Linking Garment with Person via Semantically Associated Landmarks for Virtual Try-On Supplementary Material

Keyu Yan[1,2,3*] Tingwei Gao[1*] Hui Zhang[2,3] Chengjun Xie[2†]
[1] Alibaba Group
[2]Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[3]University of Science and Technology of China, China
{keyu,hui928}@mail.ustc.edu.cn, tingwei.gtw@alibaba-inc.com, cjxie@iim.ac.cn

Overview

The Supplementary Material is structured as follows. In Section 1, the labelling rule of the proposed semantically associated landmarks is introduced in detail. Due to the page limits, it is not included in the main manuscript. The differences between our dataset and existing popular garment landmark datasets are also displayed. More qualitative visualization comparisons over the VITON-HD [1] and VITON [4] datasets are presented in Section 2. In addition to standard garments, we also especially show the try-on results of non-standard garments. Section 3 aims to provide examples in which the original try-on results of SAL-VTON are continuously controlled via manually manipulating semantically associated landmarks. In Section 4, more ablation experiment results are provided.

1. Details about the semantically associated landmarks dataset.

To obtain the semantically associated landmarks dataset, we re-annotate images on the popular virtual try-on benchmarks including VITON-HD [1] and VITON [4], and propose a new dataset named Semantically Associated Landmarks for Human and Garment (SAL-HG). The VITON-HD dataset and VITON dataset respectively contain a total of 13,679 and 16,253 image pairs, each consisting of an in-shop garment image and an image of a person wearing the garment.

A unified labelling rule of landmarks is applied for diverse styles of garments. To be specific, landmarks are defined according to different regions of garments and persons, so as to ensure that the landmarks of the same serial number on different types of garments have the same semantics. As illustrated in Fig.1, each image has 32 landmarks with several attributes (visible, occluded and absent). When a region of the garment is lacking, the attributes of such landmarks in the lacking region become absent. The sleeves of the right garment in Fig. 1 are lacking, thus the attributes of landmarks in the upper arm region, elbow region, forearm region and sleeve opening region are absent.

The described labelling rule applies not only to standard garments, but also to non-standard garments. More examples are shown in Fig. 2. Although there is considerable variation in the types of garments, the regions with clear semantics are universal. Therefore, the landmarks are labeled according to the semantic regions, so that the labelling rule can be widely applied. To improve the quality of the semantically associated landmark dataset, data labelers are required to simultaneously annotate semantically associated landmarks on the in-shop garment and the corresponding person.

In Fig. 3, some samples are selected to show the differences between different clothing landmark datasets. From the comparisons, the following observations can be made: 1) Early works such as DeepFashion [7] and ULD [9] datasets commonly focus on standard garments and label considerably few landmarks. 2) The widely used DeepFashion2 [3] dataset divides multiple types of standard garments and definite labelling rules for each garment type. As different landmarks are set for different garments types, the semantics of landmarks are not universal. For example, the landmark 12 in the short sleeve top represents the armpit but in the long sleeve top, it represents the sleeve opening. To utilize the semantics of landmarks, researchers need to first detect the different garment categories. However, there are many non-standard garments in virtual try-on that are not included in their categories. 3) The labelling rule of the FashionAI [11] dataset is similar to ours, but we have a finer division of the semantic regions and more landmarks in the semantic regions.

[*]Co-first authors contributed equally, [†] corresponding author.

