Virtual Attempt On Human Body Utilizing Single Picture For Clothes

Soriful Momin[1*], Dr Raghaw Raman Sinha[2], Dr Rajneesh Rani[3]

[1*]Dr B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, Email: sorifulmomin786@gmail.com
[2]Dr B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, Email: sinharr@nitj.ac.in
[3]Dr B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, Email: ranir@nitj.ac.in

***Corresponding Author:** Soriful Momin
*Dr B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, Email: sorifulmomin786@gmail.com

Abstract
Artificial intelligence is being implemented, Algorithms based on deep learning and computer vision are used in the newest field of study known as "human body rendering for visualization of clothes" to create realistic images of people wearing various clothes. This technology's main goal is to help customers, internet retailers, and fashion designers in illustrating how clothing will appear without physically trying it on. Using a set of images of people wearing various outfits as training data, a neural network is then used to produce new renderings of humans wearing a certain garment. The constructed images can be used in a variety of contexts, such as online shopping, fashion design, and virtual try-on. The fashion business could be completely transformed by human body rendering technology if it can eliminate the need for real prototypes and enhance customer shopping experiences.

Keywords: Random poses, posture transfer, reconstruction of apparel, and virtual try-on

INTRODUCTION

An innovative technique that combines computer vision and computer graphics approaches to create realistic and exciting visualizations of the human body of a person use of using just a single picture. [1]

In computer vision and graphics, learning human appearance from a single image is an exciting thing. The objective is to extract relevant information about a person's look from a single image, such as their clothes, hair, facial features, and body shape. Due to the complexity and variety of human appearances, this undertaking presents a number of difficulties. Researchers and industrial want to enable a variety of applications by developing algorithms and models to address this problem, such as virtual try-on, augmented reality, creating custom avatars, and image-based virtual makeovers. The way we engage with digital media and improve our visual experiences could be revolutionized by being able to deduce a person's look from a single image.

Extraction of significant visual elements is one of the essential steps in understanding human appearance from a single image.

These characteristics could include physical proportions, facial features, clothing details, and hair characteristics. For the sake of later analysis and modelling, it is essential to precisely and robustly extract these properties. In order to solve this issue, machine learning and deep learning approaches are essential. In a variety of computer vision tasks, convolutional neural networks (CNN, s) and other deep learning architectures have demonstrated outstanding performance. These approaches can learn intricate representations and patterns in human appearance by training models on enormous datasets. [2] In order to bridge the gap between online viewing and buying experiences, virtual try-on technology was developed. Virtual Try on technology creates virtual images or Image is attribute prediction, to predict specific attributes or characteristics, such as clothing kinds, colours, patterns, facial expressions, or gender, attribute prediction models are trained.

These models use the learnt representations of appearance to predict various features of a person's appearance. [3] In the field, generative models are also growing and more popular recently. New samples can be produced using Autoencoders (VAEs) and Generative Variational Adversarial Networks (GANs) that match the appearance seen in the input image. These models can create fresh samples with identical characteristics after learning the basic pattern of appearance features. Utilizing data augmentation and synthesis strategies, appearance learning models' robustness and adaptability are enhanced. Through the use of changes on the input image or the creation of synthetic data that combines appearance attributes from several sources, these strategies require producing extra training instances.

An intriguing and difficult subject, understanding human appearance from a single image requires interdisciplinary work from computer vision, machine learning, graphics. By improving our knowledge and skills in this area, we can open up fascinating applications that merge the real and virtual worlds to produce unique and thrilling experiences.

The emergence of e-commerce has completely changed how consumers purchase by putting a variety of goods at their fingertips and providing convenience. However, the inability to physically try on things before a purchase is a problem that many internet consumers have. Customers may become hesitant and apprehensive as a result of this restriction since they are unaware of how the product will fit, look, and feel in real life. Virtual Try-On (VTON) technology was developed to close the gap between online viewing and purchase experience simulations that allow consumers to Virtually try on things in order to provide them a more full and realistic purchase experience. VTON systems enable customers to see how a product would seem on them in a virtual world by utilizing computer vision, image analysis, and augmented reality (AR). Users can interact with the virtual try-on system to explore several alternatives and determine how well the product is suitable for them, it could be clothing, accessories. The fashion industry, where clients frequently want confidence regarding the fit and appearance of clothing goods, has seen strong uptake of VTON technology. Customers can better predict how clothes will fit and feel by digitally trying them on, which lowers the need for returns and exchanges. This approach enables it achievable to generate a detailed from an input that can be viewed from many perspectives and used in virtual worlds. In industries like e-commerce and fashion, shopping and fashion businesses to more effectively sell their products. Customers are now able to visualize how clothing might appear on virtually mode or themselves instead of relying solely on flat product photos. This makes purchasing more enjoyable overall and gives a more true depiction of fit, style, and look.

