Computer Vision
Final Project Report

Image-based Virtual Try-on Network with human parsing

Team 3， 0751231 曾揚 , 309505018 郭俊廷

Introduction:

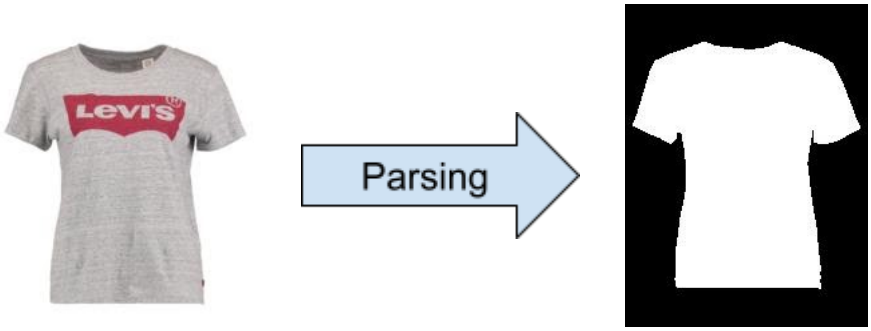
Recent years have witnessed the increasing demands of online shopping for fashion items. Image-based virtual try-on systems with the goal of transferring a desired clothing item onto the corresponding region of a person have made great strides recently, but challenges remain in generating realistic looking images that preserve both body and clothing details. Recently virtual try-on methods based solely on RGB images have also been proposed. These methods formulate virtual try-on as a conditional image generation problem, which are much less resource intensive and have the potential for widespread applications, if proven effective. However, slightly-wrong segmentation results would lead to unrealistic try-on images with large artifacts. The image quality is bounded by the traditional parser-based model. To solve this problem, we propose “teacher-tutor-student” knowledge distillation, which is able to produce highly photo-realistic images and possesses several appealing advantages compared to prior work.

Implementation procedure:

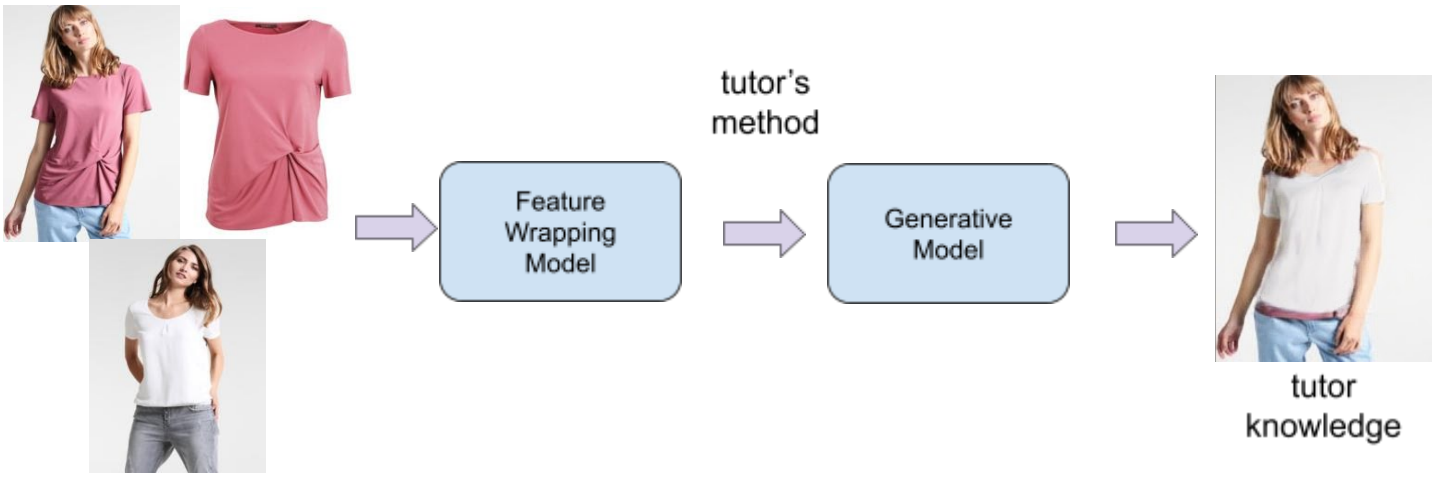
1. We use VITON (https://www.kaggle.com/rkuo2000/viton-dataset) for our datasets, and randomly split it into 14,221 image pairs for training data and 2,032 image pairs for testing data.
2. We have a Flow-based Feature Warping Model(FWM), which can find dense correspondences between human poses and clothes and a Generative Model(GM), which can concatenate the warped clothes and human pose. We will introduce details of the model architecture in the main algorithm part.

