Cross Attention Based Style Distribution for Controllable Person Image Synthesis

Xinyue Zhou[1], Mingyu Yin[1], Xinyuan Chen[3], Li Sun[1,2⋆], Changxin Gao[4], and Qingli Li[1]

[1]Shanghai Key Laboratory of Multidimensional Information Processing,
[2]Key Laboratory of Advanced Theory and Application in Statistics and Data Science,
East China Normal University, Shanghai, China
[3]Shanghai AI Laboratory, Shanghai, China
[4]Huazhong University of Science and Technology, Wuhan, China

A Network architectures

In this section, we provide the details of network structure. Table 1, 2, are the network structures of the encoder E, the generator G, respectively. In Conv and Residual Block, F, K and S respectively represent the output dimension, convolution kernel size and stride. IN and LN represent instance normalization and layer normalization, respectively.

B Comparisons with the state-of-the-arts

In Fig 1, We provide additional qualitative comparisons between our method and other state-of-the-arts(*e.g.* PATN [6], GFLA [3], ADGAN [2], PISE [4], SPGNet [1], CoCosNet [5]). Results show that our method can generate more consistent appearance and pose with the target.

C Visualization of the generated parsing maps

We also provide more visualization results of the generated parsing maps in Fig 2. It is clear that cross attention matrix can accurately predict the target parsing map regardless of diverse pose and viewpoint changes, revealing the effectiveness of the proposed cross attention based style distribution module.

D Results of virtual try-on

By exchanging the channel feature of specific semantic region in the style features, our model can achieve virtual try-on task. Additional examples of virtual try-on are shown in Fig 3.

⋆ Corresponding author, email: sunli@ee.ecnu.edu.cn.

Table 1. The structure of encoder E. In E, we put I_s^i into Pre-trained VGG19 network and take the features of the corresponding layers as side branches, then concat them together with the main branch. Note that we only show one source style I_s^i as an example, where $i = 1, 2, \cdots, 8$ is the semantic index. And all I_s^i concat together lastly.

Input	I_s^i $(256 \times 176 \times 3)$
Intermediate Layers	$\mathrm{Conv}(\mathrm{F}=64, \mathrm{K}=7, \mathrm{S}=1), \mathrm{ReLU}$
	$\mathrm{Concat}(\mathrm{Pre}-\mathrm{trained\ VGG19\ conv1\_1})$
	$\mathrm{Conv}(\mathrm{F}=128, \mathrm{K}=4, \mathrm{S}=2), \mathrm{ReLU}$
	$\mathrm{Concat}(\mathrm{Pre}-\mathrm{trained\ VGG19\ conv2\_1})$
	$\mathrm{Conv}(\mathrm{F}=256, \mathrm{K}=4, \mathrm{S}=2), \mathrm{ReLU}$
	$\mathrm{Concat}(\mathrm{Pre}-\mathrm{trained\ VGG19\ conv3\_1})$
	$\mathrm{Conv}(\mathrm{F}=512, \mathrm{K}=4, \mathrm{S}=2), \mathrm{ReLU}$
	$\mathrm{Concat}(\mathrm{Pre}-\mathrm{trained\ VGG19\ conv4\_1})$
	Avg Pooling
	$\mathrm{Conv}(\mathrm{F}=256, \mathrm{K}=1, \mathrm{S}=1)$
Output	F_s^i $(1 \times 1 \times 256)$

Table 2. The structure of the generator G. In AdaIN ResBlocks and AFN ResBlocks, the content in bracket is used as side branch to affect the main branch.