Due to the high price and technological difficulties of creating human and garment models, mass-market applications are not feasible. Early 3D computer graphics methods have limited scalability and accessibility due to the process' expensive nature.

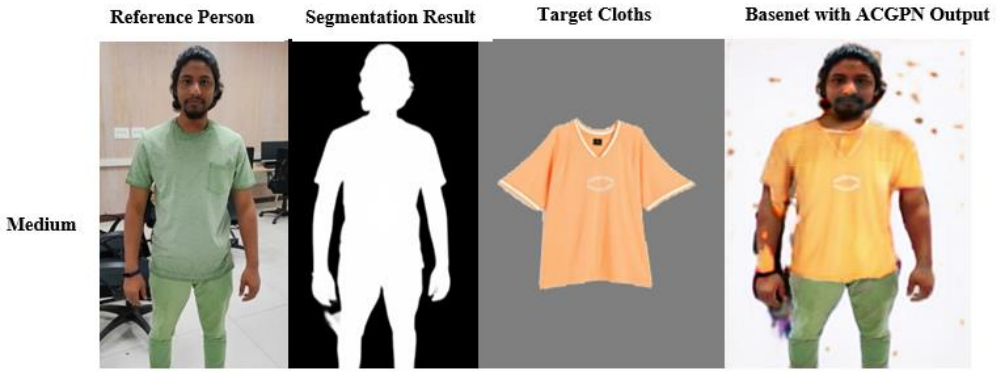

Figure -1

Figure -2

RELATED WORK

Image warping

There convolution networks based the spatial transformation's parameters from an input image A grid generator constructs a grid of sampling locations based on the expected parameters from the localization network. [4] These coordinates specify the transformation that should be applied to each pixel in the input image in order to create the output image. Find corresponding details or landmarks in the target and source images. [5] The correlation between the two images is defined by these features, which serve as control points. Based on the corresponding features found in the previous step, divide the photos into a collection of triangles. [6] The triangulation serves as the basis for pixel warping. Calculate the transformation for each triangle that converts the matching points transferring a desired image from the original image. Affine transformations, thin-plate splines, or projective transformations are all commonly used for warping.

Traditional approach to novel synthesis

Traditional rendering strategies like ray tracing and rasterization depend on complex modelling of the scene in order to produce visuals [6] geometry, substances, and lighting. However, these techniques frequently fail to capture specific

details, complex lighting effects, or realistic-looking materials. [7] By training neural networks to produce visuals directly from data-driven models, neural rendering techniques seek to get beyond these constraints. Convolutional neural networks, also referred to as CNNs, and generative adversarial networks, also known as GANs, are two examples of neural networks with deep learning that can be trained, on substantial picture [8] datasets is the basis for neural rendering. These networks acquire the ability to recognize the intricate connections between scene.

Deep Fashion dataset

Its contains more than 800,000 images covering several types of clothing, including tops, dresses, bottoms, and accessories. [9] The images are from a range of sources and show distinctive appearances, postures, and angles. there is a massive dataset here with 800000 well-posted photographs that are each labelled with 50 categories and 1000 attributes. There are an additional 300000 pairs of cross-pose photos.

For tasks like image categorization and object detection, among these subsets are the Large-scale Clothes Recognition Database, the Fashion Landmark Detection Benchmark for sitting position estimation, and the Fashion Parsing Benchmark for semantic segmentation of clothing items. Evaluation tasks like the In-Shop Clothes Retrieval Benchmark and the Consumer-to-Shop Clothes Retrieval Benchmark are also available in the Deep Fashion dataset. These challenges offer standardized evaluation criteria and processes for specific tasks, including clothing retrieval.

