# Virtual Try-On: Final Project Report

Francesco Martucci, Felicia Puzone, Francesco Sala
{*243215, 312722, 253589*} @studenti.unimore.it Github project repository:
https://github.com/felichan98/vton-cv-project

**Abstract**
*Questo abstract non mi convince del tutto, bisogna riscriverlo.* 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: given a worn cloth, the system extracts descriptors from the picture and finds similar items from a repository.

*Quest'ultima frase è da rivedere, alla luce di quello che ci disse il tutor.* The objective of the project will be for the system to more easily adapt to not as professional and noisier photos.

*N. parole: circa 150, quindi siamo ok.*

# Contents



# 1 Introduction

## 1.1 Problem statement

During the course of our project, we designed a 2D image-based virtual-try-on system (VTON).

VTON consists in generating an image of a target person wearing a given try-on garment. This kind of problem has been widely investigated because of its economic importance in the fashion industry and its level of complexity.

The main challenges and requirements[9] needed to accomplish this task are:

- 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;
- render light and shades of the final image correctly, to ensure realism.

It is evident that a trivial and simple approach doesn't work. Instead, it is required a multi-layered and deep-learning approach.

Due to the exam constraints, the project also includes a content-based retrieval system: given a worn cloth, the system finds similar items from a repository. Also this kind of problem is challenging:

- clothes can have very different shape, colors and intricate decorative patterns;
- the worn cloth images picture may be taken in uncotrolled setting, while in-shop clothing item pictures are taken in a clear and clean setting;
- if the repository contains a very large number of garments, it is necessary an efficient management of the data and efficient retrieval algorithm;
- relevance in the fashion market: garment retrieval can be a very useful e-commerce tool to enchance the customer experience.

The main contribution of our work can be summarized as an improvement of the CP-VTON+ architecture[5] using a trasformer-based generative module.

Questo paragrafo è da sistemare

## 1.2 Related works

The baseline work is the CP-VTON architecture[9]. Its main contribution was the introduction of a two-stage pipeline:

1. Warping Module: it computes a learnable Thin-Plate Spline transformation (TPS) for warping the in-shop garment in a reliable way;
2. Generative Module: it fits the warped garment on the target person.

The warping module allows the retaining of the important details of the garment, but it fails if the target person pose or the garment texture are too complex, or there are occlusions. Several works tried to improve the warping module and overcome such limitations by: integrating complementaryv modules; applying regularization techniques to stabilize the warping process during training; projection techniques of the garment details.

For the generative module, the classical approach was the U-Net architecture (feeding the person image and the warped cloth). Other works have employed a two-branch network, where one branch takes as input the person and the other the in-shop cloth and warping information. A relatively new approach apply the Transfomer-based architecture and cross-modal attention mechanisms to the inputs

Non so se aggiungere tutte le references dirette. Non so se va bene così o è necessario scen-

before feeding them to the generative network [7] [2]. Lastly, some approaches tried to estimate the person semantic layout to improve the visual quality of the generated images.

Another line of research is trying to construct better and public datasets [6]:

- increase the total number of samples;
- increase the image resolution: at now, the mostly used resolution is $256 \times 192$; although the processing is lighter, such images do not retain many cloth details;
- unpaired-images

The newest line of research is an end-to-end deep-learning pipeline, for example. Anyway, such line is outside the scope of this project.

Regarding the cloth retrieval task, pioneer works utilized a fixed set of attributes (color, length, material, etc...) hand-labeled or automatically extracted (such as SIFT/ORB keypoints). Then the system compares the query-item and shop-item according to a similarity metric. The overall performance was not satisfying, since only perceptual method aren't able to capture the higher-level depencences between clothes, that have intricate pattern and not elementary shapes. In the last years, deep neural networks have been widely applied and have pushed the research into a new phase. These new methods learns a similarity metric between real-world and shop-item images from deep features representations extracted form images; an interesting example is Exact-Street-to-Shop [3]. Like shown by [10], focusing on clothes regions and ignoring the background via a cloth parsing mechanism improves the overall performance.

aggiustare, non sono ancora del tutto soddisfatto

articolo Cucchiara su Stable-Diffusion

dovrei inserire qualche citazione?