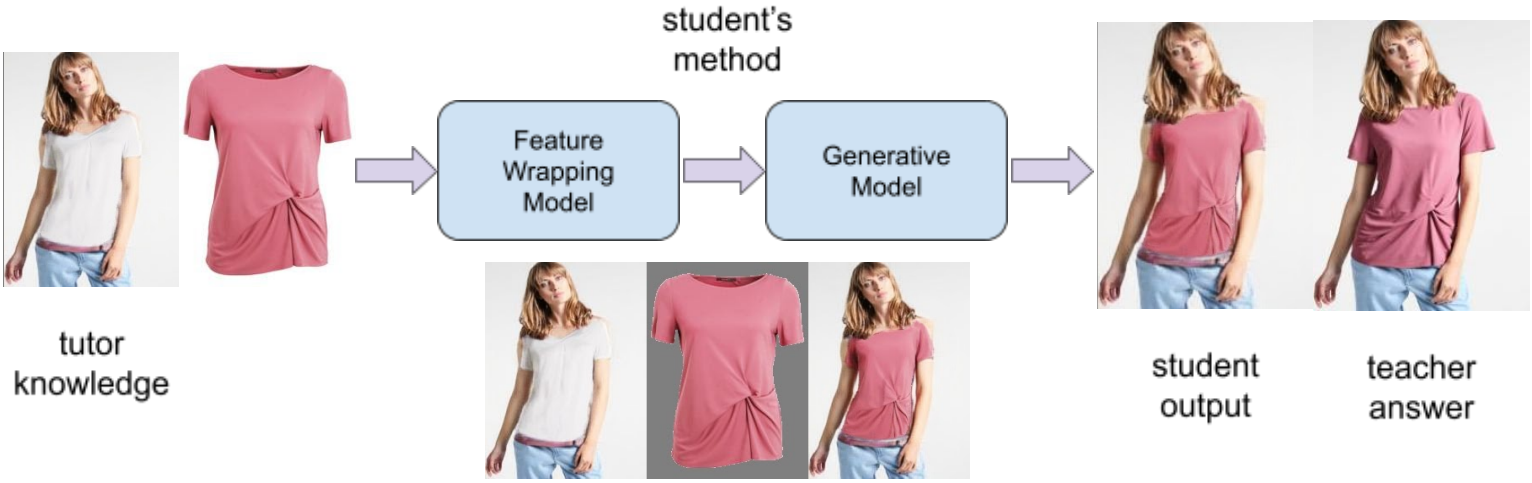
Training:

3. Suppose we have female models M_a, M_b with different human poses. Model M_a and M_b wear different fashion clothes I_a and I_b.
4. Then, we need to do clothes parsing for data preprocessing. Each image size of model and clothes is (256,193,3).



