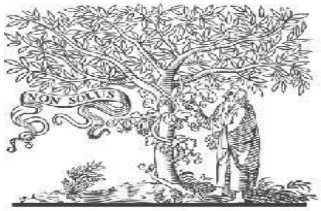
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 05th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

10.48047/IJIEMR/V12/ISSUE 04/18

Title **VIRTUAL TRYON**

Volume 12, ISSUE 04, Pages: 133-140

Paper Authors

Jeevan Babu Maddala, Sk.Lathisha, V.Mounika, V.Vamsi Krishna , V.Lakshmi Prasanna

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Virtual TryOn

Jeevan Babu Maddala[1], Assist.Professor, Department of CSE,
VasireddyVenkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

Sk.Lathisha[2], **V.Mounika**[3], **V.Vamsi Krishna**[4] , **V.Lakshmi Prasanna**[5]
[2,3,4,5] UG Students, Department of CSE,
VasireddyVenkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
jeevan@vvit.net[1], lathisha310@gmail.com[2], vintipallimounika@gmail.com[3],
vallepu670@gmail.com[4], prasannavallurulakshmi@gmail.com[5]

Abstract

Regardless of how convenient shopping clothes online is, c ustomers frequently worry about how a piece of clothing present in the image might appear while making purchases of clothing .So, enabling customers to try on clothing virtually improve their buying experience, revolutionizing how people shop while also saving the money for merchants because of reduced product returns. Without utilizing any sort of 3D data, we described a Virtual Try-On which is image-based network. Through the use of a novel Geometric Matching Module, VTON first learns a thin-plate-spline transformation for tailoring the in-store clothing to the target person's body shape. To achieve flawless integration of the warped clothing and produced image which increases the likelihood that the outcomes will be realistic, we used a Try-On Module that trains on a composition mask.

Keywords: Virtual Try-On, Thin-Plate-spline, Composition mask, Convolutional Neural Network, Geometric Matching Module, Try-On Module

Introduction

The advent of e-commerce completely altered how we shop by allowing us to buy anything we want from anywhere in the world with just a few easy clicks. The inability to try products on before making a purchase has been one of the biggest problems with online shopping, especially for clothing and accessories. Virtual try-on technology has emerged as a solution to this problem, allowing customers to preview how items might look on them before making a purchase. In this study, we give a thorough analysis of virtual try-on technology, covering its uses and difficulties. The term "Virtual Try-On" refers to a technology that lets consumers visualize how an item of clothing would look on them without trying it on. Because it improves the online shopping experience, lowers the amount of product returns, and boosts consumer happiness, this technology has gained popularity in the fashion business. To build a virtual depiction of the user and the item being tried on, Virtual tryon employs augmented reality, machine learning and computer vision technologies.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Figure 1: Each row depicts a person virtually trying on various pieces of clothing.

For developing and testing virtual try-on models, there is a dataset called Viton (Virtual Try-On). In order to give customers, the opportunity to virtually try on clothing before making a purchase, virtual try-on models are frequently developed and evaluated using the Viton dataset. Through the application of sophisticated algorithms, virtual try-on technology can accurately and realistically reflect the product on the customer's body. Because it gives customers a convenient and interactive shopping experience and eliminates the need for in-person fittings, which can be time-consuming and expensive, this technology has grown in popularity recently.

Literature Survey

The framework described in this [1] uses deep learning to simultaneously perform human body parsing and pose prediction from a single image. This study also presents the Look into Person (LIP) benchmark, which offers a uniform way to assess how well human body parsing and position estimation algorithms perform. This paper's key contribution is the creation of an integrated framework that can perform body parsing and position estimation from a single image. For both parsing and pose estimation, our model uses JPP, which only uses a single network. Because there is no need to employ multiple networks for different tasks, image processing may be completed more rapidly and efficiently. In order to do semantic image segmentation, which involves providing semantic labels to each pixel in an image, this [2] paper offers a deep learning-based method. The paper's main contribution is the creation of a deep learning-based approach for semantic image segmentation that outperforms the PASCAL VOC 2012 benchmark. The two main parts of the proposed methodology are a fully connected conditional random field (CRF) for segmentation result refinement and a deep convolutional neural network (CNN) for feature extraction. This study also introduces PASCAL VOC 2012, a brand-new benchmark dataset for assessing the effectiveness of semantic segmentation methods. To get enhanced virtual try-on results over CRFs, JPP network has been used. This research's [3] primary contribution is creating a deep learning-based method for multi-

