Supplementary Material for USC-PFN

Chenghu Du[1] **Junyin Wang**[1] **Shuqin Liu**[4] **Shengwu Xiong**[1,2,3,5,*]
[1]Wuhan University of Technology, [2]Shanghai AI Laboratory
[3]Sanya Science and Education Innovation Park, Wuhan University of Technology
[4]Wuhan Textile University, [5]Qiongtai Normal University
`{duch, wjy199708, xiongsw}@whut.edu.cn, liushuqingwtu@foxmail.com`
`https://du-chenghu.github.io/USC-PFN/`

1 Introduction

This supplementary material includes: 1) architecture details of USC-PFN, providing a comprehensive understanding of its methodology and development processes; 2) training and inference details of USC-PFN, providing a comprehensive understanding of its training and inference processes; and 3) additional qualitative results from different models, offering further insights and examples related to the research topic.

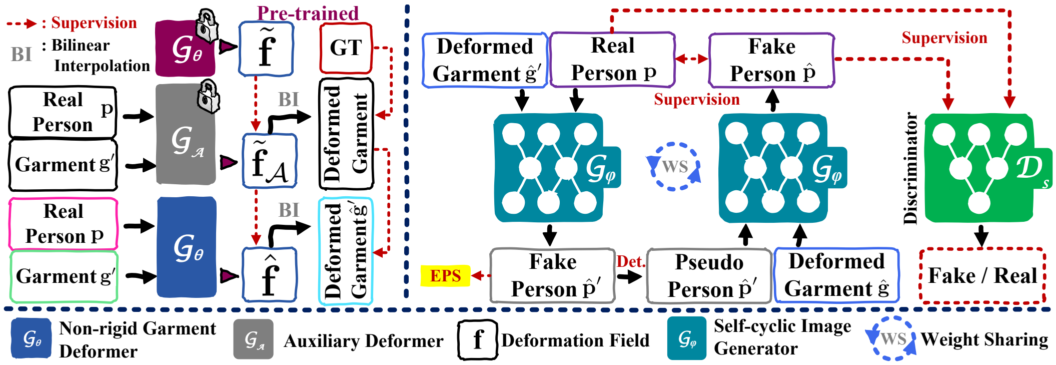
: Supervision
BI : Bilinear Interpolation
Pre-trained
$\mathcal{G}_\theta$
$\tilde{\mathbf{f}}$
GT
Real Person $\mathbf{p}$
Garment $\mathbf{g}'$
$\mathcal{G}_\mathcal{A}$
BI
$\tilde{\mathbf{f}}_\mathcal{A}$
Deformed Garment
Real Person $\mathbf{p}$
Garment $\mathbf{g}'$
$\mathcal{G}_\theta$
BI
$\hat{\mathbf{f}}$
Deformed Garment $\hat{\mathbf{g}}'$
Deformed Garment $\hat{\mathbf{g}}'$
Real Person $\mathbf{p}$
Fake Person $\hat{\mathbf{p}}$
Supervision
Supervision
$\mathcal{G}_\varphi$
WS
$\mathcal{G}_\varphi$
Discriminator
$\mathcal{D}_s$
EPS
Fake Person $\hat{\mathbf{p}}'$
Det.
Pseudo Person $\hat{\mathbf{p}}'$
Deformed Garment $\hat{\mathbf{g}}$
Fake / Real
$\mathcal{G}_\theta$ Non-rigid Garment Deformer
$\mathcal{G}_\mathcal{A}$ Auxiliary Deformer
f Deformation Field
$\mathcal{G}_\varphi$ Self-cyclic Image Generator
WS Weight Sharing

Figure 1: Overview of our self-cycle consistency network: paired garment-person images ($\mathbf{g}$, $\mathbf{p}$) and an arbitrary garment $\mathbf{g}'$ serve as the training data. **Left:** $\mathcal{G}_\theta$ is pre-trained using paired images to obtain the ground-truth deformation field $\tilde{\mathbf{f}}$ for the auxiliary deformer $\mathcal{G}_\mathcal{A}$ to optimize $\tilde{\mathbf{f}}_\mathcal{A}$. Then, $\mathcal{G}_\theta$ is further trained by using $\tilde{\mathbf{f}}_\mathcal{A}$ of $\mathcal{G}^*_\mathcal{A}$ with unpaired images. **Right:** $\mathcal{G}_\varphi$ takes unpaired ($\hat{\mathbf{g}}'$, $\mathbf{p}$) to generate a fake person $\hat{\mathbf{p}}'$. Then, the real person $\mathbf{p}$ is reconstructed by re-feeding $\hat{\mathbf{p}}'$ and $\hat{\mathbf{g}}$ into $\mathcal{G}_\varphi$, ensuring self-cycle consistency $\mathbf{p} \approx \hat{\mathbf{p}}$. $\mathcal{D}_s$ is the discriminator that learns to distinguish real and fake images. **EPS** represents enhanced pixel-level supervision and **Det.** represents detaching tensor $\hat{\mathbf{p}}'$ from the computation graph.

