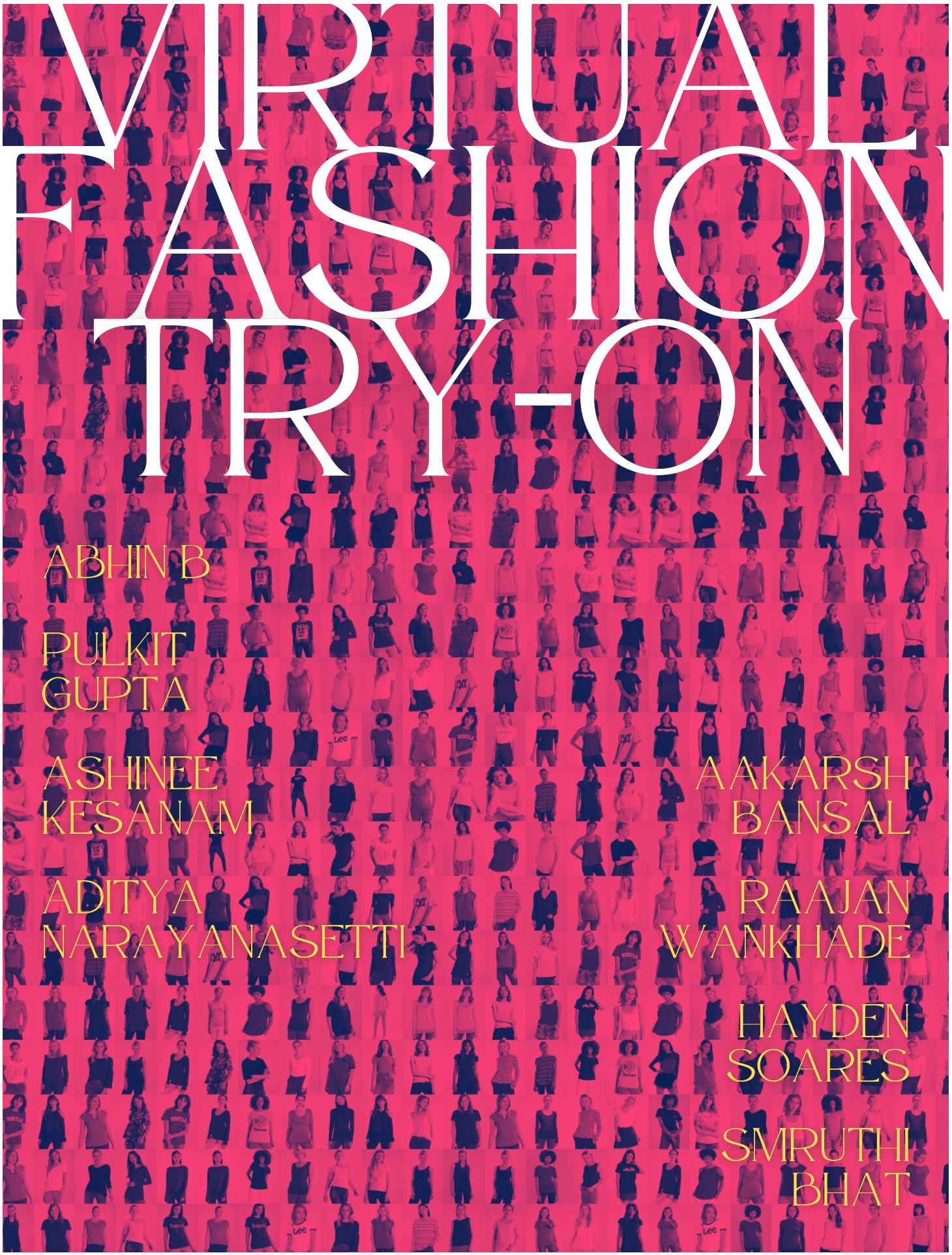


ABHINAB

PULKIT GUPTA

ASHINEE KESANAM

ADITYA NARAYANASETTI

AAKARSH BANSAL

RAAJAN WANKHADE

HAYDEN SOARES

SMRUTHI BHAT

Contents

Aim

The objective of this project is to develop a robust and user-friendly clothing style transfer system that can seamlessly transfer garments from an image to a target person. By leveraging recent advancements in computer vision, image processing, and deep learning techniques, our proposed system will enable users to experiment with different clothing styles, borrow garments virtually, and create personalized fashion statements. [1]

Introduction

The project makes use of several GAN architectures to develop a realistic Virtual Try On application. The complete model consists of 3 modules:

1. **Semantic Generation Module (SGM)**
   The semantic generation module (SGM) is proposed to separate the target clothing region as well as to preserve the body parts (i.e. arms) of the person, without changing the pose and the rest of the human body details.