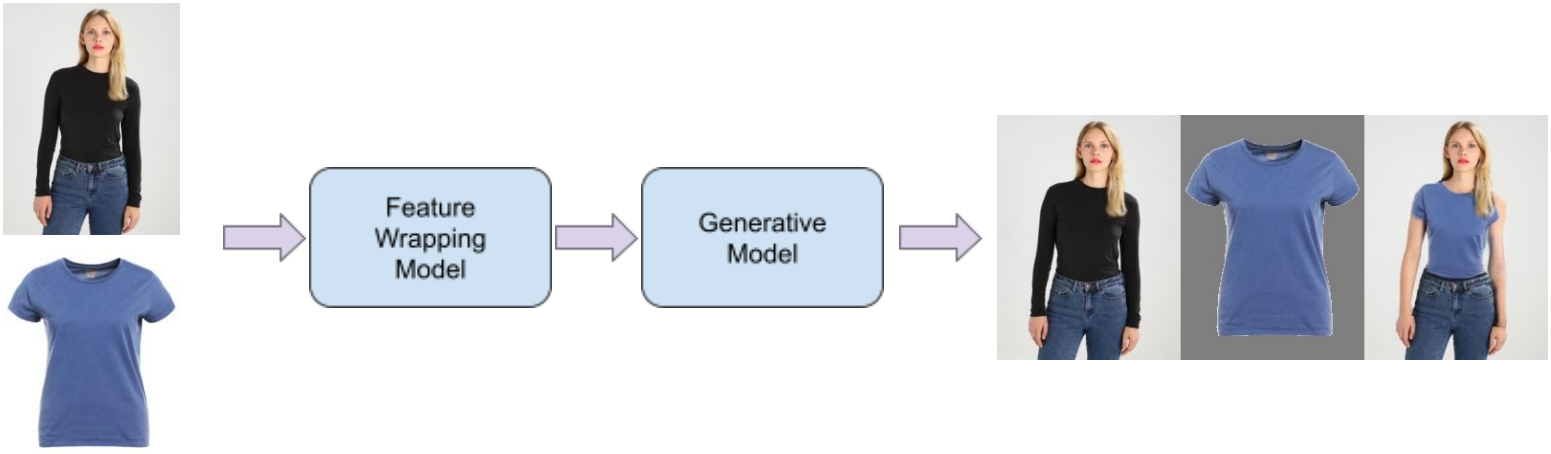
5. Input M_a, I_a, I_b to a FWM, and it can calculate the dense correspondences between M_a and I_a. Then, input these feature and human parsing M_a to a generative model, it will output model M_a wears fashion clothes I_b, we called it tutor knowledge.
6. Then, input tutor knowledge, which is M_a wears I_b, and a clothes I_a to FWM. Because we have the correct answer of M_a wears I_a as teacher knowledge. So, student can imitate to wear I_a on model M_a.





Testing:

7. For the testing part, we also need to do clothes parsing for data preprocessing. Each image size of model and clothes is (256,193,3).
8. Input model and clothes, and it will concatenate the warped clothes and human pose.



Main Algorithm:

1. Flow-based Feature Wrapping Model:
   The main purpose of Flow-based Feature Wrapping Network is to establish accurate dense correspondences between the person image and the clothes image. It can be divided into three parts, Feature Encoder, Refine Pyramid and AFlowNet.

- Feature Encoder
  Feature Encoder aims to extract features from human pose and clothes, by using encoder. First, we do downsampling to resize the image, and construct a pyramid. We set N=5, so we have 5 different sizes of images [64,128,256,256,256] and form a pyramid. Second, we use two ResBlock to extract two-branch pyramid features from 5 levels. ResBlock is the basic module that constitutes ResNet, which is a famous CNN model.

```
class ResBlock(nn.Module):
    def __init__(self, in_channels):
        super(ResBlock, self).__init__()
        self.block = nn.Sequential(
            nn.BatchNorm2d(in_channels),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels, in_channels, kernel_size=3, padding=1, bias=False),
            nn.BatchNorm2d(in_channels),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels, in_channels, kernel_size=3, padding=1, bias=False)
            )

    def forward(self, x):
        return self.block(x) + x
```