2 Architecture Details

In our architecture, any encoder-decoder network can be used as our generator. In our experiments, to achieve fast convergence of the network during the theoretical validation phase, we adopt Res-UNet [1] as Non-rigid Garment Deformer $\mathcal{G}_\theta$ (NGD), auxiliary deformer $\mathcal{G}_\mathcal{A}$, and Self-cyclic Image Generator $\mathcal{G}_\varphi$ (SIG). The discriminator is from pix2pixHD [2]. The complete code for Res-UNet can be found in our open-source repository `https://github.com/du-chenghu/USC-PFN/`.

---

[*] Shengwu Xiong is corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3 Training and Inference Details

As shown in Figure 1, our framework consists of three generators ($\mathcal{G}_\theta$, $\mathcal{G}_\mathcal{A}$, and $\mathcal{G}_\varphi$) and one discriminator ($\mathcal{D}_s$).

3.1 Pre-trained Parser for Supervision

The first thing that needs to be specifically pointed out is that **"Parser-Free Virtual Try-On"** refers to the training and inference stages, where in the input of the main generator, there is no need for human parsing (i.e., using a parser-generated human semantic segmentation map). Because [3] has confirmed that including human parsing in the generator's input, and if there are errors in this parsing, can negatively impact the results, eliminating human parsing (i.e. the parser) at the input end of the generator, is highly beneficial.

In order to facilitate the accurate localization of different semantic parts of the human body during the calculation of loss in the training stage, we utilized a pre-trained parser (denotes as τ). This parser, similar to the one used in [4, 5, 3, 6–8], is designed to predict a posterior "after-try-on" semantics, which corresponds to the semantic map of the desired try-on result. Only by inputting the person representation $\mathcal{R}$ and the target garment $\mathbf{g}'$ into τ, the target semantic map $\delta' \in \{0,1\}^{21\times H\times W}$ can be generated, which can be represented as:

$$\delta' = \tau\left(\mathcal{R}, \mathbf{g}'\right). \tag{1}$$

In the main text, we generate the result $\hat{\mathbf{p}}'$ using unpaired $(\mathbf{p}, \hat{\mathbf{g}}')$. With the help of δ', $\hat{\mathbf{p}}'_\mathbf{s}$ can be obtained through:

$$\hat{\mathbf{p}}'_\mathbf{s} = (\delta'_{arms} \cup \delta'_{neck}) \odot \hat{\mathbf{p}}', \tag{2}$$

where $\odot$ denotes entry-wise multiplication. $\hat{\mathbf{p}}'_\mathbf{g}$ can be obtained through:

$$\hat{\mathbf{p}}'_\mathbf{g} = \delta'_{garment} \odot \hat{\mathbf{p}}'. \tag{3}$$

$\hat{\mathbf{p}}'_\mathbf{c}$ can be obtained through:

$$\delta'_{identity} = 1 - \left(\delta'_{arms} \cup \delta'_{neck} \cup \delta'_{garment}\right), \tag{4}$$

$$\hat{\mathbf{p}}'_\mathbf{c} = \delta'_{identity} \odot \hat{\mathbf{p}}', \tag{5}$$

where $-$ denotes entry-wise subtraction and 1 is all-ones tensor.

3.2 Non-rigid Garment Deformer $\mathcal{G}_\theta$ (NGD)

Training. During the training phase, we first train $\mathcal{G}_\theta$ based on Markov Random Field theory using paired garment-person images:

$$\tilde{\mathbf{f}} = \mathcal{G}_\theta\left(\mathbf{p}, \mathbf{g}\right). \tag{6}$$

The reason for using paired $(\mathbf{p}, \mathbf{g})$ for training is that the clothes on $\mathbf{p}$ and $\mathbf{g}$ look the same in appearance. Therefore, we can learn implicit correlations between the two during the deformation process. These correlations could be related to color, shape, perspective changes, texture variations, shadows and lighting, depth of field, etc. It is conceivable that if $\mathbf{p}$ is replaced with the person representation $\mathcal{R}$, no relevant information other than the shape can be extracted at all.

Unfortunately, $\mathcal{G}_\theta$ does not need correlations in terms of color, shape, and texture. This results in a significant entanglement of features between target clothing and clothing worn on the person in $\mathcal{G}_\theta$'s feature space. Consequently, $\mathcal{G}_\theta$ can only be effective when the input data is paired, and it fails to deform properly when encountering unpaired samples.

