VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations

Matteo Fincato[1], Federico Landi[1], Marcella Cornia[1], Fabio Cesari[2], Rita Cucchiara[1]

[1]University of Modena and Reggio Emilia, [2]YOOX NET-A-PORTER GROUP

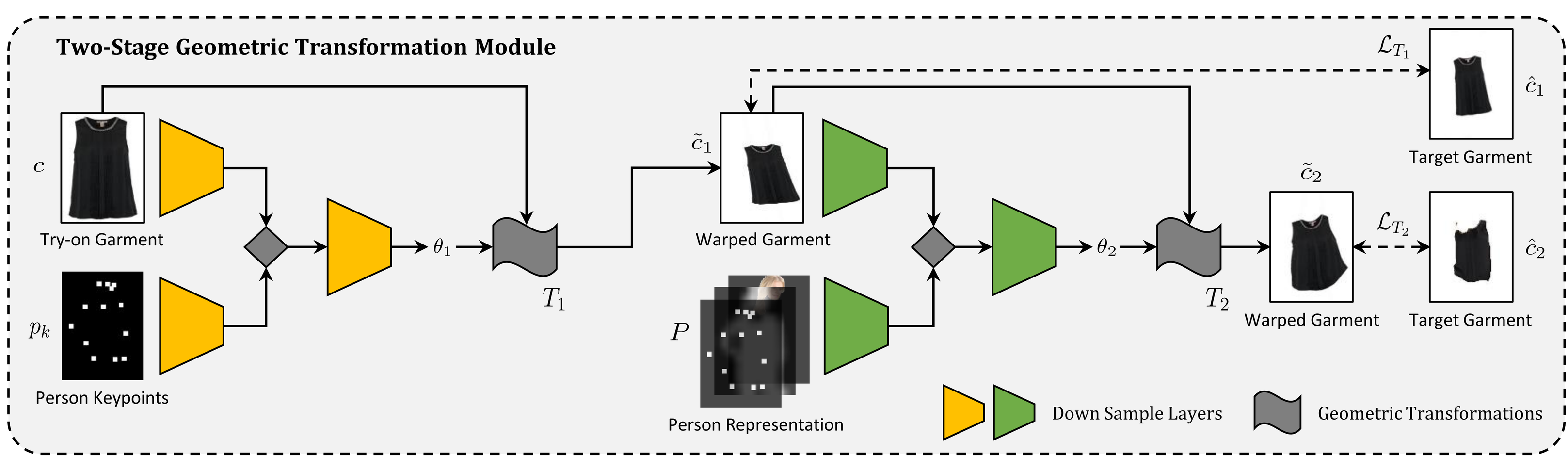
E-mail: [1]{name.surname}@unimore.it, [2]{name.surname}@ynap.com

Overview

We propose a novel solution for 2D single-pose virtual try-on which uses **multiple geometric transformations** to generates high-quality and photo-realistic images.

- Our model can generate well defined images thanks to a two-stage geometric transformation of the input garment and a generative network.
- We conduct experiments on the VITON dataset [1] and on a collected set of upper-body clothes, and we demonstrate the effectiveness of our solution both in terms of visual similarity with ground-truth images and realism of the generated try-on results.

