

Virtual Try-On: Project Milestone Report

Francesco Martucci, Felicia Puzone, Francesco Sala
{243215, 312722, 253589} @studenti.unimore.it

Abstract

Our project will be focusing on designing a 2D image-based virtual try on system. Virtual try on consists in generating an image of a reference person wearing a given try-on garment. This kind of problem has been widely investigated due to its relevance in the fashion market. VTON is also interesting as it is a challenging problem requiring a multi-layered approach incorporating at least both a geometric transformation module to warp the selected garment and a generative try-on module that creates the realistic try-on images.

The project also includes a content-based retrieval system, that extracts some descriptors from the garment picture and finds similar items from a repository.

The objective of the project will be for the system to more easily adapt to less professional and noisier photos.

Contents

1	Introduction	1
1.1	Problem statement	1
1.2	Related works	1
1.3	General approach	2
2	Technical Approach	2
2.1	Pre-processing	2
2.2	Person representation	2
2.3	Warping module	3
2.4	Generative module	3
2.5	Garment Retrieval	4
3	Intermediate/Preliminary Results	4
3.1	Pre-processing	4
3.2	Conclusion	5

1 Introduction

1.1 Problem statement

Our project will be focusing on designing a 2D image-based virtual try on system. Virtual try-on consists in generating an image of a reference person wearing a given try-on garment.

1.2 Related works

In order to solve this problem we have analyzed already existing literature and their approaches to the same problem.

The CP-VTON paper^[4] identifies four main requirements that need to be accomplished by a virtual try-on system:

- warping the garment according to the body shape and pose of the target person;
- transferring the texture of the garment on the target person without losing important details;
- merging the image of the target person with the warped result in a plausible way; and
- render light and shades of the final image correctly, to ensure realism.

One of the most important advancements made by the CP-VTON architecture is the introduction of a warping module that computes a learnable Thin-Plate Spline transformation that warps the garment in a reliable way.

We obtained access to the Dress-Code dataset^[2], collected by AImageLab, which, given its size, may potentially boost the efficacy of our system.

We also noticed that, in recent years, that there have been a lot of improvements to the quality of the generated images through the introduction of transformer-based modules. This approach is followed by the paper **Dual-Branch Collaborative Transformer for Virtual Try-On**^[1] to solve the virtual try-on problem with great results, as such we will also attempt to implement a similar transformer-based architecture.

1.3 General approach

Our system will be subdivided into different modules each one designed to solve specific step of the process.

- pre-processing: this module handles all which regards the image enhancement (denoising, light adjustment, etc...), performs the background removal task;
- person representation: this module performs pose estimation and the semantic segmentation of the person into their body parts;
- warping module: this module implements a geometric transformation that envelops the body shape of the subject with the fabric of the clothing item;
- try-on: as the last module in the pipeline, this part generates a new image by composing the warped garment over the subject and should ensure the satisfaction of the requirements stated in the above section.

2 Technical Approach

In order to get across the inner workings of each module we will separate this section into subsections related to each one.

2.1 Pre-processing

As the objective of the system is to be adaptable to dirty and noisy input images, in the pre-processing phase great care should be taken to clean such inputs. As such the pre-processing module performs different methods of input refinement.

A first denoising pass is performed utilizing the bilateral filter, after which the image goes through a light adjustment procedure which entails contrast stretching.

After these operations we perform the background removal. This is because we thought that a potential point of improvement within the virtual try-on pipeline could be the removal of the background before going deeper into the network as the background information may hinder the performance of

the subsequent modules. To perform such a task we are looking and comparing multiple already existing solutions such as:

- U-Net^[3] fine-tuning a pre-trained U-Net model using the Tik-Tok dataset;
- Detectron2^[5]: fine-tuning on Dress-Code the Detectron2 segmentation module released by Facebook.

We will compare the results and choose the best one according to state-of-the-art evaluation metrics (DICE etc...).

2.2 Person representation

This module deals with the semantic segmentation of the subject and their keypoints extraction, which will be used to as features for the warping module.

Regarding the segmentation problem we have, as of now, identified the Detectron2 segmentation module as the main candidate to solve it.

Instead, for the keypoint extraction problem we are comparing the DensePose module (from Detectron2) and the OpenPose module.