We employ an auxiliary deformer, $\mathcal{G}_\mathcal{A}$, to address this issue. Since $\mathcal{R}$ does not contain the aforementioned correlations except for the shape, we modify $\mathcal{G}_\mathcal{A}$'s input from $(\mathbf{p}, \mathbf{g})$ to $(\mathcal{R}, \mathbf{g})$ to estimate the target deformation field $\tilde{\mathbf{f}}_\mathcal{A}$:

$$\tilde{\mathbf{f}}_\mathcal{A} = \mathcal{G}_\mathcal{A}\left(\mathcal{R}, \mathbf{g}\right). \tag{7}$$

However, the limited amount of information in $\mathcal{R}$ forces us to utilize previously pre-trained $\mathcal{G}_\theta$ to supplement the aforementioned correlations for $\mathcal{G}_\mathcal{A}$. In other words, we optimize $\mathcal{G}_\mathcal{A}$'s result $\tilde{\mathbf{f}}_\mathcal{A}$ by leveraging the result $\tilde{\mathbf{f}}$ from $\mathcal{G}_\theta$, minimizing the differences between $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{f}}_\mathcal{A}$:

$$\tilde{\mathbf{f}}^*_\mathcal{A} = \text{argmin}_{\tilde{\mathbf{f}}_\mathcal{A}} \left\|\tilde{\mathbf{f}}_\mathcal{A} - \tilde{\mathbf{f}}\right\|. \tag{8}$$

After obtaining the optimal $\mathcal{G}_{\mathcal{A}}^{*}$, we continue to train the deformer $\mathcal{G}_{\theta}$ by taking unpaired $(\mathbf{p}, \mathbf{g}')$ as input and $\mathbf{f}_{\mathcal{A}}$ as supervision, to disentangle the feature entanglement of $\mathcal{G}_{\theta}$:

$$\tilde{\mathbf{f}}_{\mathcal{A}} \approx \hat{\mathbf{f}}, \text{ with } \hat{\mathbf{f}} = \mathcal{G}_{\theta}(\mathbf{p}, \mathbf{g}'), \; \tilde{\mathbf{f}}_{\mathcal{A}} = \mathcal{G}_{\mathcal{A}}^{*}(\mathcal{R}, \mathbf{g}'). \tag{9}$$

Inference. During the inference phase, we exclusively utilize $\mathcal{G}_{\theta}$ to warp arbitrary garment.

3.3 Self-cyclic Image Generator $\mathcal{G}_{\varphi}$ (SIG)

Training. During the training phase, we divided the training of $\mathcal{G}_{\varphi}$ into two steps as shown in Figure 1. In the first step, we train $\mathcal{G}_{\varphi}$ taking unpaired $(\mathbf{p}, \hat{\mathbf{g}}')$ as input to obtain the fake result $\hat{\mathbf{p}}'$:

$$\hat{\mathbf{p}}' = \mathcal{G}_{\varphi}(\mathbf{p}, \hat{\mathbf{g}}'). \tag{10}$$

Then, in the second step, $\hat{\mathbf{p}}'$ detached from the computation graph and $\hat{\mathbf{g}}$ are used as input of $\mathcal{G}_{\varphi}$ to reconstruct the real person:

$$\mathbf{p} \approx \hat{\mathbf{p}}, \text{ with } \hat{\mathbf{p}} = \mathcal{G}_{\varphi}(\hat{\mathbf{p}}', \hat{\mathbf{g}}). \tag{11}$$

In order to accelerate the convergence of $\mathcal{G}_{\varphi}$, $\mathbf{p}$ and $\hat{\mathbf{p}}'$ can be combined using τ and used as input in the second step:

$$\hat{\mathbf{p}}' = \delta_{save} \odot \mathbf{p} + (1 - \delta_{save}) \odot \hat{\mathbf{p}}', \tag{12}$$

where

$$\delta = \tau(\mathcal{R}, \mathbf{g}), \tag{13}$$

$$\delta_{identity} = 1 - (\delta_{arms} \cup \delta_{neck} \cup \delta_{garment}), \tag{14}$$

$$\delta_{save} = \left(\delta_{identity} \cap \delta'_{identity}\right) \cup (\delta_{arms} \cap \delta'_{arms}) \cup (\delta_{neck} \cap \delta'_{neck}). \tag{15}$$

Additionally, parts of the skin $\mathbf{s_p}$ can be obtained through:

$$\mathbf{s_p} = ((\delta_{arms} \cap \delta'_{arms}) \cup (\delta_{neck} \cap \delta'_{neck})) \odot \mathbf{p}. \tag{16}$$

Inference. During the inference phase, the warped garment image and the person image are fed to $\mathcal{G}_{\varphi}$ to synthesize the try-on result.