Input	P_t $(256 \times 176 \times 30)$	F_s $(1 \times 1 \times 2048)$
Intermediate Layers	$\mathrm{Conv}(\mathrm{F}=64, \mathrm{K}=7, \mathrm{S}=1), \mathrm{IN}, \mathrm{ReLU}$	$\mathrm{Fc}(2048), \mathrm{ReLU}$
	$\mathrm{Conv}(\mathrm{F}=128, \mathrm{K}=4, \mathrm{S}=2), \mathrm{IN}, \mathrm{ReLU}$	$\mathrm{Fc}(256), \mathrm{ReLU}$
	$\mathrm{Conv}(\mathrm{F}=256, \mathrm{K}=4, \mathrm{S}=2), \mathrm{IN}, \mathrm{ReLU}$	$\mathrm{Fc}(256), \mathrm{ReLU}$
	$\mathrm{F_p} = \mathrm{Residual\ Blocks}(\mathrm{F}=256, \mathrm{K}=3, \mathrm{S}=1) \times 8$	$\mathrm{F_{s'}} = \mathrm{Fc}(8192), \mathrm{ReLU}$
	$\mathrm{F_{crs}} = \mathrm{AdaIN\ ResBlock}(\mathrm{F_{s'}})$	
	$\mathrm{F_{ps}} = \mathrm{CASD}\ (\mathrm{F_{crs}}, \mathrm{F_p}, \mathrm{F_s})$	
	$\mathrm{F_{ps}} = \mathrm{CASD}\ (\mathrm{F_{ps}}, \mathrm{F_p}, \mathrm{F_s})$	
	$\mathrm{F_{p'}} = \mathrm{AFN\ ResBlocks}(\mathrm{F_{ps}})$	
	$\mathrm{UpSample}(\mathrm{scale\_factor}=2)$	
	$\mathrm{Conv}(\mathrm{F}=128, \mathrm{K}=5, \mathrm{S}=1), \mathrm{LN}, \mathrm{ReLU}$	
	$\mathrm{UpSample}(\mathrm{scale\_factor}=2)$	
	$\mathrm{Conv}(\mathrm{F}=64, \mathrm{K}=5, \mathrm{S}=1), \mathrm{LN}, \mathrm{ReLU}$	
	$\mathrm{Conv}(\mathrm{F}=3, \mathrm{K}=7, \mathrm{S}=1), \mathrm{Tanh}$	
Output	$\hat{I}_t$ $(256 \times 176 \times 3)$	

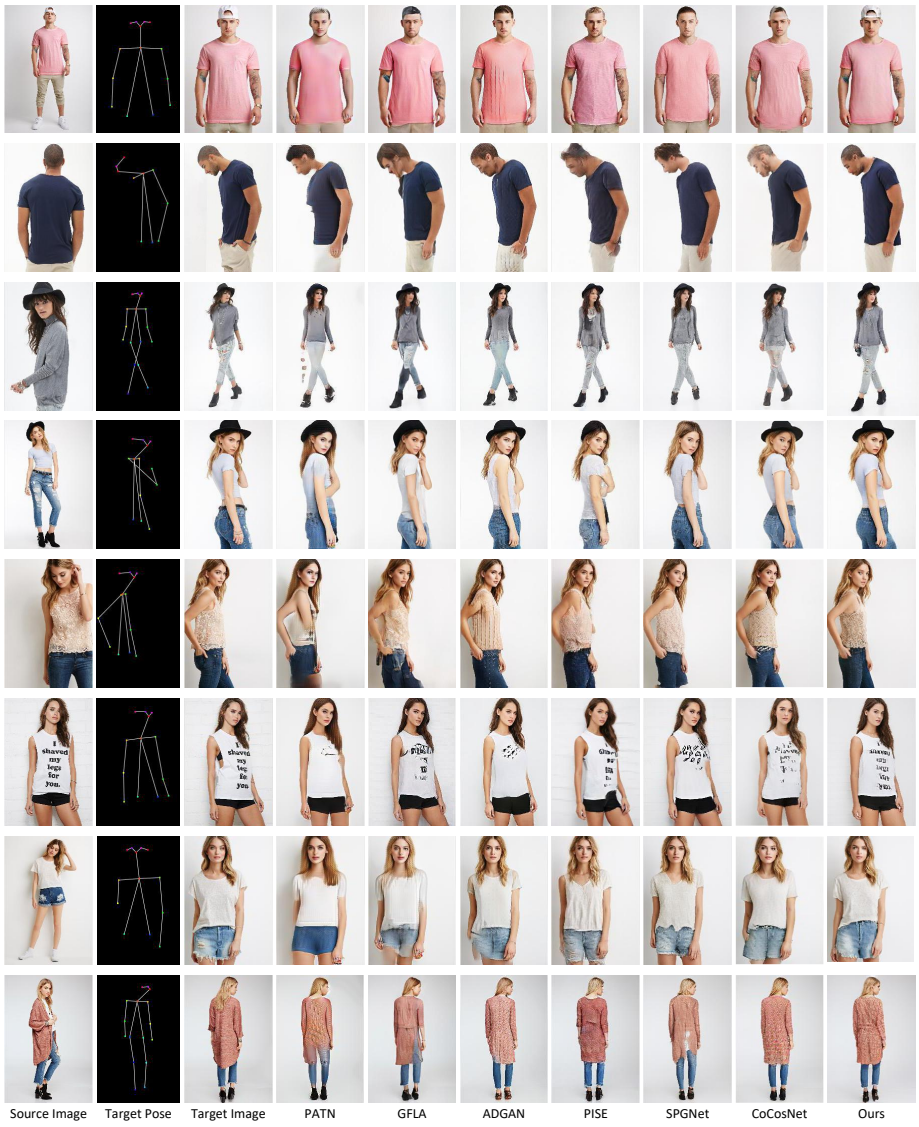

Fig. 1. Qualitative comparison between our method and other state-of-the-arts. The target ground truths and the synthesized results from each models are listed in rows.

