Cloth segmentation, styling and pose transfer for different body types

Aadhya Raj
aadhya19002@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Aparna Jha
aparna19410@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Arihant Singh
arihant19298@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Harshit Singh
harshit19424@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Smiti Chhabra
smiti19112@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Abstract

Recent years have witnessed the increasing demand for online shopping for fashion items. Despite the convenience online fashion shopping provides, consumers are concerned about how a particular fashion item would look on them when buying apparel online. Thus, allowing consumers to virtually see how the clothes look on different body types will enhance the shopping experience, transforming the way people shop for clothes. This project aims to style a piece of clothing on different body types to help with this issue and further extend it to incorporate different poses.

1. Introduction

1.1. Motivation

This pandemic saw an increase in online shopping due to mobility restrictions. Ordering something with a click from anywhere and getting it delivered wherever you want is a luxury that saves a lot of time and effort. However, one major challenge is wondering how the product would look on us. This served as the motivation for this project. This project will help us try different styles of clothes with different poses. We believe this will solve one of the significant issues with online shopping and help increase its reach, comfort, and accessibility.

1.2. Problem Statement

Using our project, we aim to solve the problem of imagining ourselves in clothes. We are trying to develop a system that segments clothes from images, styles them by allowing them to picture themselves in the custom input cloth, and applies pose transfer techniques to get different poses that output on different body types. We have developed a system that styles clothes and displays the output in different poses.

1.3. Challenges Involved

- While obtaining the cloth mask from different images from the dataset, the hair could be present on top of the dress. In that case, transferring the dress from one pose to another could lead to mismatches because the algorithm fails to determine what could be present in the location where hair was present on top of the dress.
- Transferring initial poses to complex poses could lead to unclear output poses or being transferred improperly because the poses are rarely present in the dataset.

2. Related Work

- **Paper 1** [1]: Research on Interactive cloth segmentation has been very active for the past few years. Traditional cloth segmentation methods usually rely on hand-crafted features and are often computationally expensive. But with recent technological advancements, many researchers developed methods for cloth segmentation using deep learning-based approaches. However, these approaches require a large amount of labeled data and may not be suitable for interactive applications. Researchers proposed a new method combining interactive segmentation with style transfer to counter this limitation. Style transfer involves transferring the style of one image to another while preserving its content. So here, the researchers use style transfer to generate a segmentation mask for an input image. In this paper, the proposed methods consist of two main steps:
  1) Saliency Detection: This step identifies the most

salient regions of the input image using a deep convolutional neural network. Now, the saliency map is used to guide the style transfer process.
2) Style Transfer: Now, the researchers use a cGAN (conditional generative adversarial network) to generate a segmentation mask for the input image. The results show that the proposed method outperforms many techniques regarding computational efficiency and segmentation accuracy. This proposed method can generate high-quality segmentation masks in real-time, which can be used for interactive applications, VR applications, etc.

- **Paper 2** [2]: The paper "A Review on Applications of Deep Learning Techniques in Speech Emotion Recognition" by Hanying Wang, Haitao Xiong, and Yuanyuan Caithis proposes an interactive image localized style transfer method, especially for clothes, by the use of outline image, which is extracted from content image by interactive algorithm. FIrstly a rectangle is made around the desired clothing after which an outline loss function is generated using the distance between the rectangle and the desired clothing. This method generates the new style only in the desired clothing part rather than the whole image including the background which helps in preserving the original clothing shape.

- **Paper 3** [3]: In this paper, an algorithm called “StyleBank" is proposed which is basically a CNN model composed of multiple convolutional filters having a new style per filter. For the implementation, the filter bank corresponding to the specific style to be transferred is convolved on top of the intermediate feature embedding produced by a single auto-encoder, decomposing the image into multiple feature response maps and thus providing an efficient mechanism for style transfer. StyleBank and the autoencoder are jointly learnt in the feed-forward network which helps in learning new styles by learning a new filter bank while keeping the auto-encoder fixed.

- **Paper 4** [4]: In this paper by Dae Young Park and Kwang Hee Lee, they present a novel style-attentional network (SANet) and decoders for efficient style transfer according to the semantic spatial distribution of the image. The algorithm utilizes a new identity loss function and multi-level feature embeddings to preserve the structure of the image content while posing style transfer. The SANet uses a learnable similarity kernel that represents content feature map as a weighted sum of style features, using the identity loss function during training. This helps to maintain the content structure while efficiently adapting the style onto it.