# 2 Dataset

The dataset used is Dress Code [6], introduced recently by the AImageLab of UNIMORE. Its main characteristics are:

- pubicly available;
- high-resolution images ($1024 \times 768$);
- very large dataset with respect of the publicly available ones, with about $50k$ image pairs of try-on garments and corresponding catalog images where each item is worn by a model;
- multi-category clothes: front-view and full-body of upper-body, lower-body, and full-body;
- rich annotations: for each cloth there are the dense-pose map, cloth worn by a model, keypoints of the model, label map representing the body-parts segmentation and the body skeleton.

# 3 Proposed Architecture

## 3.1 General approach

As depicted in the figure 1, our system will be subdivided into different modules each one designed to solve a specific step of the process:

- Pre-processing: this module handles all which regards the image enhancement (denoising, light adjustement, etc...) and 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 warps the fabric of the clothing item depending of the body shape and pose of the subject;
- Try-On: this module generates a new image by composing the warped garment over the subject and should ensure the satisfaction of the requirements stated in the introduction section;
- Image Retrieval: this module perform a content-based retrieval of the in-shop clothes, given a reference image.

Considering the e-commerce use case, during the inference, the customer inputs the target person image and the desired cloth image. The system selects the most similar in-shop cloth and then performs the virtual-try-on using these selected clothes.

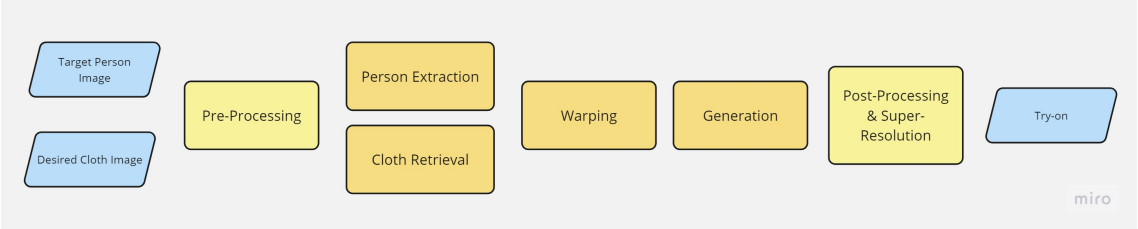


Figure 1: Overview of the pipeline

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

## 3.2 Pre-Processing

The system expects to receive two images: target person, desired cloth (it can be worn by a person or it can be cloth-only image). As the objective of the system is to be adaptable to dirty and noisy input images (so called "in-the-wild"), great care should be taken to clean such inputs.

For both of them, a first denoising pass is performed utilizing the bilateral filter, after which the images go through a light adjustment procedure which entails contrast stretching. After that, different operations are performed for each input.

Only for the target person, we perform the background removal, since the background may contain perceptual noise (in addition to the target person, there may be other people, objects etc...). This task is performed using BlaBlaNet.

## 3.3 Person Representation

This module performs pose estimation and the semantic segmentation of the person into their body parts. It is foundamental that the person representation is cloth-agnostic: the try-on task has to preserve the target person's information (face, hair, body shape and pose), while ignoring the worn cloth. We followed the CP-VTON[9] approach, so the person representation contains three components:

- Pose Heatmap: the pose estimation contains information about the person pose. It follows the keypoints representation and it has been performed using OpenPose[1]. From the keypoints list, it is computed an 18-channel feature map with each channel corresponding to one human pose keypoint, drawn as an $11 \times 11$ white rectangle;
- Body Shape: a 1-channel feature map of a blurred binary mask roughly covering different parts of human body;
- Reserved Regions: an RGB image that contains the reserved regions that are not involved in the try-on phase (for instance hands, feet, face, hair), detected using SCHP[4] that generates the segmentation mask representing the human parsing of model body parts and clothing items.