The human body via 3D reconstruction

Digital [10] representations of human beings in three dimensions are made using 3D human avatars and reconstruction methods. Applications such as computer graphics, virtual reality, augmented reality, animation, biomechanics, medical imaging, and virtual try-on experiences all depend heavily on these models and solutions. [11] By creating the SMPL model to a person's body form and then adding apparel to the 3D model, virtual try-on uses the SMPL (Skinned Multi-Person Linear) dataset to recreate the try-on process. The SMPL model captures the shape and position of the human body using a parametric mesh. Obtain the SMPL model, a statistical model that depicts the range of 3D human body shapes and poses. A parameterized representation of the human body is provided by the SMPL model, which includes information on the skinned mesh blend objects, joint locations, and body shape.

Deep generation models VAE

Deep generating models, like [12] Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), can be used to produce new and realistic samples in a variety of domains, including images, text, and music. The underlying distribution of a training dataset is captured by these models, which then produce new samples similar to the training data. A generator network and a discriminator network are the two main parts of deep generating models. The generator network creates samples that resemble the training data from random noise as input. The discriminator network gains the ability to differentiate between actual samples from training data and false ones produced by the generator.

Apparel transfer

A method called garment transfer utilising SwapNet enables the exchange of clothes across various images or people. To achieve this, SwapNet uses generative models, a deep learning-based approach, dataset is gathered that consists of matched photos of people wearing various outfits. Images of the same individual wearing one type of apparel are included in each pair, along with images of the same person wearing various kinds of clothing. With the help of the gathered dataset, a [13] SwapNet model is trained. A generative adversarial network (GAN) architecture, which comprises of a generator network and a discriminator network, is often the model's foundation. The generator creates photographs with the required clothing transfer, whilst the discriminator attempts to tell real images apart from produced ones**.**

Autoencoder to inpaint an image

Using autoencoders, imagine inpainting entails replacing any missing or damaged pixels in an image with accurate material. In order to do this goal, autoencoders, a type of neural network, learn to recreate whole images from incomplete or damaged input pairings of whole photos and their matching masked or corrupted variants. [14] To imitate the inpainting task, some parts of the masked photos have been selectively removed or changed. In order to train an autoencoder model, the prepared dataset is used. From the representation of the latent space, a decoder network reconstructs the entirety of images. in the autoencoder, it consists of a network encoder that encodes the input images and maps them to a latent area with a lesser dimension. In order to reduce the reconstruction error between the input and output images, training is done. Image Input Part The trained autoencoder model is fed a partially seen or corrupted image to conduct inpainting. The model represents the input image in latent space by encoding it.

Cloth retrieval

Resent works like as relevance feedback, assemble a collection of photographs of apparel that will be utilized for retrieving clothing. The clothing items in the dataset should come in a variety of [15] colours, patterns, and styles and be photographed from various perspectives. Relevant metadata, such as garment type, [16] color, and style, should be included to each image. It’s vital to point out that because there is such much variation in clothes appearance, style, and view, retrieving clothing using computer vision is a difficult task. Effective feature extraction methods, effective feature preprocessing approaches, and good indexing and retrieval algorithms are needed. The development of cutting-edge

techniques, including deep metric learning and attention processes, SHIFT [17] HOG [18] is a current area of research that aims to increase the precision and resilience of textile retrieval systems.

METHOD

Problem formulation for synthesizing

The process of creating new photos from an existing input view, which captures the subject from several angles or perspectives, is known as the problem formulation for synthesising distinct views of a person. The objective is to construct integrated and believable representations of the subject from unobserved angles. Common names for this activity include view generation and novel view creation.

Learning a transformation or mapping from the input view to the goal and innovative views is the main issue. The person's perspective, stance, and appearance should all be taken into consideration during this transition. The vision synthesis problem is frequently addressed using deep learning approaches like convolutional neural networks (CNNs) or [12] generative adversarial networks (GANs). Large datasets of paired or unpaired images showing the subject from various angles are used to train these algorithms. The models acquire the ability to encode the subject's visual traits and provide lifelike representations from difficult positions. Loss Function during training, a loss function is created to quantify how closely the synthesised images resemble the real-world photos taken from the target views. Commonly used loss functions that urge the generated images to be visually similar to the target views include pixel-wise L1 or L2 loss, perceptual loss, or adversarial loss.