- **Paper 5** [5]: In the paper by George. A. Cushen and Mark. S. Nixon, they present a real-time clothing segmentation model for video and single images. It initializes points on the upper body clothing instead of detecting the face of the subject using distance metrics which is advantageous because it helps prevent skin segmentation instead of clothing. It takes advantage of intensity and hue histograms for efficient segmentation.

- **Paper 6** [9]: In the paper by Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, Ping Luo, Harbin Institute of Technology, SenseTime Research, Tencent AI Lab, The University of Hong Kong, Adaptive Content Generating and Preserving Network(ACGPN) are used for virtual try-on. This method is especially effective for complex poses or poses with hidden body parts. The working is a little different as it first considers the overall structure of the image to be changed and then accordingly decides to generate or preserve image content. In short, it first looks at the layout of the reference image, then the new cloth is wrapped according to this, and lastly, all this information is put together to get the desired output.

- **Paper 7** [10]: This paper by Aiyu Cui, Daniel McKee, and Svetlana Lazebnik proposes a flexible person generation framework called Dressing in Order (DiOr), which supports 2D pose transfer, virtual try-on, and several fashion editing tasks. They represent each person as a (pose, body,{garments}) tuple. For the implementation, they ran an off-the-shelf human parser on the source garment to obtain the masked garment segment. To perform pose transfer, they set the body image and the garment set to be those of the source person and render them in the target pose. This method also has some limitations and failures. The complex or rarely seen poses are not always rendered correctly, unusual garment shapes are not preserved, some ghosting artifacts are present, and holes in garments are not always filled in properly.

- **Paper 8** [11]: In this paper by Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and

Xiang Bai on progressive pose attention transfer for person image generation, Pose-Attentional Transfer Blocks were used for transferring poses. They consider the appearance features along with the pose using progressive pose transfer. Further, a Pose-Attentional Transfer Network is proposed to make this process efficient and straightforward that can be extended to non-rigid bodies. Two datasets are used, along with a new metric kispecifically designed to evaluate the shape consistency.

- **Paper 9** [12]: In the paper Yining Li, Chen Huang, and Chen Change Loy take the source and target image and first use a variant of u-net to encode the image and target pose. Then using the reference and target pose, a 3-d flow map is created. To get an image in target pose, the features of the reference image are warped through a 3-d flow map and then to a visibility map. A visibility map is created to consider the lost pixels due to occlusions. Then concatenating the warped image features and target pose features take place to obtain the pose-transferred image.

- **Paper 10** [13]: In the paper by George. A. Cushen and Mark. S. Nixon, they present a real-time clothing segmentation model for video and single images. It initializes points on the upper body clothing instead of detecting the face of the subject using distance metrics which is advantageous because it helps prevent skin segmentation instead of clothing. It takes advantage of intensity and hue histograms for efficient segmentation.

3. Methodologies

3.1. Techniques/Algorithms

Our methodology can be divided into three steps. The first step is to create a saliency map generation and do cloth segmentation on the input reference person. The second step is to transfer the style to the input person. The last step is to blend the stylized image. Then we can move on to the pose transfer part. This part takes the stylised image as input and performs a pose transfer algorithm to that image, and outputs 40 different image having 40 different poses on the stylised image.

3.1.1 Saliency map generation and Cloth Segmentation

We used a pre-trained U-2-Net model for Cloth Segmentation. The model was trained on (this) dataset. The model accurately segments cloth components based on upper body cloth, lower body cloth, whole body cloth and background. The Saliency map generation was also done by the pre-trained U-2-Net model.

3.1.2 Style Transfer

We used the method using the algorithm proposed in the paper [7], which combines the flexibility of the neural algorithm of artistic style with the speed of fast style transfer networks to allow real-time stylization using any content/style image pair. This specific method was used for its real-time capabilities to create styles quickly between content and style image pairing.