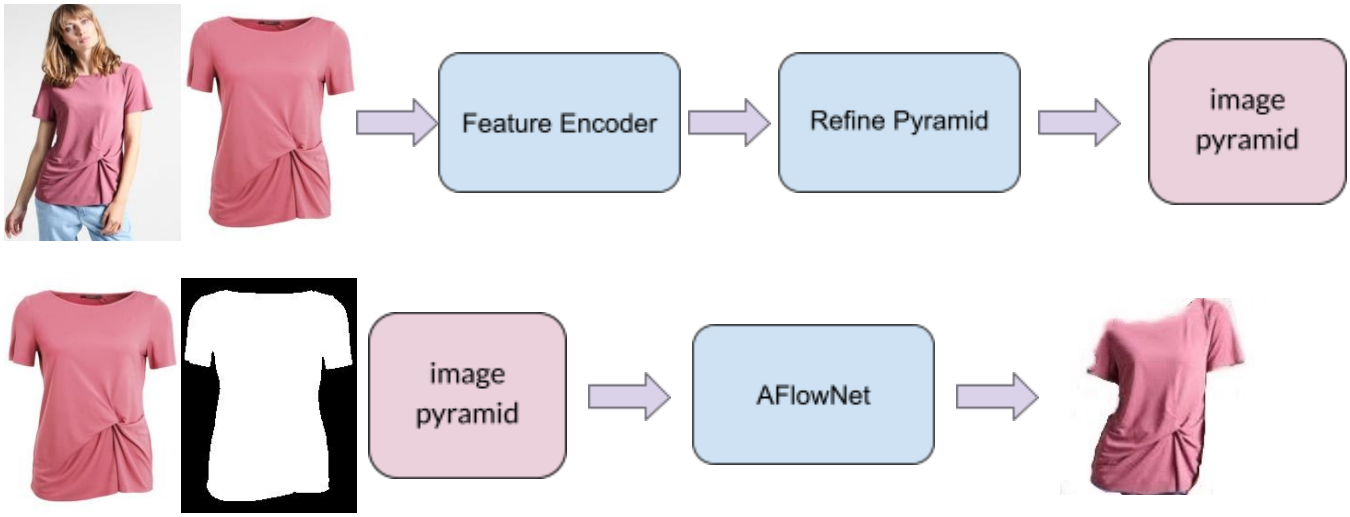
- Refine Pyramid
  Refine Pyramid aims to do “adaptive smoothing” methods. “Adaptive smoothing” is a common method in computer vision. The appearance flows between the person image and the clothes image need to be predicted accurately, or the minor mistakes should result in very unnatural warping results.

- AFlowNet
  AFlowNet aims to estimate the appearance flows from 5 levels’ pyramid features. It performs pixel-by-pixel matching of features to yield the coarse flow estimation with a subsequent refinement at each pyramid level. We input (image, image_edge, image_pyramids) to AFlowNet and output clothes images with human pose.

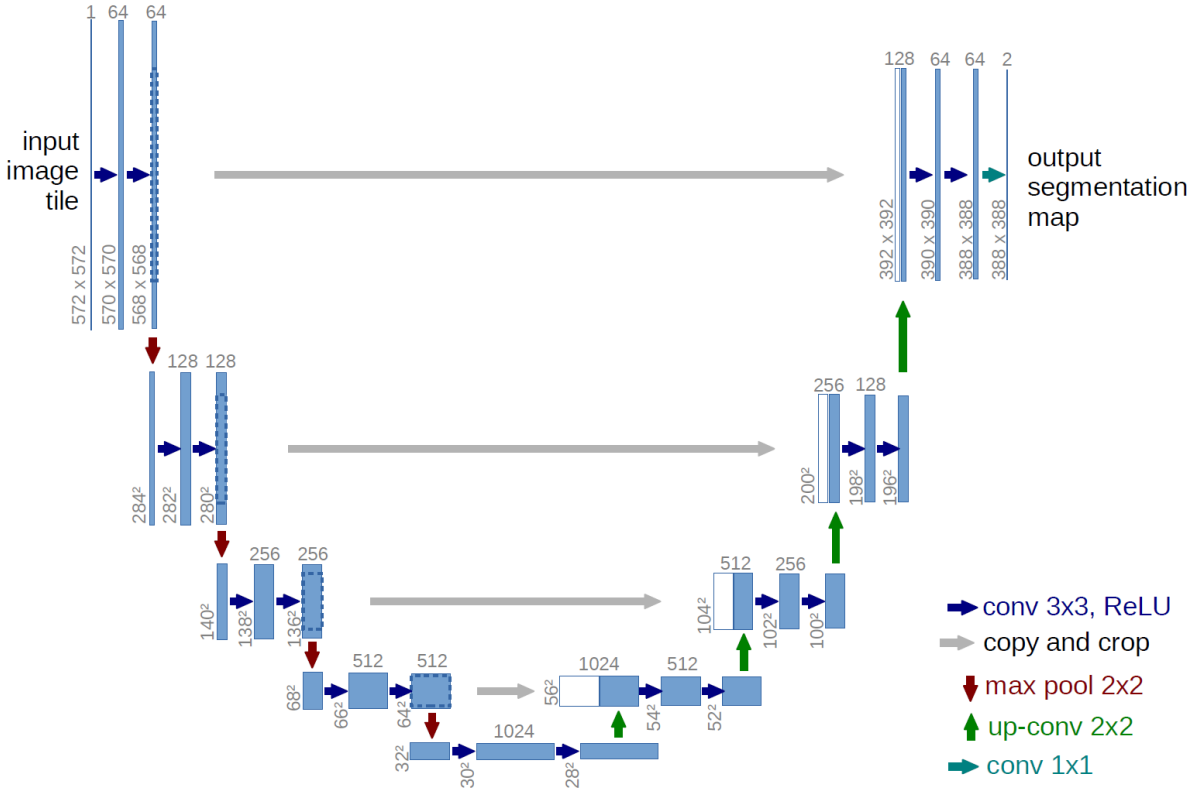
The appearance flows between the person image and the clothes image need to be predicted accurately, or the minor mistakes should result in very unnatural warping results. So, we need Loss function L_{sec}