2. **Clothes Warping Module (CWM)**
   Clothes Warping Module aims to fit the clothes into the shape of the target clothing region with visually natural deformation according to human pose as well as to retain the character of the clothes.

3. **Content Fusion Module (CFM)**
   Going beyond semantic alignment and character retention, it remains a great challenge to realize layout adaptation on visual try-on tasks. Content Fusion Module (CFM), integrates the information from previous modules to adaptively determine the generation or

preservation of the distinct human parts in the output synthesized image.

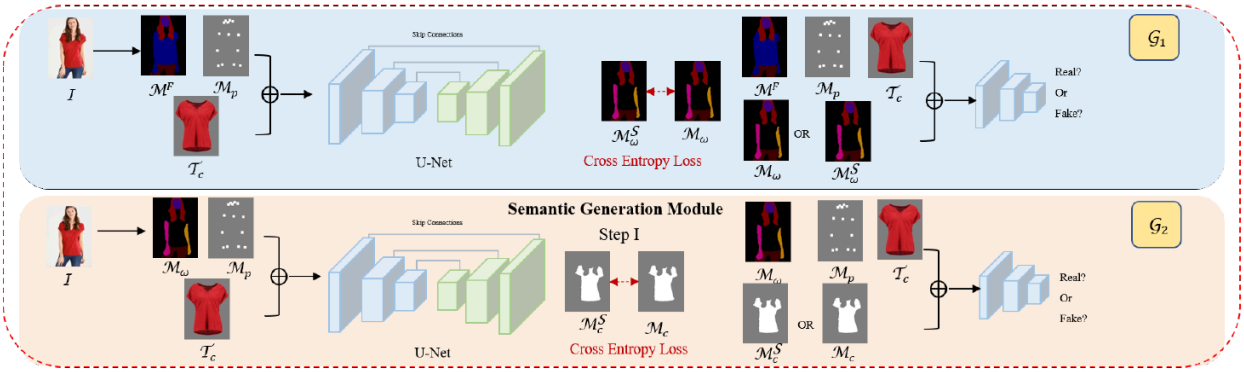
Methodology

1. Dataset

- The dataset consists of images of the person, images of the target cloth, segmentation of the person and posemaps of the person.
- All images are of dimensions 256 x 192.
- We converted the posemaps into a 4 channel input, a channel each for the points corresponding to the head, shoulders, arms and lower body. Each point was represented as a Gaussian distribution of a certain radius which was determined by the general sparsity of points in that particular channel.

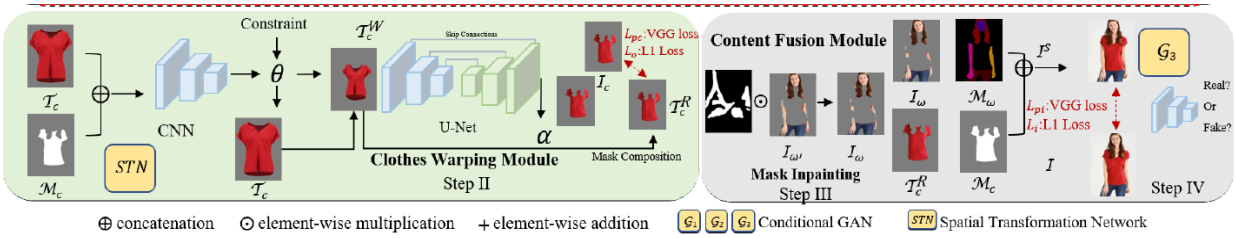
2. Architectures and Training Details

- The generator models for G1 and G2 are basically U-Nets [2] which take inputs and produce outputs corresponding to their specific task. The U-Net consists of a ResNet-50 [3] based encoder/backbone and a custom defined symmetrical decoder.