3.1.3 Blending the Stylized Image

For blending the stylised image, we exploited the saliency map, M, data. We wanted to have the stylized segmented component, S, blend with the content image, C. This required us to smooth out the edges of S when using it as a mask for C. The saliency map, M, values ranged from 0-1. When the saliency value was 1, we directly used the corresponding value from S on C. If it was between 0 and 1, we used the formula proposed in the paper [8], which is shown below.

$$I[x,y]=M[x,y]*S[x,y]+(1-M[x,y])*C[x,y]$$

3.1.4 Pose transfer

We used the pose transfer algorithm from DiOr(Dressing in Order)from paper [8] on the different body types images obtained by clothes segmentation and wrapping. In DiOr , the person is represented as {pose, body, {garments}} tuple. Pose “P” is represented with 18 key points heatmaps as defined in OpenPose. For body representation, given a source image, its segmentation map is detected by SCHP then the body feature map (Tbody) is encoded by the body encoder (Ebody), which takes only skin segments from the source image. For garment encoding, a texture encoder (Etex) is used to get its feature map (Tgk) and then runs a segmenter on the feature map to obtain a soft shape mask (Mgk). GFLA is used to transform features and masks from

the source pose of a person or garment image to the target pose. For the generation pipeline, the desired pose “P” is encoded by the pose encoder (Epose), which outputs a hidden pose map (Zpose). Then a hidden body map (Zbody) is generated using a conditional generation block (Gbody) using a hidden pose map (Zpose) and body texture map (Tbody). Then for any kth garment, the garment generator takes the garments feature map and soft shape mask together with the previous state and produces the next state. The output image is then generated from the final hidden feature map using a decoder.

3.2. Novelty

While there are several algorithms for pose transfer andcloth segmentation, we couldn't find any for different body types. Through this project, we aim towards inclusivity by accommodating different body types. Also, we couldn't find any single algorithm to perform all the tasks of cloth segmentation, styling, and pose transfer together which will be accomplished through our project.

3.3. Database and Code

The first styling part , we use the U2 net model that was trained for a total of 70000 epochs with a batch size of 9 on two different datasets - HKU-IS and DUTS-TR. For the pose transfer, we used the DIOR model that is trained on deepfashion dataset. Total 101,966 training pairs were used and 8,570 testing pairs were used.

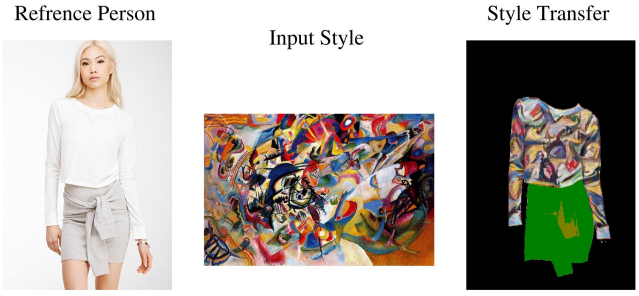
The code has been uploaded on GitHub.

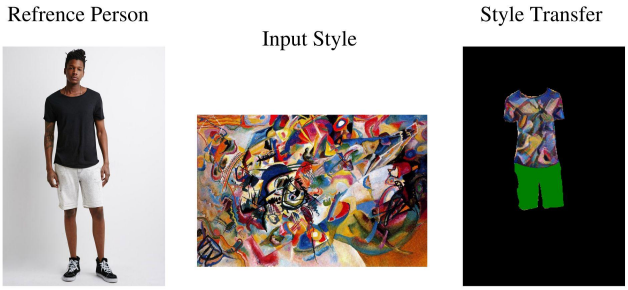
3.4. Output

This project consists of three steps. The first step is to create a saliency map generation and do cloth segmentation on the input reference person. The second step is to transfer the style to the input person. The last step is to blend the stylized image.

The output for the saliency map generation and cloth segmentation step is shown below.

The output for style transfer is shown below.

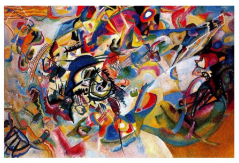




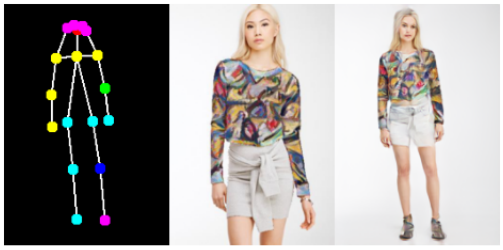
The output for blending the stylized image is shown below.