4 More Qualitative Results

We show more qualitative comparisons of the try-on results among our proposed USC-PFN and other cutting-edge methods (CP-VTON+[9], ACGPN[4], DCTON[5], PF-AFN[3], and FS-VTON[7]) on VITON Zalando dataset [10]. As shown in Figure 2. overall, our model generates better try-on images.

References

[1] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[2] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[3] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.

[4] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020.

[5] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16928–16937, 2021.

[6] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442, 2021.

[7] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[8] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2022.

[9] 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 *CVPR Workshops*, volume 3, pages 10–14, 2020.

[10] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

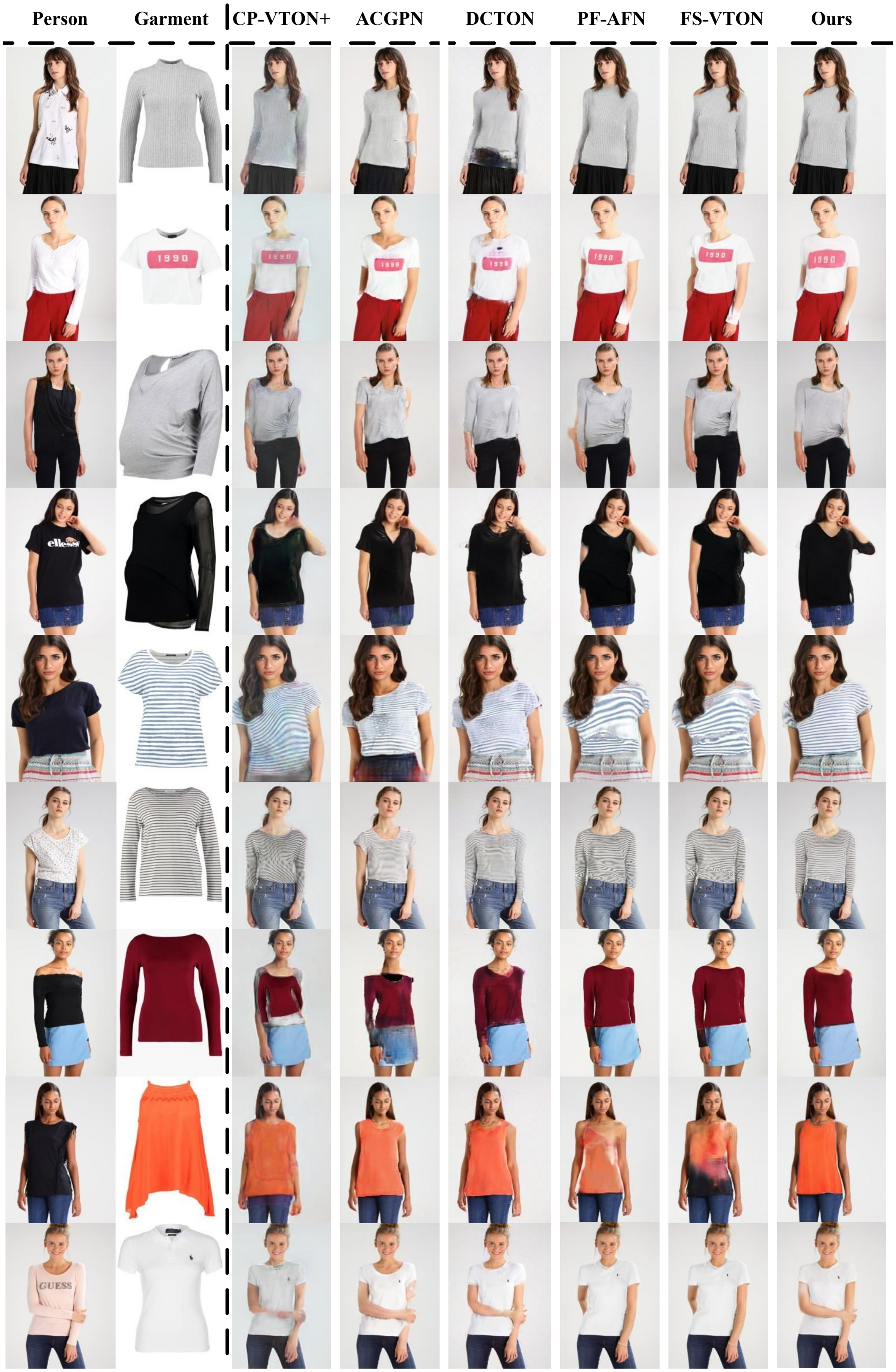

Figure 2: Qualitative results of different methods (CP-VTON+[9], ACGPN[4], DCTON[5], PF-AFN[3], FS-VTON[7], and ours) in the unpaired setting.