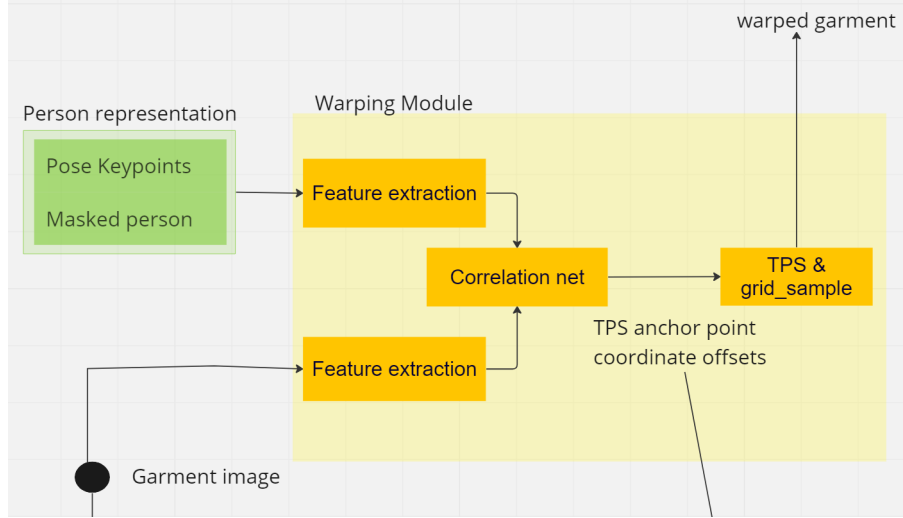


Figure 1: Warping module schema

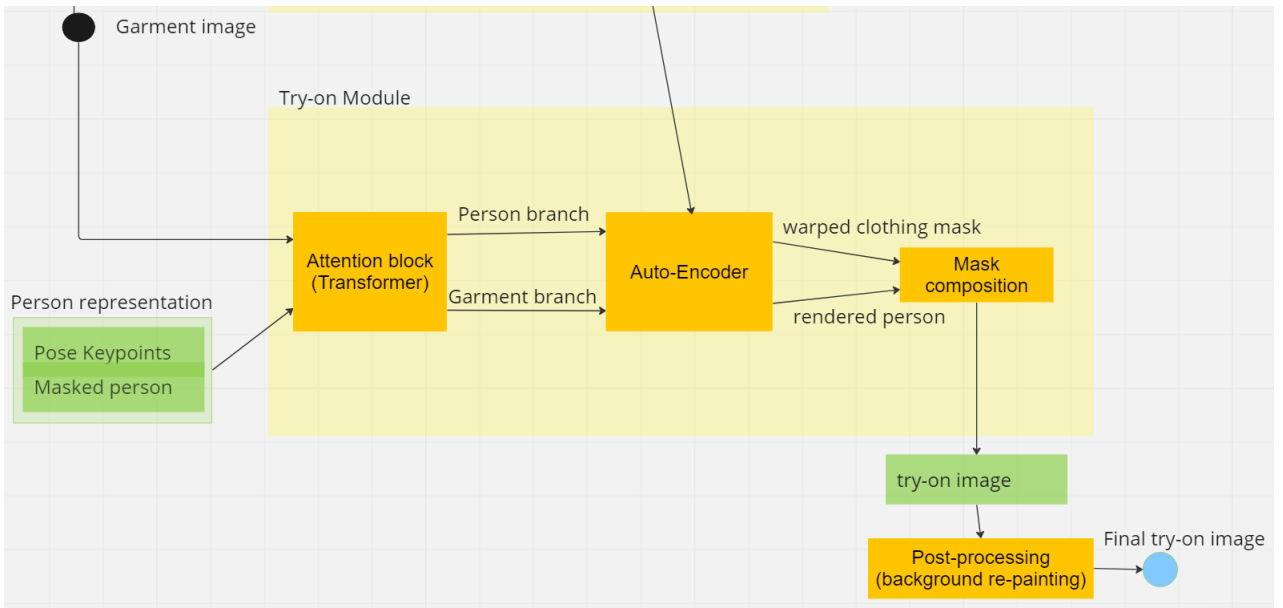


Figure 2: Generative module schema

2.3 Warping module

As depicted in the figure 1, the warping module takes as inputs the person representation feature vector and the garment image and, based on the keypoints and the segmentation mask, it warps the garment fitting it to the body shape of the subject.

The output is the set of the TPS parameters, which are used to geometrically transform the clothing item.

2.4 Generative module

As depicted in the figure 2, the generative module is the one which will be producing the

final image by merging the warped garment with the human picture.

The module will consist of a transformer-based feature extraction unit which will feed

into an autoencoder unit which will generate the output image.

2.5 Garment Retrieval

3 Intermediate/Preliminary Results

3.1 Pre-processing

Our aim is to use pictures in non-optimal lighting and perspective conditions, so due to the fact that our training dataset (Dress-Code) is mainly composed of high quality professionally taken images, we are trying to imitate this condition and apply some enhancement to the generic input picture. In particular we chose to use:

1. contrast stretching: it is a simple image enhancement technique that attempts to improve the contrast in an image by stretching the range of intensity values it contains to span a desired range of values. Table 1 depicts a sample.

2. Bilateral filtering: we attempt to remove the noise while preserving the sharp-

ness of the edges. Table 1 depicts a sample.

3. Background removal: as our dataset training images usually possess a plain and homogeneous background, in order to decrease variance between training and real-world images, and to decrease the computational training effort, we are comparing semantic background removal using different kind of segmentation techniques. Since this procedure is applied to real-world images, we trained a network over a dataset named Tik-Tok dancing segmentation that we believed to have a more natural picture acquisition of body shapes. To perform the semantic segmentation we used U-Net network and, as of now, we have trained over 10 epochs as a trial. Table 3 depicts some results.