Refrence Person

Input Style

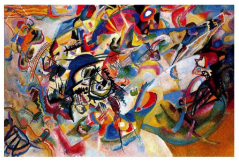


The output of Stylized Cloth

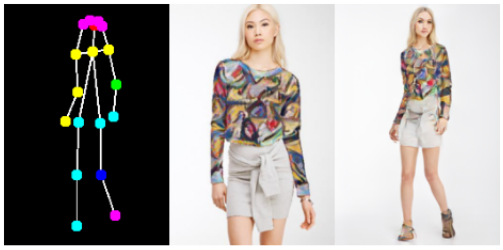


Refrence Person

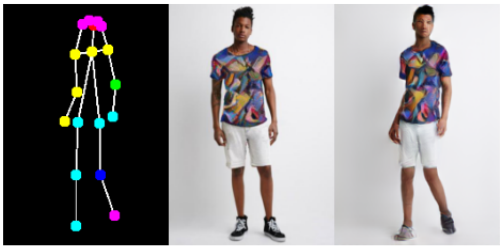
Input Style



The output of Stylized Cloth



The output for pose transfer gives us 40 images, all having different poses. Some of the output poses from the 40 generated images are shown below.



3.5. Result Analysis and Evaluation

We can compare the initial image with the final output to see which features change through the various stages, particularly how the initial image remains intact after the final output. We can evaluate the models based on the clarity of pictures formed primarily for regions where a portion of the image is replaced, like hair and smoothness for edges. Also it is viable to check the effectiveness for different poses to discover that the accuracy decreases as the complexity of the poses increases, though the results remain satisfying for almost all the forty poses tried.

4. Potential Contributions

We have successfully built a technique wherein an input image of a person and an input style, when taken, will generate an output of the human wearing the same piece of clothing with the changed style. Then we further, perform pose transfer on the stylized image to get output with 40 different poses with the stylized garment. This process works well for different body types. All the group members contributed equally and worked collaboratively to achieve this project milestone successfully.

References

[1] Liu, X., Liu, Z., Zhou, X., & Chen, M. (2019). Saliency-guided image style transfer. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 66-71). IEEE.

[2] Wang, H., Xiong, H., & Cai, Y. (2020). Image localized style transfer to design clothes based on CNN and interactive segmentation. Computational Intelligence and Neuroscience, 2020.

[3] D. Y. Park and K. H. Lee, "Arbitrary Style Transfer With Style-Attentional Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5873-5881, doi: 10.1109/CVPR.2019.00603.

[4] D. Chen, L. Yuan, J. Liao, N. Yu and G. Hua, "StyleBank: An Explicit Representation for Neural Image Style Transfer," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2770-2779, doi: 10.1109/CVPR.2017.296.

[5] George A., and Mark S. Nixon. "Real-time semantic clothing segmentation." In *Advances in Visual Computing: 8th International Symposium, ISVC 2012, Rethymnon, Crete, Greece, July 16-18, 2012, Revised Selected Papers, Part I 8*, pp. 272-281. Springer Berlin Heidelberg, 2012

[6] Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V. and Shlens, J., 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830.

[7] Liu, X., Liu, Z., Zhou, X., & Chen, M. (2019). Saliency-guided image style transfer. In 2019 IEEE International Conference on Multimedia & Expo Work- shops (ICMEW) (pp. 66-71). IEEE.

[8] Cui, A., McKee, D., & Lazebnik, S. (2021). Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14638-14647).

[9] Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., & Luo, P. (2020). Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7850-7859).

[10] Cui, Aiyu, Daniel McKee, and Svetlana Lazebnik. "Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 14638-14647. 2021.

[11] Zhu, Zhen, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. "Progressive pose attention transfer for person image generation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2347-2356. 2019.

[12] Li, Yining, Chen Huang, and Chen Change Loy. "Dense intrinsic appearance flow for human pose transfer." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3693-3702. 2019.

[13] , George A., and Mark S. Nixon. "Real-time semantic clothing segmentation." In Advances in Visual Computing: 8th International Symposium, ISVC 2012, Rethymnon, Crete, Greece, July 16-18, 2012, Revised Selected Papers, Part I 8, pp. 272-281. Springer Berlin Heidelberg, 2012