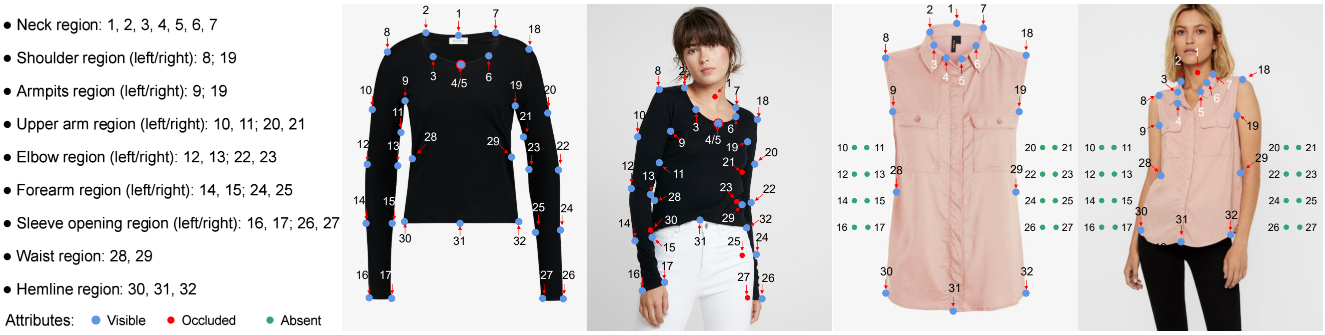
● Neck region: 1, 2, 3, 4, 5, 6, 7
● Shoulder region (left/right): 8; 19
● Armpits region (left/right): 9; 19
● Upper arm region (left/right): 10, 11; 20, 21
● Elbow region (left/right): 12, 13; 22, 23
● Forearm region (left/right): 14, 15; 24, 25
● Sleeve opening region (left/right): 16, 17; 26, 27
● Waist region: 28, 29
● Hemline region: 30, 31, 32
Attributes: ● Visible ● Occluded ● Absent

Figure 1. The labelling rule of SAL-HG. The different colors of landmarks represent different attributes. For easy observation, two landmarks where the positions coincide are indicated by blue dots with a red border, such as landmark 4 and landmark 5 in the left garment.

