VITON: An Image-based Virtual Try-on Network

Radek Bartyzal

GLAMI AI

20. 7. 2020

Virtual Try On

Companies:

- TriMirror, Fits Me

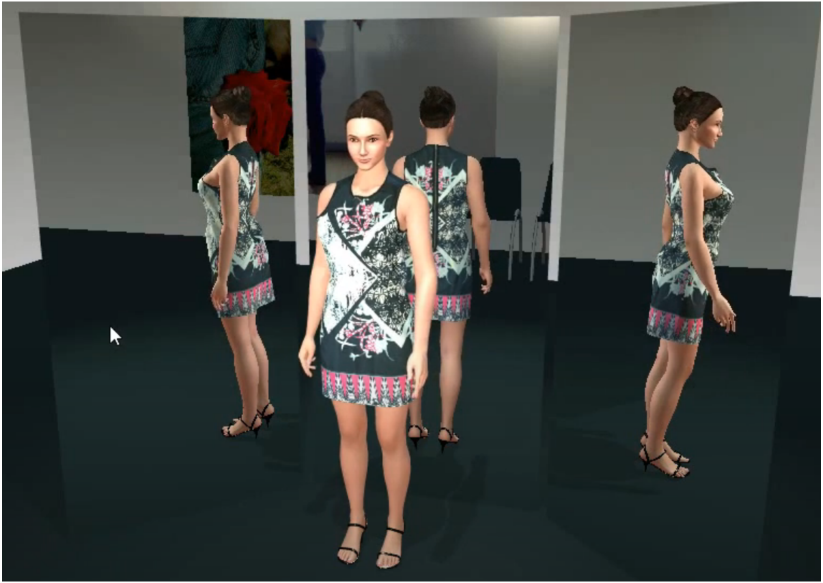
The key enabling factor behind them is the use of 3D measurements of body shape:

- captured directly by depth cameras
- inferred from a 2D image

Relevant work:

- infer 3D clothing model from 1 image [3]
- a lot of other methods that fail to produce realistic images

Example of TriMirror



Motivation

Idea:

- do not model the 3D objects, keep it all 2D
- keep the person's face + body parts = make it personal
- produce photo-realistic images = no simple avatars

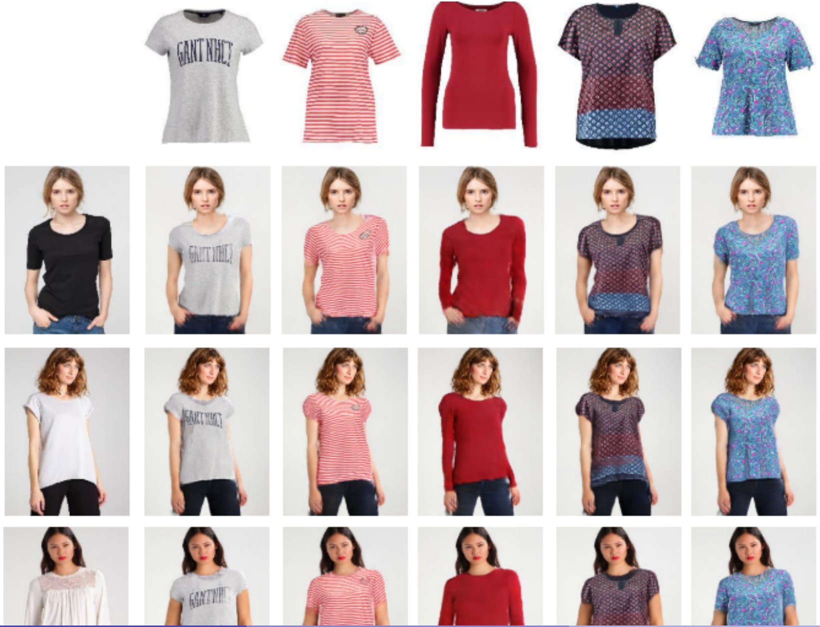
Input:

- 1 (good quality) photo of a person in any clothing
- 1 product image of a piece of clothing (white background)

Output:

- 1 image = original photo with product image put on

Example (cherry picked)



Training

Inputs:

- reference image I with a person wearing c
- product image c

Steps:

1. create clothing-agnostic representation (P) of person in I
2. synthesize the reference image I with an encoder-decoder = input is P and c, output is attempted reconstruction of I (I') + cloth mask (M)
3. use cloth mask M and product image c to generate warped product image c'
4. refinement net: input = c', I', output = 1-channel mask α
5. result = $\alpha \cdot c' + \alpha \cdot I'$