person pose estimation that delivers innovative results on the COCO benchmark. The described method can handle numerous people in the same image and estimate their poses in real-time. This paper also introduces COCO, a brand-new benchmark dataset containing ground-truth annotations for body parts and poses and photos of numerous people. This paper [5] suggests a self-supervised learning approach for human parsing that can discover the underlying structure of human body parts and their connections. The paper's main contribution is the application of a unique "structure-sensitive learning" self-supervised learning method, which enables the model to learn the structural links between various body components without the need for explicit annotations. The technique accomplishes this by taking advantage of the training data's built-in spatial and semantic relationships between various bodily parts. This paper [4] presents a deep learning-based method for creating pictures of people in various positions. The article's important focus is the use of a pose-guided picture creation method, which makes it possible to create realistic photographs of people in various stances without having to use a lot of training data. Our model has a major advantage over [4] is that it can generate a try-on image from a source image and a target garment image in a single step, whereas in [4] , it usually requires a two-step process that involves creating a person image conditioned on a pose and then using that image to create a try-on image.

Methodology

We treat the image-based virtual experiment task as a conditional imaging problem. In general, given reference images L_i of a person wearing clothing C_i and a target clothing C, our model's intended purpose is to create a fresh new image L_o of the user in a new clothing C_o, where the shape of the body and pose are preserved, the attributes of the target garment C are inhibited, and the impact of C_i's old garment is discarded. Training takes place with the trial triplets (L_i, C, I_t).Here C is paired with It wearing C_t, is simple but not practical because it is challenging to obtain these triplets and I_t represents the reality of L_o . In order to simplify things, we equalize L_i and I_t .This indicates that pairs of C, I_t are adequate. However, direct training (I_t, C, I_t) impairs the generalizability of the model in the testing phase when only disconnected inputs (L_i, C) are available. Previous work addressed this problem by constructing a clothing-agnostic person representation P to remove the influence of the source clothing C_i. This representation is used in our approach and we improved it by omitting fewer details from the reference image of the person. The substantial spatial inaccuracy is one of the difficulties in image-based virtual experimentation between the in-store garment and the user's body. Existing network topologies (such as ResNet, UNet, and FCN), which

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

cannot accommodate significant spatial deformations, produce ambiguous experimental results in conditional image generation. We bring forth a geometric matching module (GMM) to create a deformed image of the clothing by directly matching the input clothing C with the human representations P. A pixel-by-pixel L1 loss was used to directly train the end-to-end neural network known as GMM. The Try-On module combines the distorted clothing C and the created personal image I_r to create the final test results,Io.

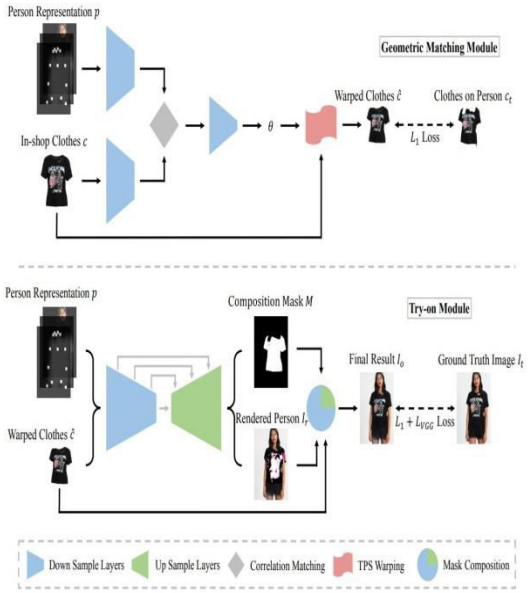


Figure 2: Architecture of GMM and TOM.

Virtual TryOn Network

Person Representation: Both our geometric matching module and try-on module make use of this person representation. In order to preserve as much of the input person image information as possible, which primarily comprises the face, hair, shape of the body, and pose, this depiction seeks to leave away the effects of the old clothing C_i, such as its tone, form, and texture. It consists of the following three parts:

1. Posture heatmap: A feature map which is a collection of 18 activation maps produced by applying 18 different filters to an input image. It is represented by an 11 x 11 rectangle and has each channel matching to a key point in a human position.