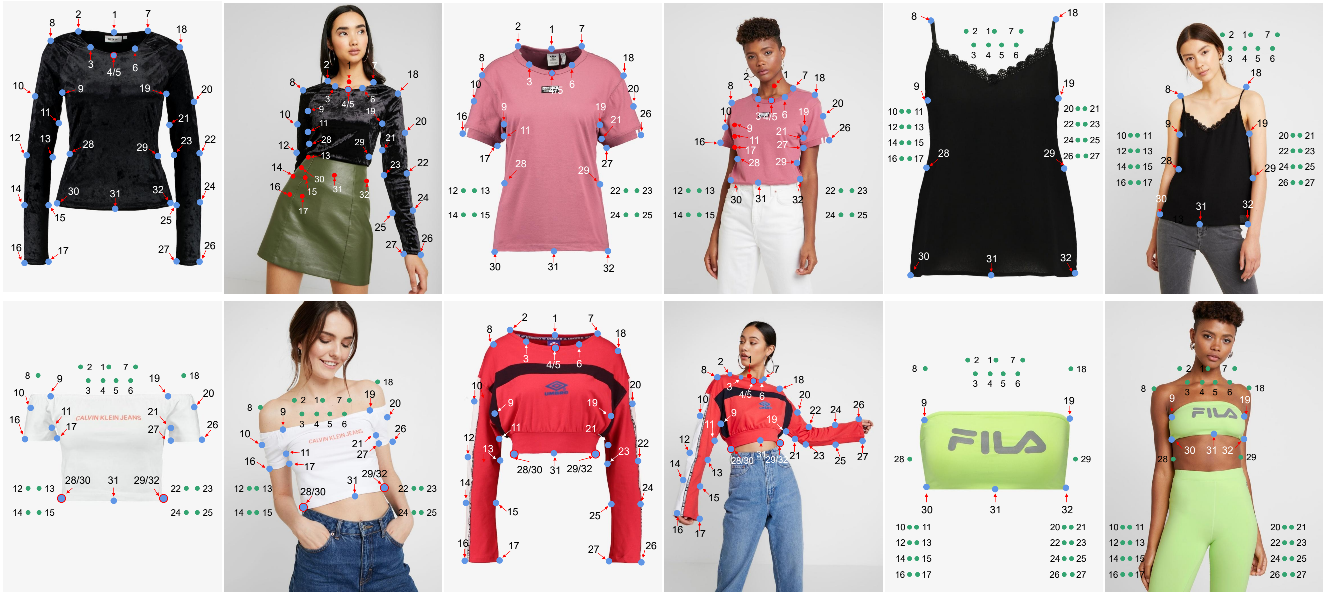

Figure 2. Examples of SAL-HG. The first row shows examples of **standard garments** that are well-considered in existing clothing landmark datasets. The second raw shows examples of **non-standard garments** in the virtual try-on datasets.

2. Qualitative comparison over VITON and VITON-HD datasets.

The visual comparison between our method and state-of-the-art high-resolution virtual try-on methods (VITON-HD [1] and HR-VITON [6]) on the representative VITON-HD dataset is shown in Fig. 4. The former two rows in Fig. 4 are the try-on results of standard garments such as T-shirts, vests, and long-sleeve tops. Compared with other methods, our method can preserve details and synthesise more photo-realistic results in the case of simple standard garments.

To verify the effectiveness of the semantically associated landmarks dataset utilising a unified labelling rule for diverse styles of garments, the try-on results of several representative non-standard garments are displayed in Fig. 4. It is clearly seen that the proposed method can synthesize photo-realistic results with accurate garment shapes, regardless of the garment type. Non-standard garments are covered in our dataset while existing popular garment landmark datasets do not contain these hard samples, which makes them difficult to handle these hard situations well. Therefore, it can be affirmed that the proposed semantically associated landmarks dataset is more versatile for virtual try-on.

Moreover, the visual comparison results on the VITON testing dataset in a standard to non-standard manner are shown in Fig. 5. It can be clearly seen that our method performs the best compared with other competitive methods (CP-VTON+ [8], ACGPN [10], PF-AFN [2] and StyleFlow [5]), indicating the superiority of our proposed method.