The Clothes Wrapping

Module (CWM) is a method that simulates the act of wrapping clothes around a virtual body model to allow for virtual clothing try-on. It makes use of physics-based modelling and the idea of cloth simulation to produce accurate garment draping and fitting on a digital avatar or human body model.

The "Cloth-Model Wrapper" [18] (CMW) is a well-liked Clothes Wrapping Module implementation. The CMW software module connects with modelling or animation tools and offers features for simulating the application of virtual clothing to a virtual body.

The wrapped clothing is seen and depicted either in real-time or as a pre-rendered animation. This covers methods for accurate and visually appealing apparel rendering, lighting, shading, texture mapping, and rendering.

Virtual try-on systems and applications connected to fashion require the Clothes Wrapping Module, such as CMW. It enables users to picture how various clothes would fit and appear on a virtual body without actually trying them on. It gives users a realistic portrayal of clothing and assists them in making judgements about the clothes they want by imitating the draping process.

Segmentation using BasNet

A deep learning model called BasNet was created for the task of segmenting images, with an emphasis on the task of object and border segmentation. According to the research paper titled BASNet Boundary-Aware Salient Object Detection it stands for Boundary-Aware Salient Object Detection."

The [20] BasNet model performs pixel-level segmentation of objects in an image using a deep neural network architecture. It seeks to distinguish the background from the conspicuous objects by precisely identifying them and their borders. Various computer vision tasks, including object detection, image manipulation, and scene comprehension, can benefit from this.

Encoder: Using convolutional layers, the encoder network receives an input image and extracts high-level feature representations. It extracts structural features—information about object boundaries, textures, and shapes—from the input image.

Decoder: The decoder network reconstructs the segmentation map using the encoded information obtained from the encoder. It combines low-level and high-level feature representations via upsampling and skip connections, which enables accurate localization of object boundaries.

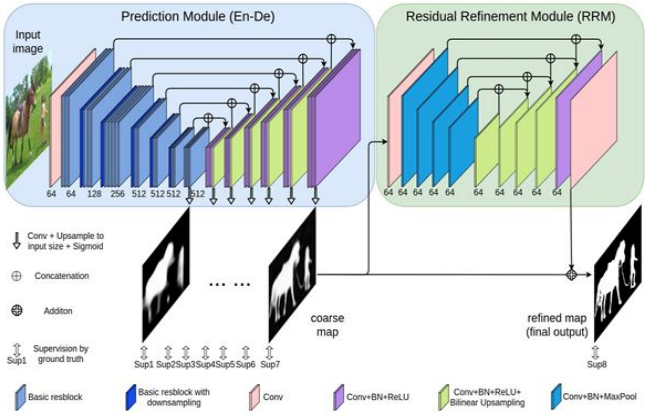

Figure – 3, Architecture of Bas Net

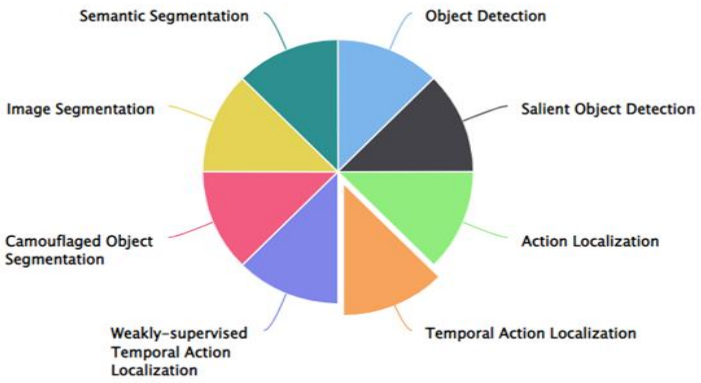

Figure -4, use case of Bas Net

Models pertaining to Semantic Generation.

Image visual try-on, which seeks to project a desired outfit's image onto a reference Figureure, has gained popularity recently. Prior artworks typically concentrate on maintaining the personality of a clothing image when warping it to an arbitrary human position (for example, texture, logo, embroidery). However, creating photo-realistic try-on photos still proves to be difficult when the reference person has prominent occlusions and human poses. We suggest an innovative visual try-on called A Network for Adaptive Content Creation and Preservation to address this problem. In specifically, ACGPN evaluates whether image information has to be generated or maintained based on the [20] semantic arrangement of the reference picture that will change after try-on long sleeve shirt-arm, arm-jacket, expected semantic layout, which results in a try-on that is photorealistic and has rich apparel details. ACGPN typically consists of three main parts. In order to gradually forecast the intended semantic layout after try-on, a semantic layout synthesis module first makes use of semantic segmentation of the reference image. A second-order difference constraint is introduced to the clothes warping module, which then warps an image of clothing in accordance with the resulting semantic pattern. Thirdly, a content fusion inpainting module combines all available data (such as a reference image, a semantic layout, and warped clothing) to produced.

