth to be searched and cloth retrieved similar to the search utilizing a share feature and colour. It was considered first to convert RGB to HSV colour space. Colour histogram is measured for the representation of colours. In contrast, the high and low threshold is obtained for method representation in the analysis.

Manhattan distance is determined to identify differences between the cloth image of the user searching for and the image stored in a database. The frame allows users to find a clothes image or a product that is similar to what is desired by uploading or not using text-based input. The system was developed using Windows 8.1 8GBRAM and Intel core i5 Processor 2.3 GHz. Content-based image retrieval(CBIR). MATLAB 's scripting language is the main language used here. Adobe Photoshop CS6 was used to design the interface for the system.

The CBIR system was tested with 50 photographs of five categories of garments, including T-shirts, dresses, pants and jackets. As input to the System Query by Example (QBE), only ten images per category are used. The top ten images were recorded for each query image and accuracy calculated at recall level 11. In all retrieval levels, the HSV histogram can often retrieve the relevant image. The average accuracy of the HSV histogram is 100%. Low-high thresholds are used for the distribution of shape. This proposed method can locate the desired picture in higher positions, i.e. (Recall level 0 to level 7). The high-level method's average precision is 96.36%.

The method of colour moment produced the desired image successfully up to the second level, but on the third level, the precision decreased. The average accuracy at the moment of colour is 66.55%. From its beginning, the Canny edge descriptor was the less efficient process. The unrelated pictures were taken back by the Canny edge descriptor. The average accuracy of canny edge amounted to 61,45%. It can, therefore, be concluded that it is possible to obtain an image of the clothing based on content to make shopping easier for the user. It may also contain details, such as price, size, branding and other information, and description of the garment. This report (Megha, 2018) has also developed a search system for smarter purchases of clothes. It was designed to draw out and contrast four main characteristics of the image of the clothing (i.e., the outline of the clothing, the texture colour, and the features).

A score for each property is calculated based on similarities between the picture of the user and the images stored in the database. The last scoring is generated to collect each property's weighted score. The highest picture is chosen and delivered to the user. The system

functions well according to the results of the test. Even if improvements are still required, the Gabor filter is effectively used for texture detection.

Nevertheless, it is not necessarily accurate as the pattern is not essential to the style of garments. Fourier Descriptor is used to detect a picture that works very well. SIFT is based on function detection that also benefits from the invariance of the scale. The RAMSAC algorithm was executed, but it was found to be less useful. The method of comparison of the colours must be improved since it is highly diverse and relevant to the colour amalgamation of the garment. The future work will concentrate on the areas where the modern mobile smart shopping network could be invented.

3.5 Image-based Garment Try-on

Currently, the Try-on (Bonetti, 2018) virtual garment fitting system is created by attaching virtual garments to the person's image. The appeal of this technology grows daily and leads to research, but it continues to change. The reason for this paper is not just to transform the clothing chosen into the most appropriate share, but also to maintain the identity of the clothing that is covered in the user's picture. A new learning Characteristic Preserving Virtual Try-on Network (CP-VTON), has been developed in paper to solve every real problem in this area.

First of all, with a new Geometric Matching Module(GMM), CP-VTON discovers the transformation of the finely layered spline for transforming the inside wear into the user's body shape. As in the past, the Geometric Matching Model is stronger than Internet correspondence computing. Secondly, an attempt is made to reduce the limit artefact of deformed clothing and achieve more realistic results. Try-on kit that composes the mask to fit the deformed clothes with the picture given to guarantee the smoothness of the test module. There the user creates a reference image of himself using the usual clothing and the chosen image of the user.

Additional CP-VPON is used to synthesize a new image of the user using the option of clothes in which the pose and shape of the body structure of the resulting image are established. The characteristics of the clothes chosen are reduced here, eliminating the effect of traditional clothing. In both experiments, the dataset obtained by Han et al. is used. This consists of nearly 19,000 women's front view and top garment sets. 16,253 were split train and test set of cleaned pairs. Two methods were tested on the composition of the mask I CP – VTON (No mask) and ii) CP-VTON (with mask) L1 loss, of which 14,221 were used in training and 2032 pairs for testing. This was inferred that the user essentially wears chosen clothes, and the CP – VTON picture (mask) has been made. In contrast to without L1 (renders) and without L1 (masks, the target tissue mask offers both quantitative and qualitative results.

This experiment demonstrates that the entire CP-VTON pipeline is beneficial in virtual testing for clothing and retains key characteristics (texture, emblem and sticking) of the in-shop garment. The explanation for this was to make the resulting image (Target clothes + user picture) more realistic in the paper (Yang,2020). The Adaptive Content Generating and preserving Network (ACGPN) research paper uses a more sophisticated approach than CP-VTON. The ACGPN anticipates the relevant specifics from the user picture layout that may alter in the future after virtual attempt. The new selected clothing illustration jacket will be put on the picture (long sleeve shirt-> Arm, Arm- > jacket) if the consumer is using long shirts. It will determine if the quality of the image will be conserved in order to gain valuable knowledge from its structure for future predictions.