These feature maps are all scaled to a fixed resolution and concatenated together to form the cloth-agnostic person representation map $p$ of $k$ channels, where $k = 18 + 1 + 3 = 22$. In CP-VTON, the resolution adopted is $256 \times 192$, instead we used $512 \times 384$.

This representation is utilized in both the warping and try-on module.

## 3.4 Cloth Retrieval

The goal of this module is to match a real-world example of a garment item to the same or similar item in an online shop. We take inspiration from Exact-Street-to-Shop approach that makes two interesing points:

- mixing up the geometric/perceptual and the deep-learning works pretty well;
- tackle the cloth-matching problem as a binary classification (match or not).

inserire immagine

As depicted in the figure.., we built a reference repository of in-shop cloth items, using the cloth-only images in DressCode. For each cloth item we extract a new feature representation following these steps:

- for each image channel (DressCode images are in the RGB format):
  - run ORB algorithm to compute the keypoints and the corresponding descriptors;
  - compute the histogram on the 256 different color levels;
  - concatenate all the values in a 1D tensor;
- concatenate all the channel tensors in a 1D tensor and save it.

The noise-clean desidered cloth image is fed to a sub-module that detects the cloth area (as a bounding box), using SCHP. From this area is extracted a new feature representation (query-feature) following the same steps presented for the cloth-only images. Then, the comparisons are performed:

- for each in-shop cloth:
  - concatenate the query-feature with the in-shop cloth feature in a 1D tensor;

  - fed the tensor to the deep neural network;

- sort and select the $k$-best matched in-shop clothes according to the matching-score given by the network.

As depicted in the figure..., the deep neural network we designed is composed by a sequence of Fully-Connected and Normalization layers, that outputs two numbers representing, after softmax normalization, the probability of no-matching and the probability of matching. The higher one is considered the final classification result.

To train this network, we used the regularized cross-entropy loss:

$$L(\theta) = \lambda_1 \cdot L_1(c_{warped}, I_{c_t}) + \lambda_{reg} \cdot L_{reg}. \tag{1}$$

During training, the network tried to classify both positive examples (the paired couples of worn-cloth and cloth taken) and negative examples (not paired couples). Since the number of potential negative examples was far more the number of positive examples, we used data augmentation to creating synthetic positive examples from existing one: we applyed gaussian noise to the feature representations.

inserire l'immagine della rete

sistemare la formula

in Experiments, accennare al dropout

## 3.5 Warping

We follow the warping module proposed in CP-VTON+ [5]. This module transforms the input try-on garment $c$ into a warped image of the same item that matches the body pose $p$ and shape $m$. As warping function we use a thin-plate spline (TPS) geometric transformation, which is commonly used in virtual try-on models *qua va la cit.33*. Inside this module, we aim to learn the correspondence between the inputs $(c, p, m)$ and the set of parameters $\theta$ to be used in the TPS transformation. Intuitively, the TPS fits a surface (try-on garmet) through points (feature regression output points) minimizing an energy function. Formally, it is a poli-harmonic spline function which maps a set of points $(x, y)$ on their correspondences $(x', y')$ sampled from input images. Under these terms, the warping module aims to learn how to perform such transformation.

As depicted in figure blabla, the module extracts an encoded representation of the try-on garment $c$ and an encoded representation of the person representation through two separate convolutional networks. Then, it is computed a correlation map between these representations. This correlation map is used to predict the spatial transformation parameters $\theta$ corresponding to the $(x, y)$-coordinate offsets of TPS anchor points.

To train this network, we used the following loss:

$$L(\theta) = \lambda_1 \cdot L_1(c_{warped}, I_{c_t}) + \lambda_{reg} \cdot L_{reg}; \tag{2}$$

$$L_{reg}(G_x, G_y) = \sum ii \sum ii \sum iii; \tag{3}$$

where $L_1$ indicates the pixel-wise $L_1$-loss between the warped result $c_{warped}$ and the ground truth $c_t$. $L_{reg}$ indicates the grid regularization loss.