$$L_{sec} = \sum_{i=1}^{N} \sum_{t} \sum_{\pi \in N_t} P(f_i^{t-\pi} + f_i^{t+\pi} - 2f_i^t)$$

where f_i^t denotes the t-th point on the flow maps of the i-th scale. N_t indicates the set of horizontal, vertical, and both diagonal neighborhoods around the t-th point. The P is the generalized charbonnier loss function.



2. Generative Model
   The main purpose of Generative Model is to synthesize the try-on image. The basic module of generative model is Res-UNet, which is built upon a UNet (https://arxiv.org/pdf/1505.04597.pdf) architecture, in combination with residual connections, which can preserve the details of the warped clothes and generate realistic try-on results, which can preserve the details of the warped clothes and generate realistic try-on results. Each Res-UNet contains five ResUnetSkipConnectionBlock, and each ResUnetSkipConnectionBlock contains several ResidualBlocks.



- Loss Function
  We have two parts of the loss function here.

$$L_l = ||s_I - I||_1$$

L_l is the pixel-wise L1 loss, where s_I is the student answer and I is the real image representing the teacher's answer.

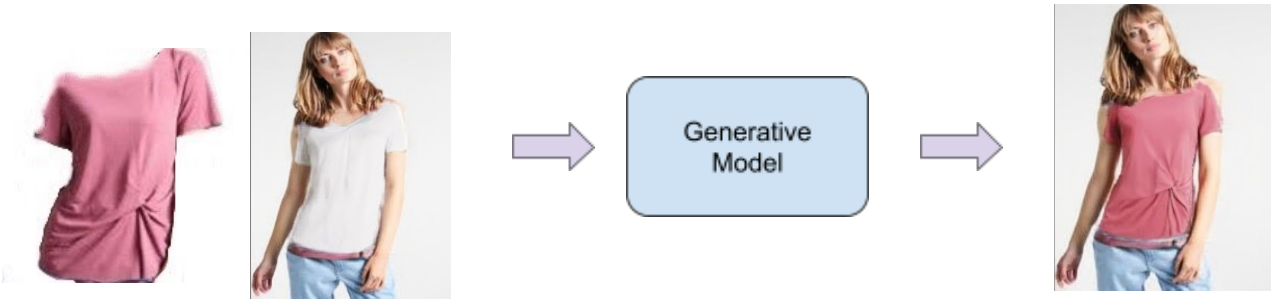
$$L_p = \sum_m ||\phi_m(s_I) - \phi_m(I)||_1$$

L_p is the perceptual loss to encourage the visual similarity between the tryon image, we use VGG19 with m-feature map here.