This conservation will lead to practical and successful descriptions of the virtual photo try-on and clothing. ACGPN consists of three features: -First, the Semantic Generation Module (SGM), creating, through meaningful segmentation, the mask of the human body and the mask of the target clothing area, creating the meaningful alignment of the dimensional structure. Secondary, following a deformed garment mask, the Clothing Warping (CWM) module created to deform the target garment image. In (CWM), Thin-Plate Spline (TPS) implementation for geometrical standardization, but still maintaining primary image clothing characteristics Third step is the Model Content Fusion (CFM), which gathers information from the last element so that the preservation and development of specific human body components in the output image is scalable. The details used here are the same as the paper. Conclusion ACGPM produces realistic virtual test results while maintaining the basic features of the clothing (texture, logo and border) and the identification components of the individual (pose, Body parts, bottom clothes).

Three modules (GMM), (CWM) and (CFM) has been specifically designed. The VITON data collection, which included three different forms of difficulty testing, was used to assess the ACGPN. The findings thus suggest a more significant superiority over ACGPN in terms of the user's analysis, image quality and quantitative measures over the conventional method of virtual research.

4 Research Methodology

The objective of this paper is to impose a virtual garment on the user's input image. To obtain the results, several steps need to be carried out, such as Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. Based on the requirements as mentioned above, Cross Industry Standard Process for Data Mining(CRISP-DM) methodology is used, as it is a well-known approach for planning a data mining project. In the following was explained how the research project planning is related to the CRISP-DM methodology and what tool and innovative approaches has been applied to bring the solution to the results.

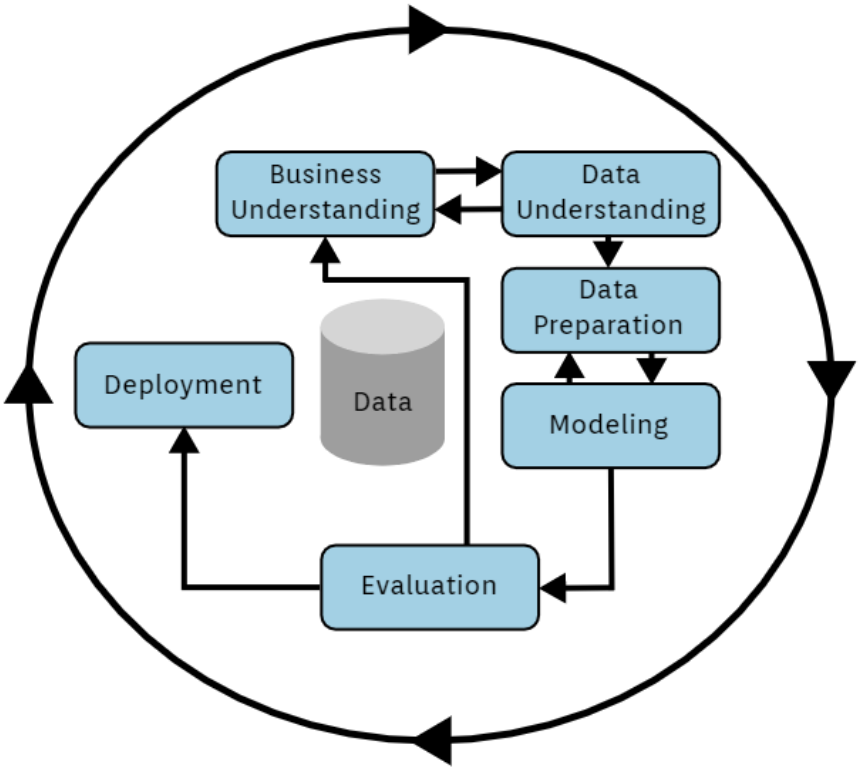

Fig 2. Working of CRISP-DM

4.1 Business Understanding

The online shopping industry is booming day by day. There are many applications where we can virtually try-on different things like spectacles, jewellery, make-up, hair colour and many more. There are some applications like (Zyler) which allows user to click their picture and upload it. This picture is further segmented, and only the face part is taken with the help of face keypoints generated through open pose. The face part image is further imposed on a dummy body available in different sizes and colours. From a business point of view if there was a system which took the whole body picture and would impose clothes on the user's image as per the user's choice, then it will surely be profitable. A system like this is yet not released. A user will be happier if the clothes are imposed on their body image. They would see how that garment would look on them. This idea will reduce the return and refund policy for online clothes shopping application as before buying it, they can try it on themselves virtually.

4.2 Data Understanding

In our project, we used a pre-trained model- Adaptively Generating Preserving Image Content(ACGPN) to generate than the image of the user with the selected garment imposed on it. For this VITION dataset has been used which consists of 19,000 image pairs, each of which consists of front view image of women and a top clothing image. This dataset consists of different types of images from different difficult level. It has easy pose in which the model's arms are straight towards the ground. Medium pose in which the model's arms are in the pocket and hard pose in which models' arms are folded or have a similar pose.