### 3.6 Generation

This module aims at fitting the warped garment on the target person. Previous works like CP-VTON and CP-VTON+ directly concatenate the person image $p$, the warped clothing image $\hat{c}$, and the warped clothing mask image $\hat{cm}$ . Then the concatenated input is sent to a UNet[8] to generate a composition mask $M_o$ and a rendered person image $I_R$. The main limitation of this approach is the inability of the convolution operator to model global long-range dependences. For this reason, we follow the CIT[7] approach that leverage the Transfomer-based architecture and cross-modal attention mechanism that are able to capture the dependence among these three input images.

As shown in the figure 2, the three inputs goes through the patch embedding, for making the image data compatible. Then each goes through a 1D temporal convolution to ensure the relation modeling of each element with its neighbor elements. Then the Interactive-Transformer II is utilized for modeling the global long-range correlation. The output $X_{out-II}$ of Interactive Transformer II is obtained after a linear projection and a reshape operation. Then $X_{out-II}$ is utilized for two proposes; one is to activate the important region of the overall input by adding $X_{out-II}$ to $I(p, \hat{c}, \hat{cm})$; another is to guide the final mask composition as follows:

$$\begin{aligned} I_R^{global} &= sigmoid(X_{out-II}) \times I_R, \\ I_o &= M_o \times \hat{c} + (1 - M_o) \times I_R^{global} \end{aligned} \tag{4}$$

where $\times$ represents the element-wise multiplication and *sigmod* indicates the sigmoid activation function.

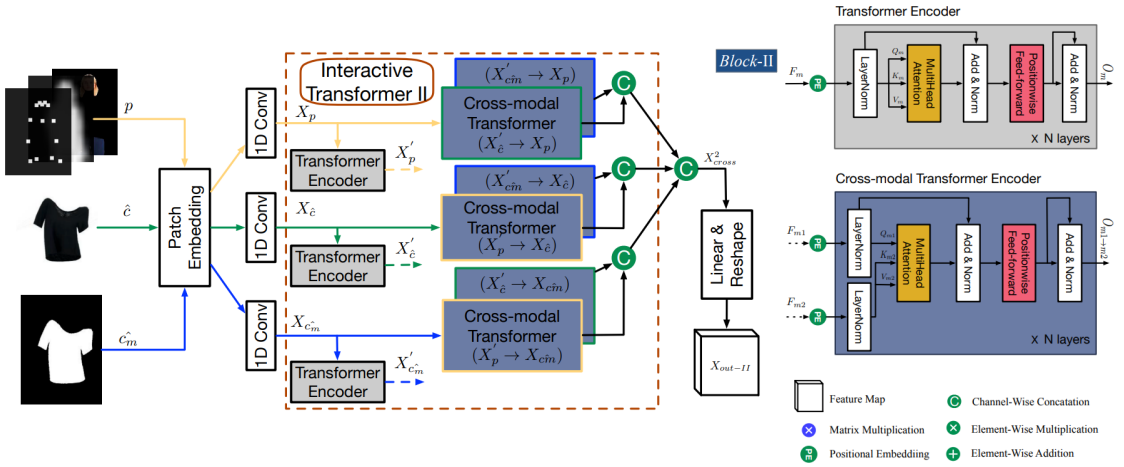


Figure 2: On the left: CIT Reasoning Block architecture, also called Interactive-Transformer II. On the right: the Transformer Encoder and Cross-Modal Transformer Encoder architetures and a legend of symbols.

### 3.7 Post-Processing and Super-Resolution

# References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] 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.

[3] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to buy it: Matching street clothing photos in online shops. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3343–3351, 2015.

[4] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing, 2019.

[5] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[6] 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.

[7] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *arXiv preprint arXiv:2104.05519*, 2021.

[8] 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.

[9] 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.

[10] Zhonghao Wang, Yujun Gu, Ya Zhang, Jun Zhou, and Xiao Gu. Clothing retrieval with visual attention model, 2017.