2. Shape of the body: A feature map of a binary mask which is hazy or blurry that roughly covers the various body parts. The input data is represented as a one-dimensional array or sequence, and each element in the array corresponds to a single feature.

3. Reserved regions: An RGB image that includes the parts of the face and hair that must remain blank to preserve a person's identification.

Geometric Matching Module: Three stages make up the traditional method for the geometry estimating task of matching images. Initial stage is descriptors like Scale-Invariant Feature Transform, which is a feature extraction algorithm used in computer vision to identify and describe local features in images. In the next stage, in order to create a set of preliminary correspondences, the descriptors are matched across both the input images. Lastly, these relationships are utilized to accurately predict the geometric model's parameters. Our new Geometric Matching Module is created in response to this

work. Here P is used to indicate the person representation, target clothing by C and warped clothing by ^C. Four parts make up our GMM.

1. For the purpose of retrieving P and C 's high-level characteristics, two networks are utilized.

2. To merge two attributes into one tensor and give it as input to the regressor network, correlation layer is used.

3. A regression network to forecast the parameters of the spatial transformation.

4. Thin Plate Spline which is represented as T.

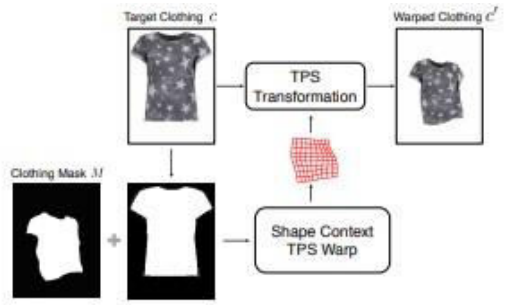


Figure 3: Warping of a clothing image.

Try-on Module: Just putting warped clothing ^C onto the target person's image It is one simple fix. One more option is to use encoder-decoder network, namely UNet which has symmetric architecture, where the decoder part has the same number of layers as the encoder part, which is preferable for rendering seamless, smooth visuals. Our TOM tries to combine the benefits of the two strategies. Concatenating the depiction of the person P and the twisted clothing ^C as an input to UNet, then it produces an image of a person I_r and anticipates a composition mask represented as M at the same time. Finally, both I_r and ^C are united in one using the mask M to amalgamate the final output I_0.

Implementation

During our experiments, we kept λ_{L1} and λ_{vgg} at a value of 1, while setting λ_{mask} to 1 when using a texture mask. The batch size for training the geometry matching module and the try-on module was four, and we trained both modules for 200,000 steps using the Adam optimizer with β_1 = 0.5 and β_2 = 0.999. We set the initial learning rate to 0.0001 and kept it constant for the first 100,000 steps. After that, we decreased the learning rate linearly until it reached zero for the remaining steps. All input and output images were resized to a resolution of 256 × 192.

Regarding the geometric matching module, the feature extraction networks for human representation and clothing were structurally similar, consisting of four 2-chain subsampling convolutional layers and two 1-chain layers with varying filter numbers. The regression network contained two 2-chain convolution layers, two 1-chain layers, and a fully connected output layer. The only difference between the two networks was the number of input channels, and the output size of the fully connected layer was 50, which predicted the x and y coordinate offsets of the tps anchor points.

For the try-on module, we used a 12-layer unet network with six 2-bit subsampling convolutional layers and six upsampling layers. To reduce the occurrence of "checkerboard artifacts," we replaced the 2-band deconvolution layers used in upsampling with a combination of nearest interpolation layers and 1-band deconvolution layers. The number of filters for downsampling convolutional layers varied, while the number of filters for upsampling convolutional layers was fixed.

Dataset

The dataset used here is VITON , which is present in the Kaggle repository. VITON dataset was created by Richard Kuo. 16,253 pairs of images constitutes this dataset. A training set and a testing set have been created from them. 14,221 pairings make up the training set, whereas 2,032 pairs make up the testing set. Images are 192 x 256 pixels in size.