To generate the output Deep Fashion has provided test dataset which consists of different variables such as: -
Test_img:- Images of User
Test_color: - garment image (to be imposed on the user's image)
Test_edge: - consists of garment image edges.
Test_label: - human parsing representation of test_img.
Test_pose: - body key-points of test_img
Test_colormask:- consists of strokes
Test_mask:- consists of a random mask.

Test_colormask and test_mask are used to shade the image to make it incomplete so that the network learns to paint it.

4.3 Data Preparation

To impose clothes on the user's image, we have used Deep Fashion-Try-On pre-trained model, also known as ACGPN model. To generate test data for this model, libraries like Open pose and Self-Correcting Human Parsing strategies have been used. First, we take input images from the user; this input image is resized to 192x256 by using Pillow library since it is the requirement of ACGPN model. To generate test_label, Self-Correction Human Parsing strategies have been used. They have provided a state-of-the-art trained model on three popular datasets consists of the different label system. In our research, 'LIP' pre-trained model has been used. The LIP pre-trained model uses LIP dataset, which consists of more than 50,000 images; LIP is the biggest individuals human parsing dataset. This dataset focuses more on real complex situations. LIP consists of 20 different labels includes ['Background', 'Hair', 'Glove', 'Sunglasses', 'Upper-clothes', 'Left-shoe', 'Right-shoe', 'Socks', 'Pants', 'Left-leg', 'Right-leg', 'Left-arm', 'Face', 'Right-arm', 'Jumpsuits', 'Scarf', 'Skirt', 'Dress', 'Coat', 'Hat'] . By using this strategy, test_label has been generated with .png label file.

To generate test_pose, Open pose library has been used. It generated the body key point of the image from test_img files and saved it into the test_pose. Here Open pose Demo has been implemented, and COCO-Model dataset has been used (pose-iter-44000.caffemodel).
Test_color and test_edge are taken from VITION dataset itself depending upon the same file name. It is a process where the user decides which garment they want to virtually try-on. The test_colormask and test_mask are taken from the Testing VITION dataset as it is.

4.4 Modelling

ACGPN model was published five months ago, i.e. on March 2020, there is not much information available on how to generate data variables form custom images. By understanding the data, the structure we found out that the test_label can be generated by using Self-Correction Human Parsing strategies and the segmentation labels of ACGPN is applied in this model. Similarly, we found out the test_pose are the body key points of the input images which can be generated by using Open pose. All these data variables are then imported into ACGPN model for testing. For inference, we made sure that the input images are in a different pose to check the ability of ACGPN to impose garment on the different pose.

4.5 Evaluation

The ACGPN model was tested using different body pose so that the evaluation can be in a qualitative manner.

5 Design Specification

5.1 UNET

For segmentation of Bio-medical images, Olaf Ronneberger et al. created Unet. Unet architecture consists of two approaches. The first approaches are the contradiction path (also known as the encoder) which captures all the background information of the picture. The encoding is the standard array of convolutional and max-pooling layers. The second approach is the approach for the symmetric expanding (also knowns as a decoder) which uses transposed convolutions to extract exact locations in the images. Hence it is an end-to-end fully convolutional network (FCN), i.e. it includes only convolutional layers and no dense layers that allows it to accept images of all size.

5.2 VAE (Variational autoencoder)

A variational autoencoder (VAE) is known as a directed probabilistic graphical model (DPGM) which forms an auto-encoder like architecture and whose posterior is estimated by a neural network. The highest layer of graphical design model (z) in VAE, where the generative process starts is treated as the latent variable. The complex process of data generative is described by g(z), which the structure of a neural network is modelled. In short, VAE is the neural network which learns or trains to reproduce its input. It measures the probability density function (PDF) of the data, which is training. When VAE is trained with natural images without any random gibberish, it gives high probability value to that image and the low probability value to the image with random gibberish.

5.3 CNN (Convolutional Neural Network)

The convolutional neural network has achieved excellent reorganization performance for video and images (Krizhevsky et. At 2012). This is because there are several public images repositories available. The convolutional neural network can be built using many different parameters depending on what is the requirement of the task. The selection of feature consists of one of the significant obstacles and often affecting the performance of a convolutional neural network. For image classification, there are many techniques available, including one of which is the leading technique, which offers reliable and high performance in comparison with other available technique. Classification capacity is used for the performance assessment among the models. By increasing or decreasing the width and breadth of the image data, also known as fine-tuning, can be used to enhance the model.

Correct assumptions are made with pixels and statistics from the image attributions. A Seven-way type soft-max activation layer is provided for the final layer of CNN architecture to classify images into seven different types of skin lesions. The fundamental theory behind CNN success is that the higher level of attributes has better knowledge about the data at the multilevel representation.