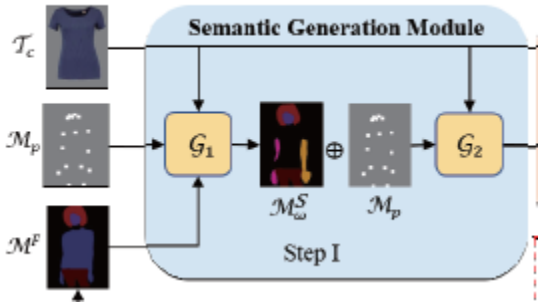


- For G3, instead of using transpose convolution layers we instead used bilinear upsampling and then used down-convolutional layers to achieve the same shapes; this was done to remove checkerboard artifacts that emerged during training. [4]

- The discriminators are again based on the ResNet-50 model and include some custom layers to match the image shapes.
- In G3, we used a critic instead of a discriminator, removed the sigmoid at the end of the discriminator to allow for a wider range of output values.
- We have used a batch size of 8 throughout the training processes.
- The loss function used for G1 and G2 is a weighted sum of binary cross entropy and pixel-wise cross entropy.
- For G3 we used a combination of the WGAN loss function [5] and perceptual loss [6], this was because using the loss function mentioned in the paper led to mode collapse and generation of images belonging only to a particular color.
- Adam optimizer has been used for G1 and G2, while RMSProp was used for G3.



3. Semantic Generation Module (SGM)



- The Semantic Generation Module consists of 2 GANs, G1 and G2. we train a body parsing GAN G1 to generate M_Sw (synthesized body part

mask) by leveraging the information from the fused map MF, the pose map Mp, and the target clothing image Tc.

- Using the generated information of body parts, its corresponding pose map and target clothing image, it is tractable to get the estimated clothing region. In the second stage, M_Sw, Mp and Tc are combined to generate the synthesized mask of the clothes M_Sc by G2.

4. Clothes Warping Module (CWM)

- Clothes Warping Module aims to fit the clothes into the shape of the target clothing region with visually natural deformation according to human pose as well as to retain the character of the clothes.
- This uses STN (Spatial Transformer Network) [7] to transform the cloth image to the required orientation for warping.
- This still does not encompass the actual warping output, as it is just a transformation on the cloth, hence a U-Net architecture is used to get composition masks to be multiplied with STN output to get the final warping image.

5. Content Fusion Module (CFM)

- Going beyond semantic alignment and character retention, it remains a great challenge to realize layout adaptation on visual try-on tasks.
- Content Fusion Module (CFM), integrates the information from previous modules to adaptively determine the generation or preservation of the distinct human parts in the output synthesized image.
- The training process uses mask inpainting to force the model into learning the reconstruction of non-targeted details. The reconstruction loss is a sum of:
    i. a modified version of perceptual loss from pretrained VGG16 model, and

ii. the WGAN loss.

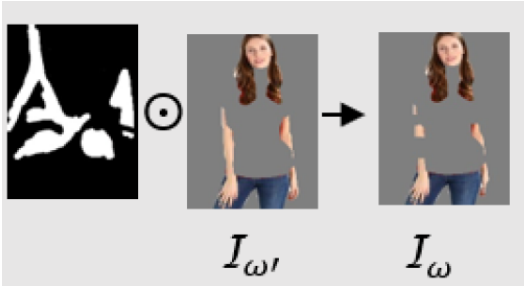


Fig. Mask Inpainting with a Random Mask

Results

We have achieved appreciable results using our models, here are some examples:

 +  --> 

Conclusion

The proposed system uses multiple Generative Adversarial Network (GAN) architectures to achieve the development of the Virtual Try On application. It uses three modules: Semantic Generation Module (SGM), Clothes Warping Module (CWM), and Content Fusion Module (CFM). SGM separates the target clothing region while preserving the user's body, CWM fits the clothing onto the user considering pose and style, and CFM refines the output for a natural look.

Limitations:

1. Computational Cost: Training GANs requires significant computational resources. Optimizing the training process for efficiency would be beneficial. Our training process involved several optimisations and workarounds for lack of resources.
2. Limited dataset: The current dataset only contains images of female models visibly in their youth. Including more images of people of varying ages and genders and body types would improve the generalisability of the application.

Future work:

1. Improved pose estimation: The current system uses 4 channels for its pose estimation model. Including finer details like elbows and wrists could enhance results.
2. Leveraging improved, more sophisticated models involving novel techniques to make even more realistic images.

Applications:

1. E-Commerce: Customers can virtually try on clothes, leading to more informed purchases and reduced return of goods.
2. Fashion design: Designers can visualize their creations on different body types, streamlining the design process.
3. Entertainment: VTO can be integrated into social media platforms or gaming applications for users to experiment with different looks.

References

1. Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content | IEEE Conference Publication
2. U-Net: Convolutional Networks for Biomedical Image Segmentation | SpringerLink
3. [1512.03385] Deep Residual Learning for Image Recognition
4. Deconvolution and Checkerboard Artifacts
5. [1701.07875] Wasserstein GAN
6. Perceptual Losses for Real-Time Style Transfer and Super-Resolution | SpringerLink
7. Spatial Transformer Networks