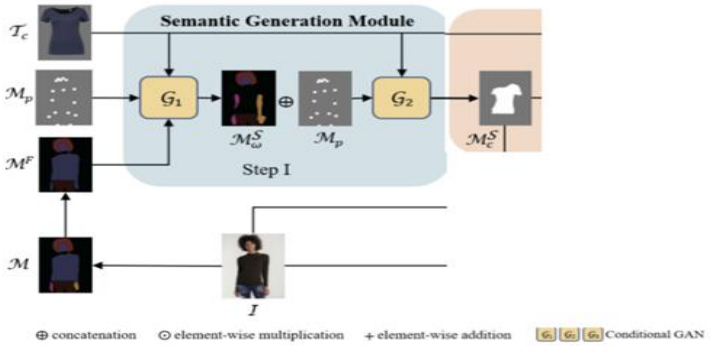

Figure -5

EXPERIMENTS

Dataset

ACGPN (Attribute-Controlled Clothing Generation and Person Image Editing) was used in a specific experiment using the VITON dataset. ACGPN is being compared against many techniques in the trial, including VITON, CP-VTON, and VTNFP (Virtual Try-On through Neural Feature Points). The VITON dataset, which contains over 19,000 image pairs, is used in this experiment. Each pair consists of an image of a woman from the front and an image of top garments. The dataset is downsized to 16,253 pairs after invalid image pairs are removed. A training set of 14,221 pairings and examine set of 2,032 pairs are created from these pairs. By analyzing visual outcomes and doing quantitative comparisons, ACGPN is contrasted with alternative approaches. Despite the absence of the VTNFP's official code, the visual outcomes provided in comparison purposes. The project probably demonstrates the potential of ACGPN in terms of virtual try-on, image manipulation, and clothing production. The research article or experiment report's appendix or supplemental material contains the comprehensive ACGPN try-on results. It is advised to consult the original research publication or experimental literature that outlines the ACGPN method and its application on the VITON database in order to have a more thorough understanding of the particular experiment and its outcomes.

Implementation Details.

Previous author Han Yang done good work virtual Try on, but there some gap so we can try to improved it, [21] He does his work ACGPN. And SGM, [18], CWM, and CFM, all those go to generator like [22] U-NET articture, we are replacing [20] BasNet articture after that use [23] pix2pixHD, then structure of STN and CWM. Share same spirit of U-Net it design.

Due to the ability of these type of architectures to simultaneously capture high-level global contexts and low-level details, we chose to implement our segmentation prediction module in an encoder-decoder form. The final layer of each decoder stage is monitored by the GT, which was influenced by lessen over-fitting The encoder includes six stages made up of fundamental res-blocks and an input convolutional layer. ResNet-34 [Figure -4] is used to create the input CNN, s layer and the first four stages.

We are implement [24] VITON Dataset using BasNet implementation.
Result of Segment input and output image in medium and Hard pose.

Figure -6

Result of BasNet + APCGN

Figure -7

Qualitative Result

1st time we work APCGN with BasNet using VITON Dataset we get good result BasNet perform like U2Net but BasNet provide more image Details less time as we can see the result Figure-1,2,7 and Table -1.

Figure -8

Table 1

Method

Method	All	Easy	Medium	Hard	IS
VITON[12]	0.783	0.787	0.779	0.779	2.650
CP-VTON[42]	0.745	0.753	0.742	0.729	2.757
VTNEP [49]	0.803	0.810	0.801	0.788	2.784
ACPGN+	0.825	0.834	0.823	0.805	2.805
ACGPN*	0.826	0.835	0.823	0.806	2.789
ACGPN	0.845	0.854	0.841	0.828	2.829
BasNet + ACGPN	0.851	0.852	0.845	0.795	2.831

Table 1
The SSIM [25] and IS [26] result of six methods,
ACGPN+ and ACGPN* for ablation research

GitHub–
https://github.com/sorifulM/BasNet_ACGPN

User Study
A well-liked dataset for applications involving virtual try-on is the [24] VITON dataset. It is made up of pairs of photos, each of which features a person wearing a distinct outfit. The collection offers the details required to carry out virtual try-ons by combining the apparel from one image onto the subject in the other. Its work for clothes we can try some image if person wearer hodies its not properly replace [Figure-8] so we try better extraction and better mapping another we can extent this data for shoes, cap, watch, for virtually Try on task.