Two-Stage Geometric Transformation Module



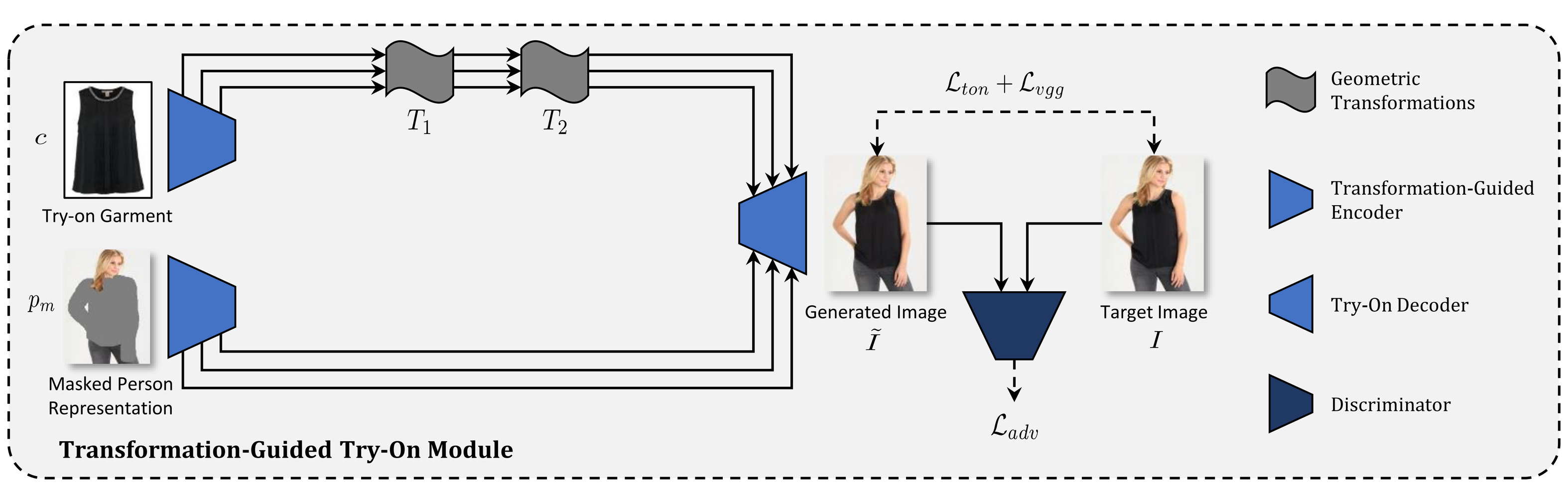
We employ two different geometric transformations, namely **affine** and **thin-plate spline**, to warp the in-shop image c of a particular garment.

- Given an image c and a pose heatmap p_k, we compute the parameters $\theta_1 = \{\boldsymbol{A}, \boldsymbol{b}\}$ for the affine transformation T_1:

$$\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{0} & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ 0 & 0 & 1 \end{bmatrix}$$

- Given the input $\tilde{c}_1 = T_1(c, \theta_1)$ and a 22-channel structure person representation P, we predict the parameters θ_2 to compute the thin-plate spline transformation. We the generate the final output $\tilde{c}_2 = T_2(\tilde{c}_1, \theta_2)$.
- The loss used to train this module is $\mathcal{L}_{GT} = \lambda_1 \mathcal{L}_{T_1} + \lambda_2 \mathcal{L}_{T_2}$ where $\mathcal{L}_{T_1}$ and $\mathcal{L}_{T_2}$ are L_1 distances between the results of the two learned transformations and the corresponding ground-truths.

Transformation-Guided Try-On Module



We generate an output image $\hat{I}$ representing the reference person wearing c by employing a U-Net architecture [3] consisting in two main components.

- **Transformation-Guided Encoder:** We apply the previous learned spatial transformations in the clothes branch, separated from the person branch:

$$T(E^i(c), \theta_1, \theta_2) = T_2(T_1(E^i(c), \theta_1), \theta_2)$$

- **Try-On Decoder:** The final result $\tilde{I}$ is guided by a pixel-level L_1, a perceptual loss [2] and an adversarial loss:

$$\mathcal{L}_{TON} = \rho_1 \mathcal{L}_{ton} + \rho_2 \mathcal{L}_{vgg} + \rho_3 \mathcal{L}_{adv},$$

where ρ_1, ρ_2 and ρ_3 are weighting coefficients.

Warping Results

The affine transformation helps the TPS generating better warped clothes that are closer to the target body pose while reducing artifacts and distortions.

Model	FID	KID	IS
CP-VTON (TPS only) [4]	101.12	6.80±0.67	3.31±0.35
VITON-GT (Affine + TPS)	**59.53**	**3.27±0.48**	**3.40±0.22**

Try-On Results

Our VITON-GT better preserves textures and details of the original clothes, thus increasing the realism of generated images.

Model	SSIM	MS-SSIM	FID	KID	IS
CP-VTON [4]	0.789	0.838	19.04	0.93±0.18	2.61±0.14
VITON-GT (no FT, no Adv. Loss)	0.879	0.919	15.32	0.58±0.19	2.72±0.14
VITON-GT (no Adv. Loss)	0.879	0.921	13.01	0.36±0.12	2.73±0.09
VITON-GT	**0.886**	**0.925**	**12.45**	**0.32±0.12**	**2.76±0.11**

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. VITON: An Image-based Virtual Try-On Network. In *CVPR*, 2018.

[2] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[3] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015.

[4] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018.