5.3.1 Partial convolutional layer

Partial convolutional was initially proposed to treat the incomplete input data, for example-holes in images. It normalizes performance to compensate for the incomplete data portion. A partial convolution layer consists of masked and rest convolution, accompanied by a mask-update setup. The patch pixel of the picture is not noticed during convolution.

5.3.2 Max pooling layer

Max pooling or minimum pooling (also known as pooling operation) in every patch of every feature map it measures the largest or maximum value. The findings are pooled feature map or down-sampled displaying the most present feature in the region but not the average feature present if it is the case of average pooling. This has been shown to perform well for computer vision activities such as picture recognition in action than standard pooling.

5.3.3 ReLu Activation Layer & dropout

ReLu is a non-linear method for accessing the neural network of multi-layer or deep neural network. The performance of ReLu is the highest value between zero and the input value. The output is always zero if the input is positive for an input value, and when the input value is negative. Which means all the non-negative values are first kept in the memory, and then this value is transformed to zero. For shutting down neuron of random seeds, the dropout layer is used. To increase the effectiveness to generalize the model, this action is taken.

5.4 VGG

VGG is a model that recognizes the object and allow up to 19 layers. It was constructed like deep CNN. It also often reaches the baseline for a variety of activities and data other than ImageNet. Till now, VGG is the most commonly used architecture for image recognition.

5.4.1 Input

VGG takes RGB images of 224x224 pixels. The writer also cut down the 224x224 patch core of every image for ImageNet Competition just to retain the same input size of the image.

5.4.2 Convolutional layer

In VGG, the convolutional layer uses 3x3 small receptive field. It is the smallest size gains (left/right and up/down). The ReLu unit also allows 1x1 convolutional filters which further functions as a linear transformation of the input. To preserved the spatial resolution after convolution, the convolutional strides are set to pixel 1

5.5 GAN

Generative Adversarial Network is a deep learning-based generative model that is used for unsupervised learning. This Is a system where two neural networks compete against each other to create or generate variation in the data. It was first described in a paper in 2014 by Goodfellow, and a standardized and much stable model theory was proposed by Alex Bradford 2016, which is also known as Deep Convolutional General Adversarial Network (DCGAN). Most of the GAN today uses DCGAN. The GAN architecture consists of 2 sub-models known as the generative model and the discriminative model. The general network that takes a sample and generates a sample of data. Discriminator network describes whether the data is generated or is taken from the real sample using binary classification problems with the help of sigmoidal function that gives the output of 0 to 1.

Generative means that the model follows the unsupervised learning approach and is a generative model. In adversarial setting, the model is trained with the adversarial setting, and the network simply means for the training of the model we use a neural network as an artificial intelligence algorithm. The generative model analyzes the distribution of data in such a way that after the training phase the probability of the discriminator making a mistake it maximizes and the discriminator, on the other hand, is based on the model that will estimate the probability that the sample is coming to form the real data or not in the generator. In this Pix2Pix Model, GAN is used. The model Pix2Pix is a form of conditional GAN or CGAN where the output images are generated depending upon the input, which in this case is the source image.

The discriminator is equipped with source images and target image both to decide whether the image of the target has a reasonable transformation of the source image. The generator to generate a reasonable image in the target domain, the generator being trained by adversarial loss. The L1, loss which is calculated between generated image and expected image also updates the generator. This additional loss helps to generator model to generate a reasonable transformation of source picture.

5.6 ResNet

For Numerous computer vision jobs, the residual network is the backbone; it is a classic neural network. ResNet allows 150+ layers while training the extreme deep neural network, and that is the speciality of it. ResNet is an “Exotic architecture” which depends on the micro-architecture modules (also referred to as “network-in-network architecture”) as opposed to multiple sequential architectures like AlexNet, OverFeat and VGG.
Building Blocks are referred to as the micro-architecture, which is used to create a network. With many numbers of micro-architecture building blocks with standard layers like CONV, POOL. Results in the macro-architecture (i.e. the need network itself). ResNet architecture was first introduced in 2015 publication “Deep Residual Learning for Image Recognition”, which reveals that use of regular SGD (and a sensible initialization function) can be used to train the intense network by making use of the residual model.

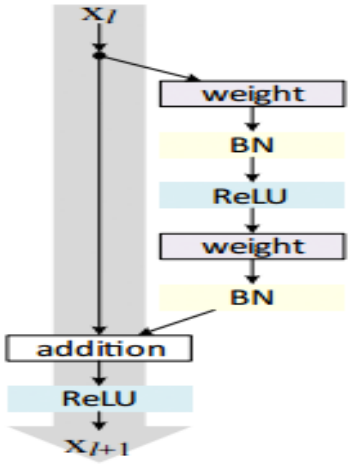

Fig3. Architecture of ResNet

Although ResNet is much deeper than that model VGG16 and VGG19, the size of the model is currently slightly smaller because of the usage of global average pooling instead of fully connected layers. Hence the ResNet50 model size is decreased to 102mb.