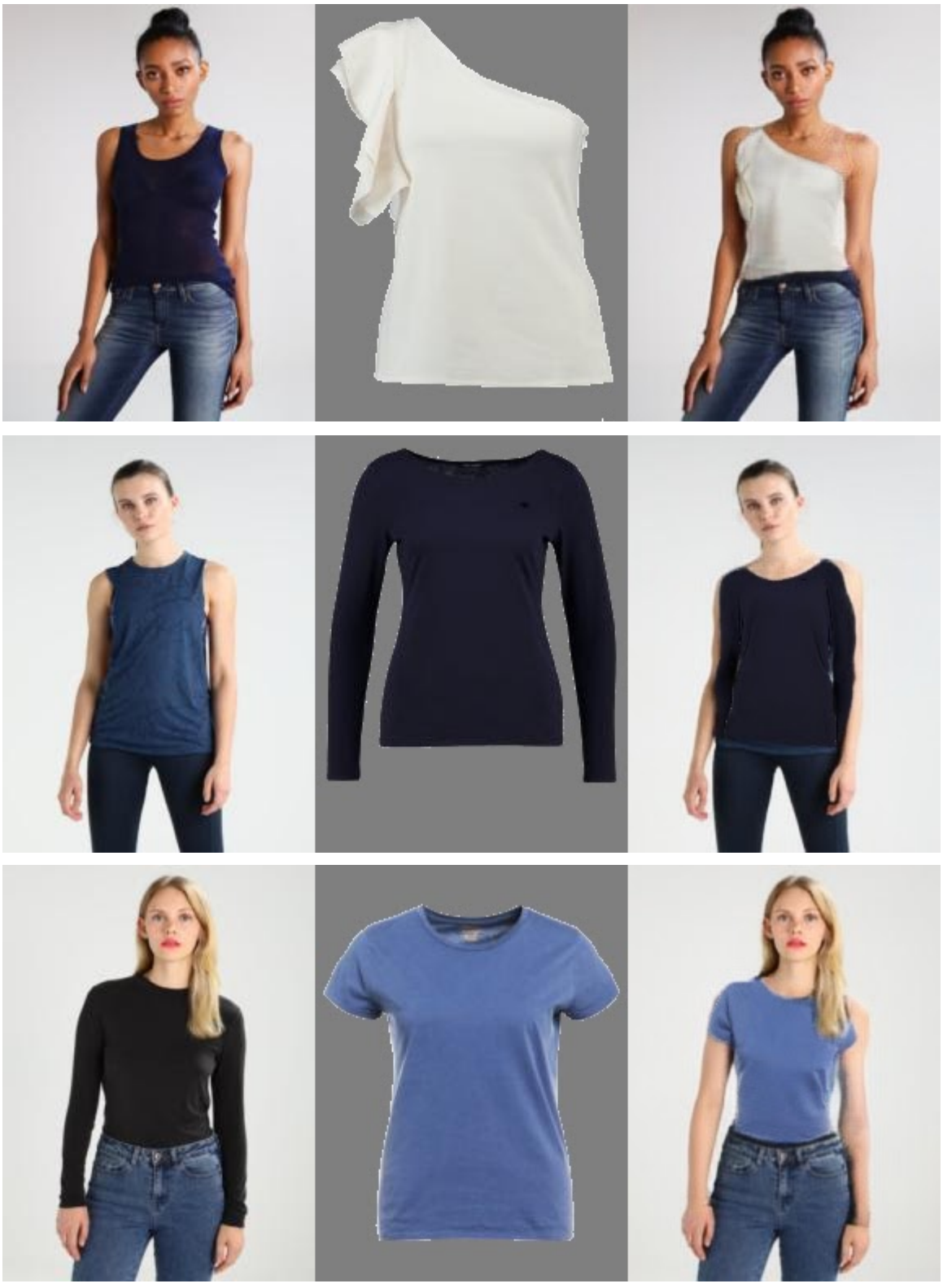
We combine loss L_{sec} in AFlowNet parts. The object function can be written as,

$$minimize \quad L_l + L_p + L_{sec}$$

- Knowledge distillation
  In contrast, the input of the student network is only the fake image and the clothes image. Thus, in most cases, the extracted features from FWM usually capture richer semantic information and the estimated appearance flows are more accurate, thus can be used to guide student networks. However, if the parsing results are not accurate, the FWM would provide totally wrong guidance, making its semantic information to student and predicted flows irresponsible.



Experimental result :




KARL
LAGERFELD
KARL
LAGERFELD


Levi's
Levi's

Conclusion:

In this work, we implemented an image-based virtual try-on model, by using a Flow-based Feature Wrapping Model and a Generative Model. Our approach treats the fake images produced by the parser based network (tutor knowledge) as input of the parser-free student network, which is supervised by the original real person image (teacher knowledge) in a self-supervised way. Besides using real images as supervisions, we further distill the appearance flows between the person image and the clothing image, to find accurate dense correspondence between them for high-quality image generation.