Clothing-agnostic person representation

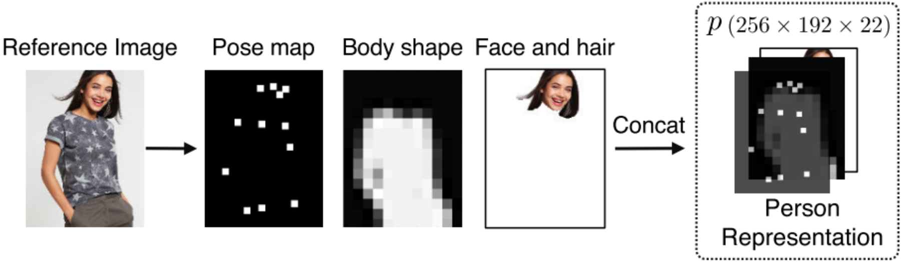


Figure: Given a reference image I, we extract the pose, body shape and face and hair regions of the person, and use this information as part of input to our generator.

- Pose: SOTA pose estimator
- Body: downsampled mask (1=human, 0=not) SOTA human parser
- Face: extract face+hair from human parser

Generate coarse image + clothing mask

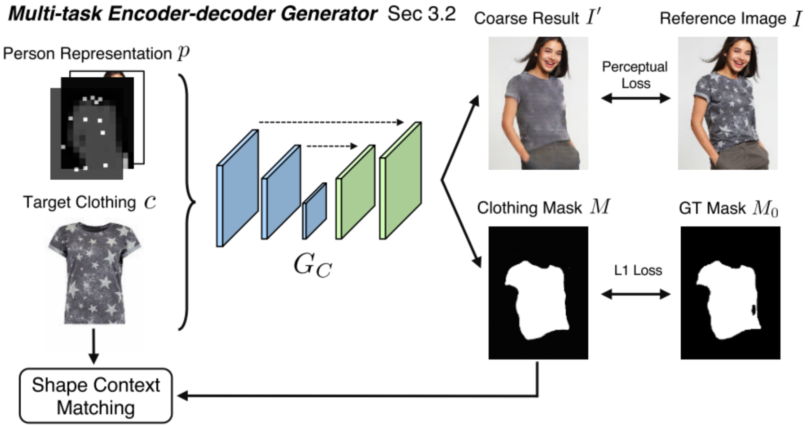


Figure: G_C = CNN U-Net with skip connections.

Generate warped product image

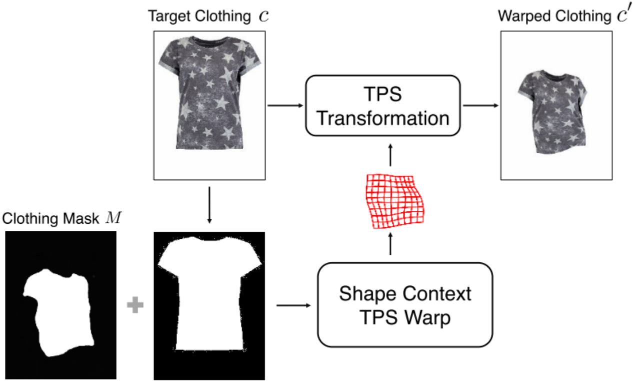


Figure: Warp the clothing item by estimating a thin plate spline (TPS) transformation with shape context matching.

Refinement network

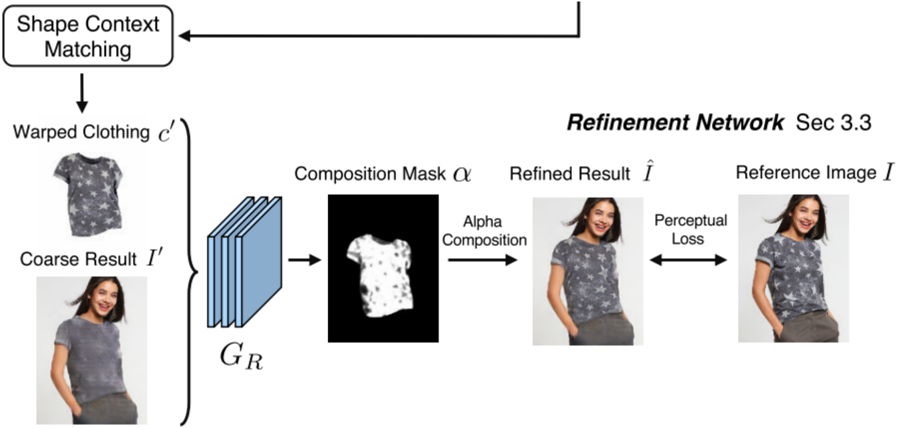


Figure: Simple CNN.