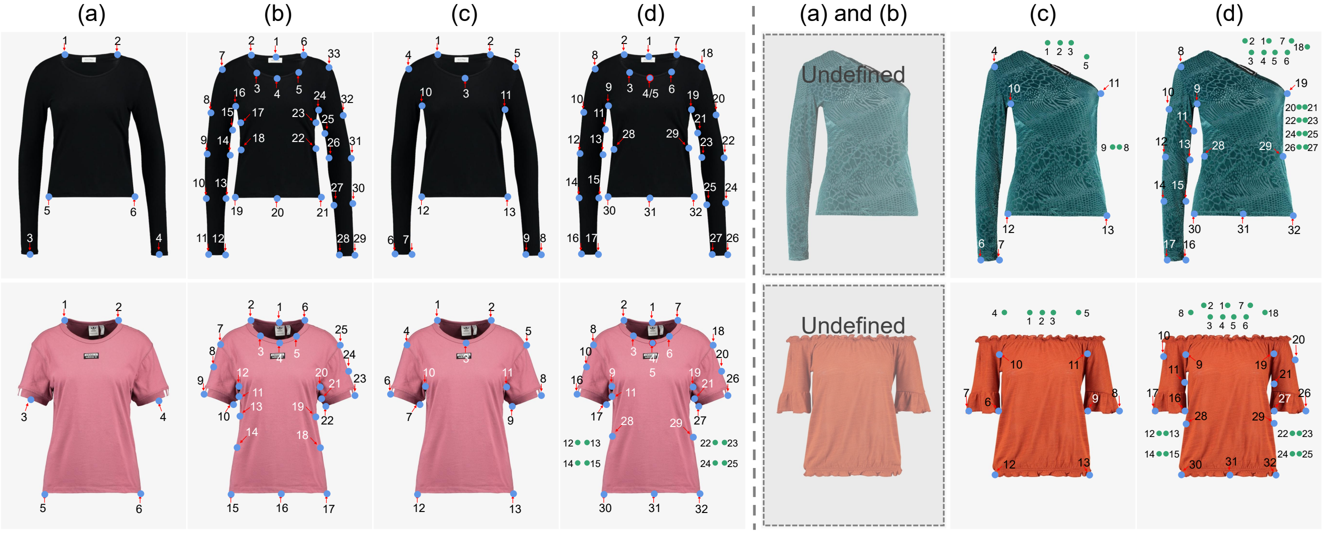

Figure 3. Comparison of labelling rules on the standard garments (left) and non-standard garments (right) for different landmark datasets. (a) DeepFasion and ULD, (b) DeepFashion2, (c) FashionAI and (d) Ours SAL-HG.

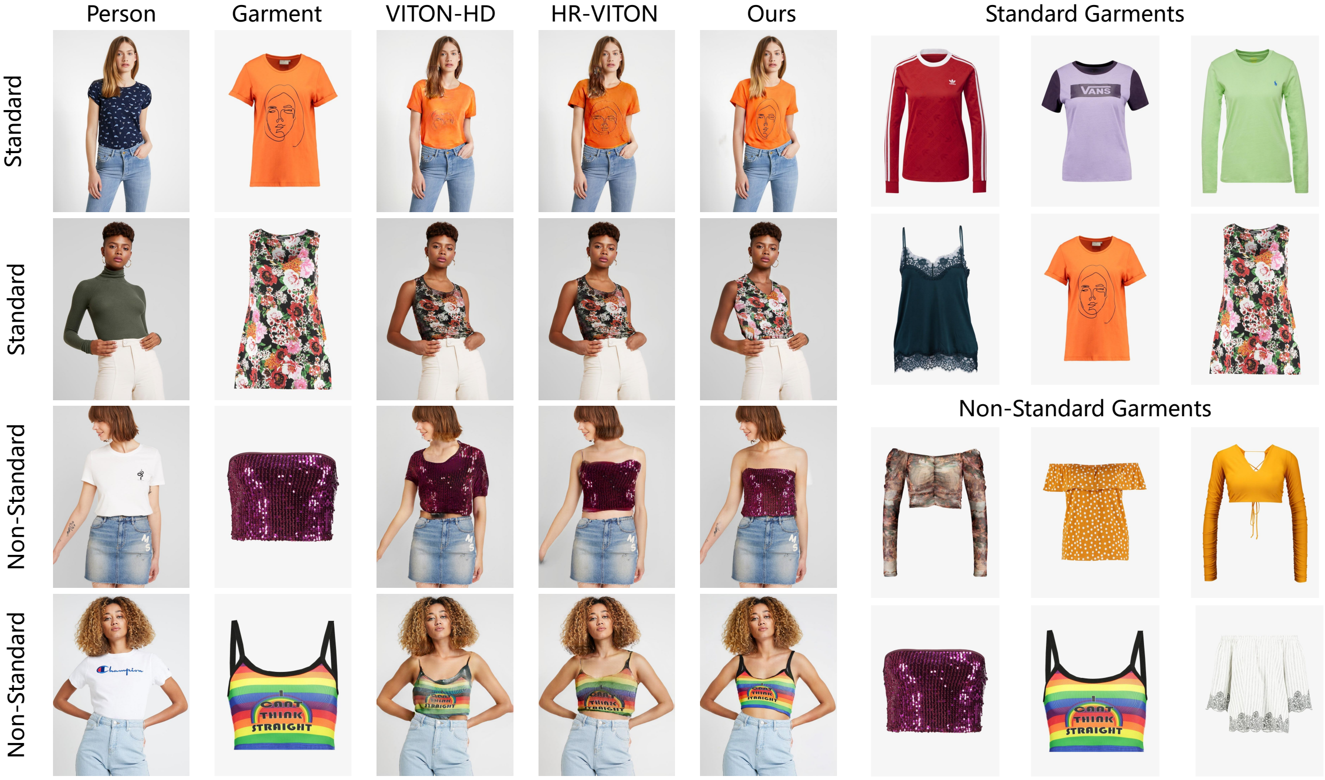

Figure 4. Qualitative results from different models (VITON-HD, HR-VITON and ours) on the VITON-HD testing dataset.