Results

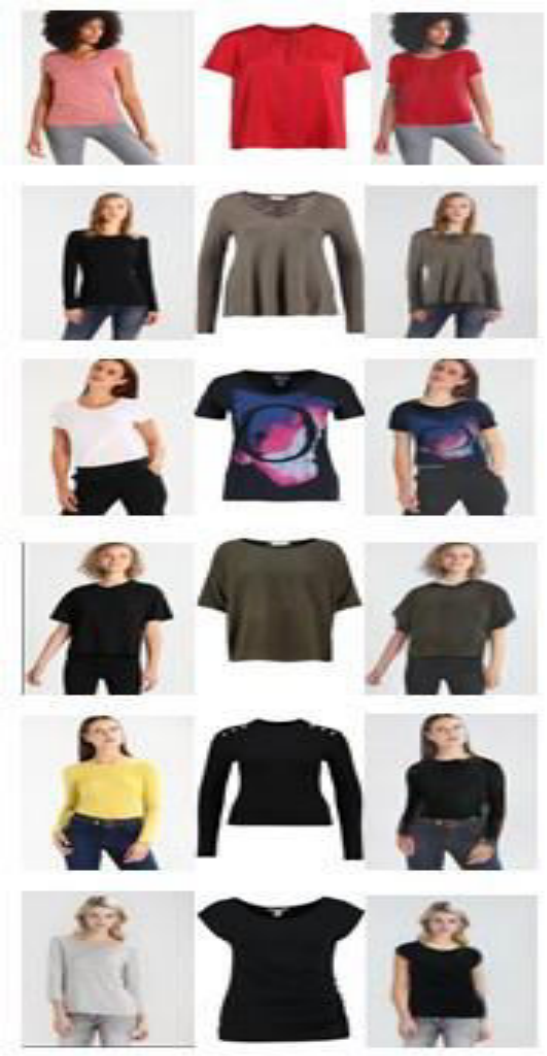


Figure 4:Rendering target clothing on to a person using our Virtual tryon network.

The metrics used to evaluate our model are SSIM, LPIPS, Inception Score.The SSIM Index is a metric used to measure the quality of an image.The Learned Perceptual Image Patch Similarity (LPIPS) is a method used to evaluate the perceptual similarity of two images.

SSIM	LPIPS	IS(mean ± std)
0.7788	0.1387	2.7800 ± 0.0584

Table 1: quantitative assessment on the VITON dataset.

Conclusion

In summary, virtual try-on technologies are transforming the fashion sector by giving customers a more enjoyable, personal, and useful shopping experience. These developments enable users to virtually try on clothing before making a purchase by recreating the look and fit of clothing on their bodies using computer vision, machine learning, and some other reducing techniques.

Limitations

Some poses don't produce reliable outcomes like our system suffers from different style of cloth and human pose especially cross-armed positions in three dimensions.However, some clothing items, such as those with complex detailing or texture, could be more challenging to replicate.

Future Scope

To handle occlusion in virtual try-ons, we can investigate the use of the reference person's 3D structure.The use of 3D structure might solve many issues as well.

References

[1] J. Dong, Q. Chen, X. Shen, J. Yang and S. Yan, "Towards Unified Human Parsing and Pose Estimation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 843-850,2014

[2] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2018.

[3] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021.

[4] W. Zhao, Q. Xie, Y. Ma, Y. Liu and S. Xiong, "Pose Guided Person Image Generation Based on Pose Skeleton Sequence and 3D Convolution," *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020, pp. 1561-1565, 2020.

[5] K. Gong, X. Liang, D. Zhang, X. Shen and L. Lin, "Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6757-

6765,2017.

[6] S .Belongie ,J. Malik , J. Puzicha , "Shape matching and object recognition using shape contexts. IEEE transactions on pattern analysis and machine intelligence" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27 , No.11 PP. 1832 – 1837, 2005.

[7] X. Han, Z. Wu, Z. Wu, R. Yu and L. S. Davis, "VITON: An Image-Based Virtual Try-on Network," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7543-7552, 2018.

[8] Y. Choi, M. Choi, M. Kim, J. -W. Ha, S. Kim and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8789-8797, 2018.

[9] Q. Chen and V. Koltun, "Photographic Image Synthesis with Cascaded Refinement Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 1520-1529, 2017.

[10] K. Gong, X. Liang, D. Zhang, X. Shen and L. Lin, "Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6757-6765, 2017.

[11] P. Isola, J. -Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5967-5976, 2017.

[12] C. Lassner, G. Pons-Moll and P. V. Gehler, "A Generative Model of People in Clothing," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 853-862, 2017.

[13] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng and S. Yan, "Perceptual Generative Adversarial Networks for Small Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1951-1959, 2017.