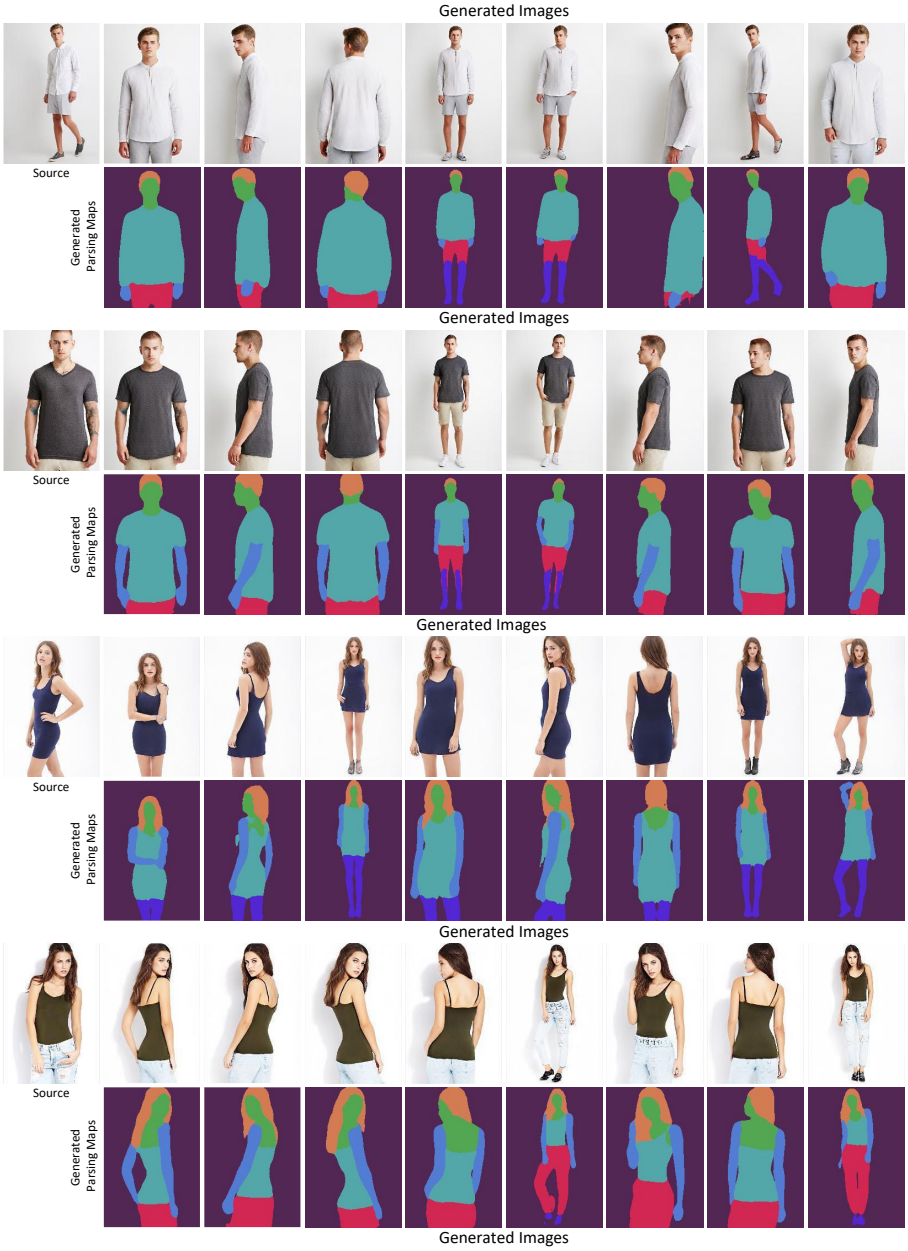

Fig. 2. Given the source image, our model is able to transfer the pose as required. The synthesized person and visualization of the generated target parsing maps are shown.