CONCLUSION
In this work we applied novel approach BASNet to Generate Mask less time and more Person Details [Figure - 7]
Our aim to generate more realistic Virtual Try on.
We demonstrate the efficiency of BasNet with ACGPN and the value of the VITON dataset by comprehensive experiment, which may greatly assist future research.

REFERENCES

1. Hongqiang Zhu, " Research and Development of Virtual Try-On System Based on Mobile Platform" IEEE Transition, 2017. [1]
2. Xintong Han, "An Image-based Virtual Try-on Network" IEEE Transition 2018[2]
3. Ian J. Goodfellow, "Generative Adversarial Nets" arXiv 2014. [3]
4. Jimmy SJ Ren, "Shepard convolutional neural networks". In Proc. NIPS, pages 901– 909, 2015. [4]
5. Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. "Sparsity invariant cnns. In International Conference on 3D Vision "(3DV), pages 11– 20. IEEE, 2017. [5]
6. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. "Free-form image inpainting with gated convolution". arXiv preprint arXiv. [6]
7. Buehler, C., Bosse, M., McMillan, L., Gortler, S.J., Cohen, M.F. "Unstructured lumigraph rendering" In: SIGGRAPH ,2001. [7]
8. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F. "The lumigraph. In: SIGGRAPH". p. 43–54,1996. [8]
9. Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang, Wang, Xiaoou Tang, "Powering Robust Clothes Recognition and Retrieval with Rich Annotations", IEEE transaction, 2016. [9]
10. Helena A. Correiaa, José Henrique Brito, "3D reconstruction of human bodies from single-view and multi-view images" Elsevier, 2023. [10]
11. Thai Thanh Tuan, Matiur Rahman Minar, Heejune Ahn "Multiple Pose Virtual Try-On Based on 3D Clothing Reconstruction" IEEE transition, 2023. [11]
12. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets in Advances in neural information processing systems", pages 2672–2680, 2014. [12]
13. Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu "Image Based Garment Transfer", ECCV, (2018), [13]
14. Mohammad H. Givkashi1, Mahshid Hadipour, Arezoo PariZanganeh, Zahra Nabizadeh4, Nader Karimi, Shadrokh Samav "Image Inpainting Using AutoEncoder and Guided Selection of Predicted Pixels", arXiv,2021. [14]
15. Wang and T. Zhang. "Clothes search in consumer photos via color matching and attribute learning". In ACM MM, pages 1353–1356, 2011 [15]
16. H. Chen, A. Gallagher, and B. Girod. "Describing clothing by semantic attributes". In ECCV, pages 609–623. 2012. [16]
17. D. G. Lowe. "Distinctive image features from scale-invariant keypoints". IJCV, 60(2):91–110, 2004. [17]
18. N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In CVPR, pages 886–893, 2005. [18]

19. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. “In Advances in neural information processing systems”, pages 2017–2025, 2015[19]
20. Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, “Boundary-Aware Segmentation Network for Mobile and Web Applications” IEEE transition, 2021, [20]
21. Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, Ping Luo “Photo-Realistic Virtual Try-On by Adaptively Generating” CVPR, 2020, [21]
22. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “Convolutional Networks for Biomedical Image Segmentation”, arXiv,2015, [22]
23. Dan Jin, Han Zheng, Qingming Zhao, Chunjie Wang, Mengze Zhang and Huishu Yuan “A Deep-Learning-Based Image Enhancement Approach”, PDPI, 2021, [23]
24. Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis “An Image-based Virtual Try-on Network”, CVPR, 2018, [24]
25. Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. [25]
26. Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In NIPS, pages 2226–2234, 2016. [26]
27. Cloth reference from ecommerce link - https://www2.hm.com/en_in/productpage.1165808002.html https://www2.hm.com/en_in/productpage.1191242001.html