5.7 STN Network (Spatial Transformer)

Suppose we add a distorted image as input. In that case, the spatial transformer will modify the input before transferring it to the CNN model, so the output is we get from STN is the definitive version of the image. In the collection of different modules, the spatial transformer is another block of LEGO, where it discards the invariance of spatial from the input image by implementing a learnable affine transformation which is further followed by the interpolation. The STN block is applied before CNN and does not need to do anything after applying as it mostly works by itself.

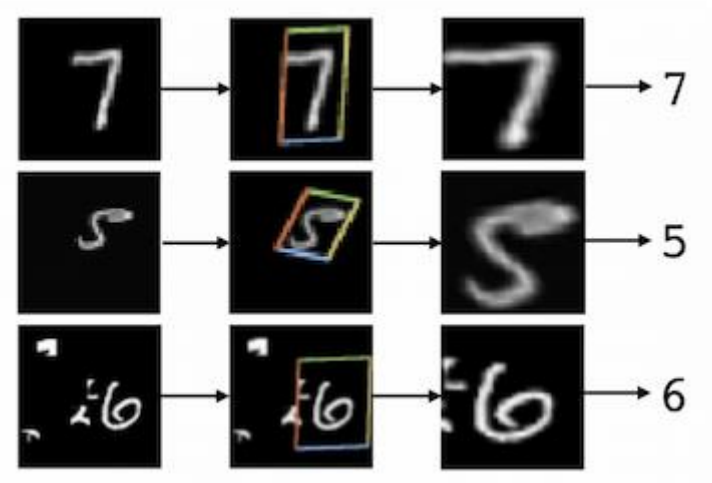

Fig4. STN-FCN AFFINE

5.8 DeepLabV3

As the primary feature extractor DeepLabV3 uses pre-trained model ResNet-101 from ImageNet with atrous convolution. Atrous convolution is the replacement for the downsampling layer. The last block of ResNet in the modified ResNet model makes use of atrous convolution with various dilution rates. To modify the ResNet block, DeepLabV3 uses atrous spatial pyramid pooling and bilinear upsampling, which is for the module decoder, which is on the top. The depth-separable convolution with strides will override all the max grouping operation. After 3x3 depth convolution each additional batch normalization and ReLu activation it has applied. The model depth is expanded without modifying or attending the network structures entry flow.

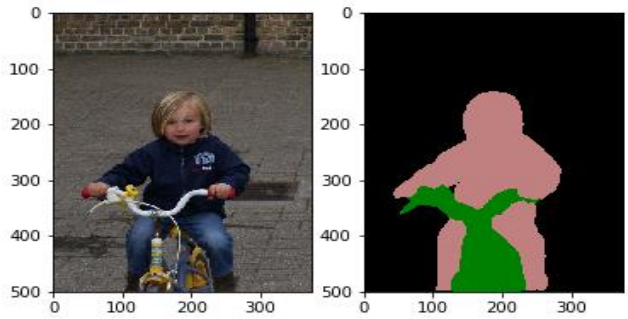

Fig 5. Semantic segmentation using DeepLabV3

5.9 MobileNetV2

MobileNetV2 is substantially improved than MobilNetV1 and supports state-of-the-art for different mobile visuals recognition which includes classification of images, detection of the objects and semantic segmentation. For Tensorflow Slim Image recognition library, this method was published. Based on the idea of MobileNetV1, an efficient building block MobileNetV2 uses depthwise separable convolutional. Although MovbilNetV2 has two newly added features. 1) In between layers, there are the linear bottleneck and 2) Between the bottleneck, there is a shortcut. Below is the fundamental structure of MobileNetV2.

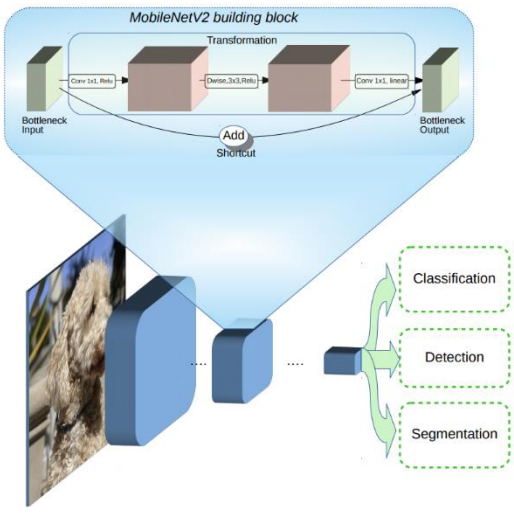

Fig: 6 - Overview of MobileNetV2 Architecture. Blue blocks represent composite convolutional building blocks, as shown above.

The assumptions are that the bottleneck represents the input and output of the algorithm. In contrast, the internal layer represents the power of the algorithm to transform from definitions of the lower level, such as pixels to descriptors of higher level such as image categories. In conclusion, shortcuts, as with remaining conventional links allow quicker training process and better precision.