3. Try-on image editing via semantically associated landmarks.

In the main manuscript, the extended experiments show that manipulating semantically associated landmarks in different semantic regions can achieve different garment editing effects. And the editing effects of different regions can be combined to obtain the combined results.

Furthermore, the editing effects can be continuously controlled by manipulating semantically associated landmarks. As

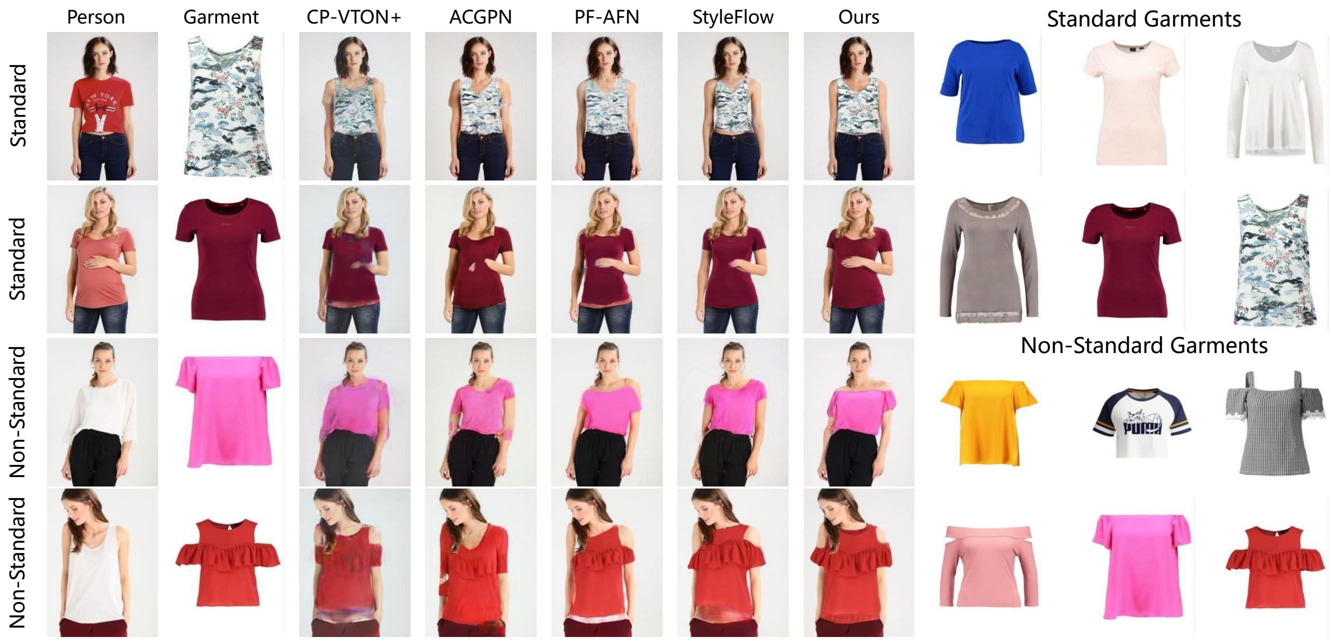

Figure 5. Qualitative results from different models (CP-VTON+, ACGPN, PF-AFN, StyleFlow and ours) on the VITON testing dataset.