Effect of removal of body and pose

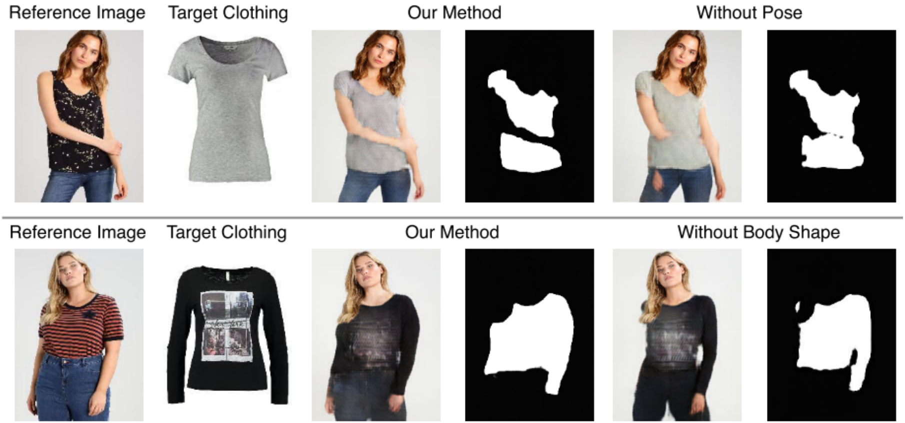


Figure: For each method, we show its coarse result and predicted clothing mask output by the cor- responding encoder-decoder generator.

Failure cases

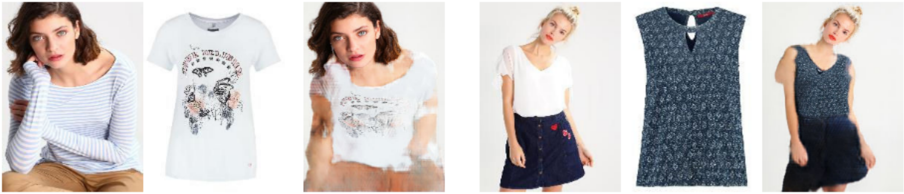
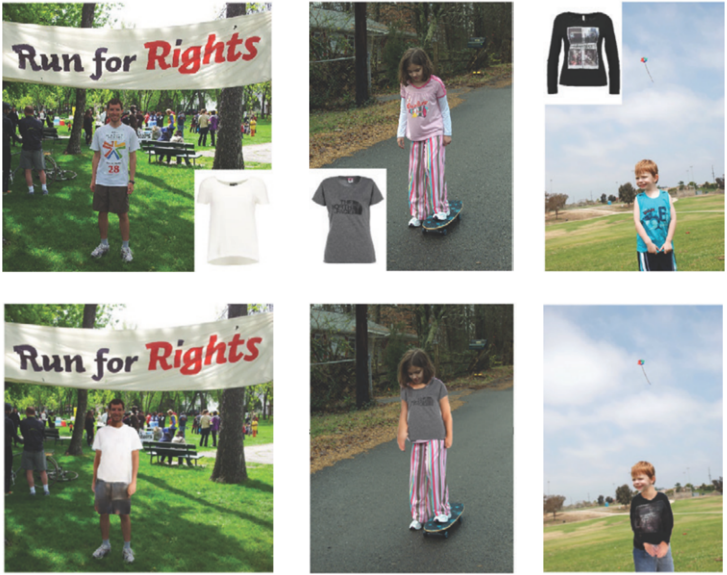


Figure: Failure cases of our method due to rarely-seen poses (example on the left) or a huge mismatch in the current and target clothing shapes (right arm in the right example).

Results on real photos from COCO dataset



Conclusion

What is required:

- trained pose estimator
- trained human parser = pixel-wise image segmentation of body parts

New version O-VITON:

- similar principle
- requires training:
  - pixel-wise semantic segmentation of body parts + clothing
  - DensePose network which captures the pose and body shape

Sources

1. Han, Xintong, et al. "Viton: An image-based virtual try-on network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. https://arxiv.org/abs/1711.08447

2. TriMirror video https://www.youtube.com/watch?v=vYJ19Z9i-zY

3. S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. Berg, and M. C. Lin. Detailed garment recovery from a single-view image. In ICCV, 2017