6 Implementation

6.1 Body key-points extraction using Open pose

Body key-points can be extracted using Open pose. The open pose can be used in a programming language like C++, JavaScript, and python. Since the proposed research project uses python programming language for implementation, the python library is used for Open pose to extract key-points as well. To bring all the variables together, Google Colab has been used. To install open pose into the system, the system needs the proper configuration of CUDA and CudNN 7.5 and windows 10. The system should have a higher level of the graphic card. Since this criterion is not met for our system, Open pose and Nvidia-smi 450.57 has been installed in Google Colab.
Similarly, Cmake -3.172 has been installed as it is an extensible open-source system which runs the process which is built in an operating system. Different libraries like OpenCV, Python3, OpenCL Generic and some dependencies are installed. Cloning of Open pose has been done so that it is faster to use its functionality. A model for COCO data and Pretrained weights is downloaded and uploaded using the library. The open pose gives 14 key points for input images; this key point is generated and stored in the .json file for further implementation. The key points generated is in 2d form, which is converted into regular key points. This key point is stored in ACGPN model as test_pose.

6.2 Self-Correction Human Parsing for generating labels.

Self-Correction Human Parsing is an out-of-the-box human parsing representation extraction released in Oct 2019. This human parsing model tracks images of single, multiple humans as well as video. To extract the human parsing representation (test_label) trained SCPH model is used. LIP Dataset is used for implementation. Dataset Setting has been changed in the final file as we need a parsing representation that would work only on required parts of the body for research. To generate colourmap, different segmentation labels are changed. Since we are interested in imposing clothes on the upper body of the user, preference given to the body parts are changed in the colour palette. The following parts are as per priorities given to the colour palette, i.e., Hairs, Upper-clothes, Face, Left-arm, Right-arm and background. Following is the result generated.

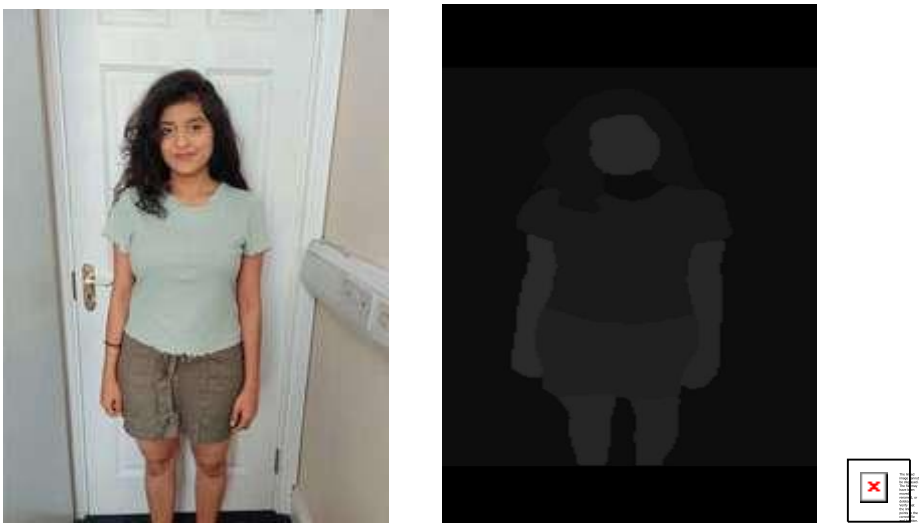

Fig 7. Human parsing representation generated from custom image.

6.3 Adaptive Content Generating and Preserving Network

The ACGPN anticipates the essential details that can be changed in future after virtual attempt from the user image model. When the user chooses a shirt with a long sleeve, then the currently chosen sample shirt (long sleeve shirt-> arm, arm- > jacket,) from the standard arm picture will be added. This decides how the contents of the image will be held for future analysis in order for the model to gather essential details. This preservation can contribute to practical virtual photographic trial and useful descriptions of clothing. The ACGPN has three features: the-First, Semantic Generation Module (SGM), which produces a practical segmentation of the mask of the human body parts and mask of the area of the garments, providing the critical alignment of the dimensional structure. Secondly, in conjunction with the deformed fabric mask, the Clothing Warping Module (CWM) creates the target image deformed. In the CWM (Think Plates Spline) presentation, geometrical similarity will be provided, but the main features of the picture of the clothing will be retained. The third stage is the Content Fusion Model (CFM), which unites the previous element knowledge in order to discover the preservation and development of specific components of the human body in the combined image flexibly. Analysis ACGPM creates realistic simulated images while preserving the core aspects of clothing (texture, logo and brotherhood) and the parts of human identity (pose, body parts and the base clothes), respectively. Three modules (GMM), (CWM) and (CFM) has been specially designed. The VITION dataset has different pictures depending on the pose. After adding test labels and test_pose, we inferred the model and received the results. While inferencing the model, code as only one cloth needs to be imposed, the index number has been changed in the aligned_datset.py.