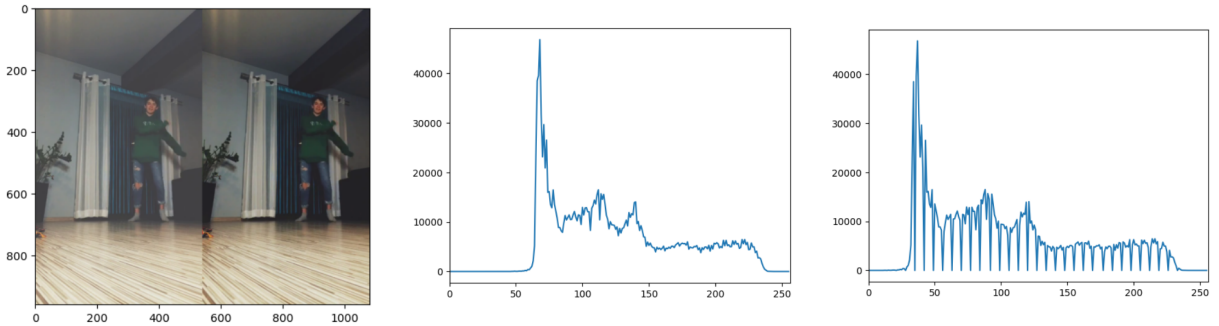


Table 1: Sample from Tik-Tok Segmentation Dataset pre-processed with contrast stretching.

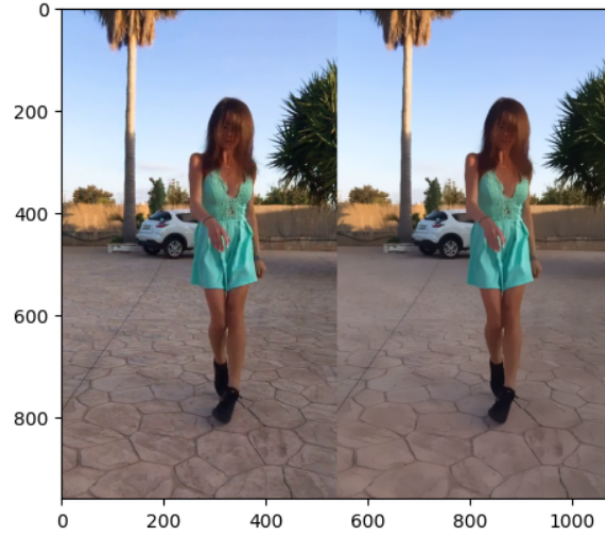


Table 2: Sample from Tik-Tok Segmentation Dataset pre-processed with bilateral filter.

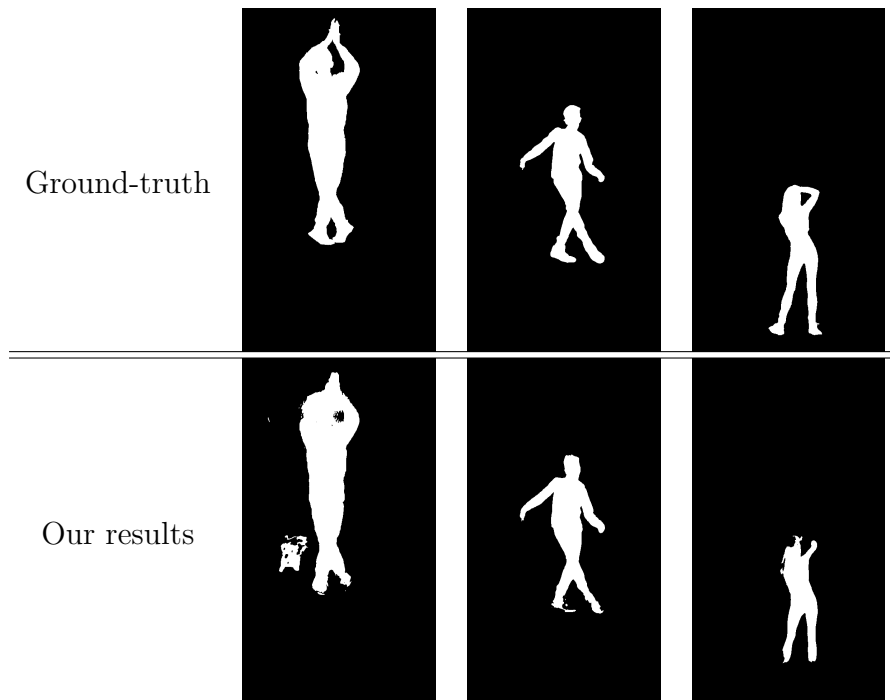


Table 3: Samples of segmentation maps from Tik-Tok Segmentation Dataset.

3.2 Conclusion

Cosa succederà nelle prossime puntate ?...?

References

- [1] Emanuele Fenocchi, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Fabio Cesari, and Rita Cucchiara. Dual-branch collaborative transformer for virtual try-on. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2246–2250, 2022.

- [2] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision*, 2022.
- [3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
- [4] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.
- [5] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.