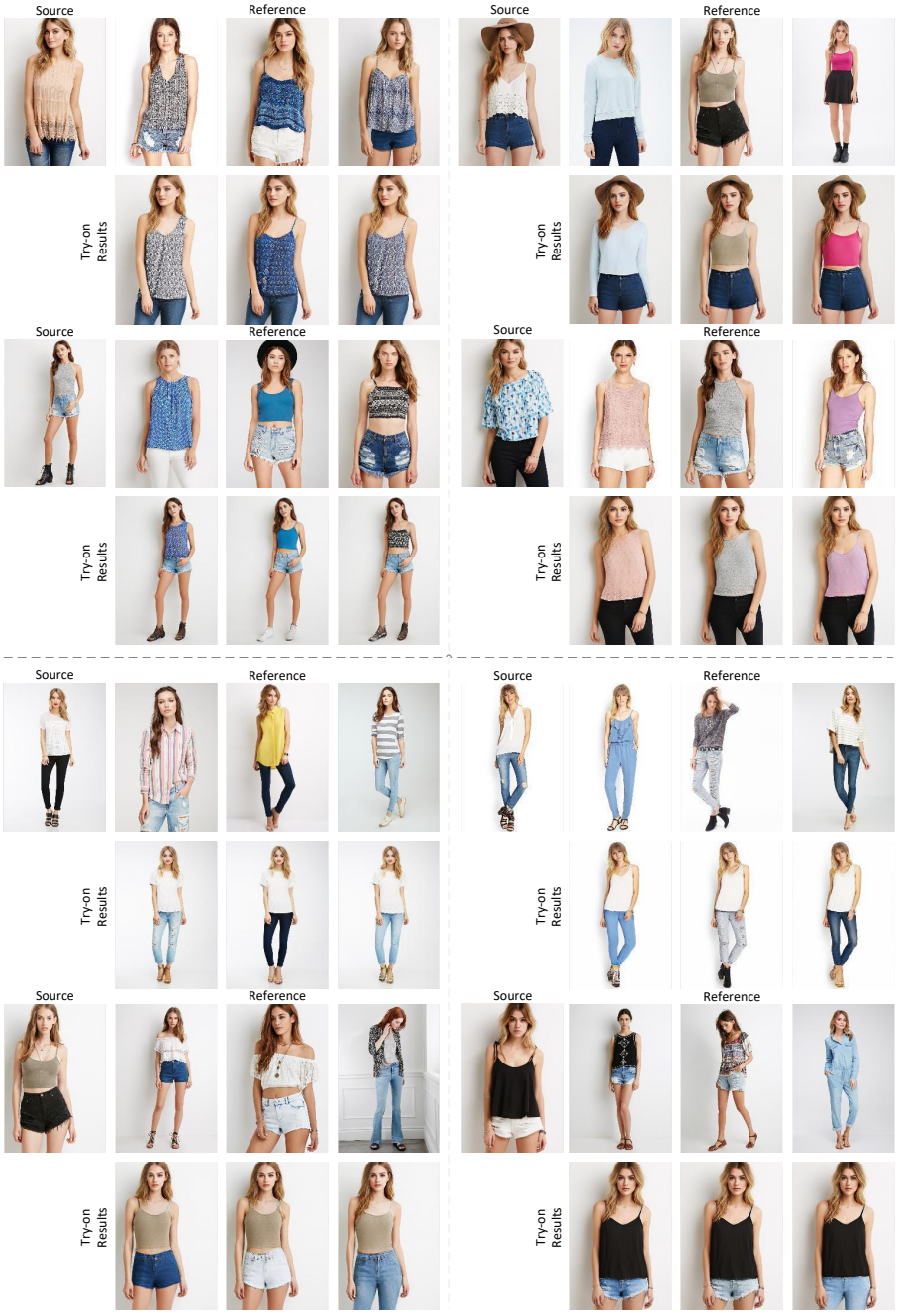

Fig. 3. Given the source image and reference images, our model is able to perform virtual try-on task. The top half is the results of trying on the upper-clothes and the bottom half is the results of trying on the pants.

References

1. Lv, Z., Li, X., Li, X., Li, F., Lin, T., He, D., Zuo, W.: Learning semantic person image generation by region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10806–10815 (2021)
2. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5084–5093 (2020)
3. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7690–7699 (2020)
4. Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7982–7990 (2021)
5. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5143–5153 (2020)
6. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2347–2356 (2019)