Fig 8. Imposed clothes on the custom image using ACGPN

7 Evaluation

Qualitative results will evaluate the performance of ACGPN. For testing the performance of ACGPN, having input images in different poses has been ensured.

7.1 Imposing Virtual garment on Easy pose

Fig 9. Qualitative results of Easy pose

By imposing Virtual garment on the easy pose, it can be visualizing that the ACGPN model is working precisely. In FIG. Evaluation can be made that the three-fourth sleeves garment when imposed on full sleeves garment, it does not affect the body part and is adequately imposed on the user. The segmentation results take proper body part of the user as the pose is easy. The result images are clear and have no blurring.

7.2 Imposing garment on Medium pose.

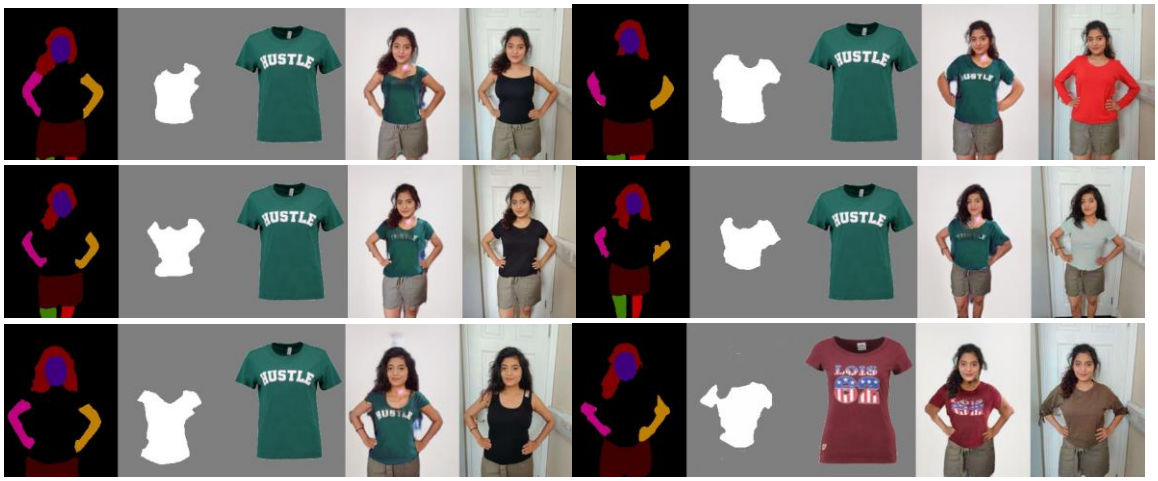

Fig 10. Qualitative result of Medium Pose

In the medium pose, we can visualize that the segmentation result takes proper body parts of the user image. However, the garment imposed on the user's body is not precise enough. There is a colour mixture of hand and garment. Boundary blurring can be visualizing and also the cluttered texture. The garment colour is generated in background. The hands of the user are not properly segmentized.

7.3 Imposing garment on Hard pose.

Fig 11. Qualitative result of Hard pose

In the hard pose, ACGPN model does not work good on custom images. We can visualize that their segmentation representation has some "broken arms", the arms are not segmented properly. There is lots of distortion in the pixels, the clothes imposed are overlapping the

arms, and some strange points are generated near the neck area. The imposed garment should have the same shape as the target clothes, but the imposed garment takes the shape of the garment which the user has worn. From this, the conclusion can be made that ACGPN does not work precisely on hard pose images.

7.4 Discussion

ACGPN model was published five months ago, i.e. on March 2020, there is not much information available on how to generate data variables form custom images. By understanding the data, the structure we found out that the test_label can be generated by using Self-Correction Human Parsing strategies and the segmentation labels of ACGPN is applied in this model. Similarly, we found out the test_pose are the body key points of the input images which can be generated by using Open pose. All these data variables are then imported into ACGPN model for testing. For inference, we made sure that the input images are in a different pose to check the ability of ACGPN to impose garment on the different pose.

The results we got from ACGPN was conclusive by not measurable; they can only be visualizing. ACGPN itself consists of many other pertained models, so find accuracy for it is a difficult task but is not impossible. A change in weights of the Pix2PixHD model, batch size and epochs can give us a good result for Hard pose images also. The experiment gave a solution for using ACGPN on Custom images.

8 Conclusion and Future Work

In this research project, we implemented ACGPN (Adaptively Content Generating and Preserving Network) model for imposing Virtual Garment on users Input Images. The Input Images was divided into 3 categories, i.e. Easy pose image, Medium pose image, Hard pose image. The easy pose in which the model's arms are straight towards the ground. Medium pose in which the model's arms are in the pocket and hard pose in which models' arms are folded or have a similar pose. To use ACGPN on our custom data, we figured out that the test_label can be generated from Self Correction Human Parsing strategies and test_pose can be generated from Open pose. First, the Input Images size is resized to 192x256 since that is the requirement of ACGPN model. The output of this is stored in test_label folder. To generate Body Key points (i.e. test_pose), we implemented Open pose on Input Images and those key points are saved as .json files in test_pose folder.

The test_color is the images of the garment which need to be imposed on the user's image. Test_edge is the test_color's garment edge. Test_colormask and test_mask are to shade the image to make it incomplete so that the network learns to paint it. Tets_img is the input image. All these test files are imported in the ACGPN model for testing. In the ACGPN, we evaluated results by visualizing (i.e. Qualitative Results). We visualize that the model worked accurately on Easy pose image and good on Medium pose image. However, the results for Hard pose images did not turn out good. So, we can conclude our research objective that yes, we can use ACGPN to impose virtual garments on the custom images.

As the paper was published 5 months ago (i.e. March 2020) there is no much information about how the model can be evaluated further. If we can add different weights to Pix2PixHD or batch size or testing using the different number of epochs, the output for Hard pose images can also be accurate. This model can be further converted into the Torchscript using ImageNet for conversion. This Torchscript can be imported in Android or iOS applications using Pytorch Mobile. If the implementation works successfully, online shoppers

can sit at home and try clothes virtually on their mobile application. This will reduce the return and refund problems faced by the company as well as the customer.

Acknowledgements.

I express my sincere gratitude to Prof. Christian Horn for her unrelenting support and encouragement. Thanks again for her immense support, motivation and guidance on my every phase of my thesis.

References

Andriluka, M., Roth, S. and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 623–630.

Bonetti, F., Warnaby, G. and Quinn, L. (2018) 'Augmented Reality and Virtual Reality in Physical and Online Retailing: A Review, Synthesis and Research Agenda', pp. 119–132. doi: 10.1007/978-3-319-64027-3_9.

Bulat, A. and Tzimiropoulos, G. (2016). Human pose estimation via convolutional part heatmap regression, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9911 LNCS: 717–732.

Chen, Y., Shen, C., Wei, X., Liu, L. and Yang, J. (2017). Adversarial posenet: A structure-aware convolutional network for human pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1221–1230.

Germann, M., Popa, T., Ziegler, R., Keiser, R. and Gross, M. (2011). Space-time body pose estimation in uncontrolled environments, 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, pp. 244–251.

Gomes, D. G., Calheiros, R. N. and Tolosana-Calasanz, R. (2015). Introduction to the special issue on cloud computing: Recent developments and challenging issues, *Computers & Electrical Engineering* 42: 31–32.

Hu, P. *et al.* (2019) 'A new method to evaluate the dynamic air gap thickness and garment sliding of virtual clothes during walking', *Textile Research Journal*, 89(19–20), pp. 4148–4161. doi: 10.1177/0040517519826930.

Hu, P. *et al.* (2020) 'A generic method of wearable items virtual try-on', *Textile Research Journal*, pp. 1–14. doi: 10.1177/0040517520909995.

Kune, R., Konugurthi, P., Agarwal, A., Rao, C. R. and Buyya, R. (2016). The anatomy of big data computing, *Softw., Pract. Exper.* 46(1): 79–105.

Mustaffa, M. R. *et al.* (2019) 'Dress me up!: Content-based clothing image retrieval', *ACM International Conference Proceeding Series*, pp. 206–210. doi: 10.1145/3309074.3309121.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(3): 257–267.

Ramanan, D., Forsyth, D. A. and Zisserman, A. (2005). Strike a pose: tracking people by finding stylized poses, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, pp. 271–278 vol. 1.

Shen, Ju & Su, Po-Chang & Cheung, Sen-ching & Zhao, Jian. (2013). Virtual Mirror Rendering with Stationary RGB-D Cameras and Stored 3D background.. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. 22. 10.1109/TIP.2013.2268941.

Shen, J. *et al.* (2013) 'Virtual mirror rendering with stationary RGB-D cameras and stored 3-D background', *IEEE Transactions on Image Processing*. IEEE, 22(9), pp. 3433–3448. doi: 10.1109/TIP.2013.2268941.

Tang, W., Yu, P. and Wu, Y. (2018). Deeply Learned Compositional Models for Human Pose Estimation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11207 LNCS: 197–214.

Wang, B. *et al.* (2018) 'Toward characteristic-preserving image-based virtual try-on network', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11217 LNCS, pp. 607–623. doi: 10.1007/978-3-030-01261-8_36.

Wei, S., Ramakrishna, V., Kanade, T. and Sheikh, Y. (2016). Convolutional pose machines, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724– 4732.

Yang, H. *et al.* (2020) 'Towards Photo-Realistic Virtual Try-On by Adaptively Generating$\leftrightarrow$Preserving Image Content'. Available at: http://arxiv.org/abs/2003.05863.

Yang, L. *et al.* (2017) 'Fully convolutional network with superpixel parsing for fashion web image segmentation', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10132 LNCS, pp. 139–151. doi: 10.1007/978-3-319-51811-4_12.