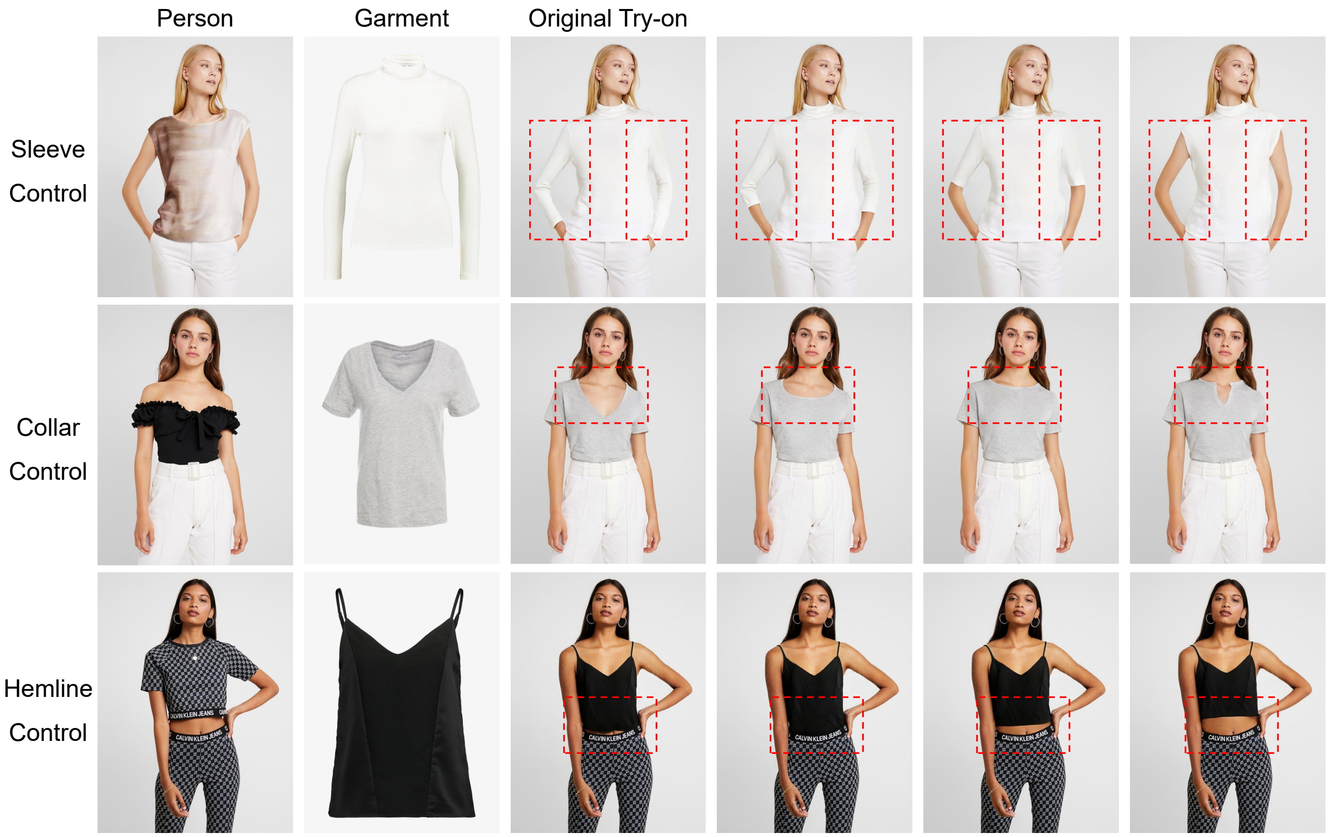

Figure 6. Control the try-on results of SAL-VTON via different semantically associated landmark settings. The shape or length of local garment regions on the try-on image can be **continuously controlled**.

shown in Fig. 6, the continuously controlled regions are highlighted by red boxes. In the first row, the length of the sleeves on the original try-on result is continuously shortened. In the second row, the shape of the collar is successfully controlled. In the third row, we continuously adjust the position of the hemline by changing the semantically associated landmarks of the hemline and waist regions.

4. More ablation experiment results.

In this section, we provide more ablation experiment results that are not included in the paper. An ablation study using fewer landmarks are implemented. We remove the landmarks (10,11,14,15,20,21,24,25) on the upper arm and forearm regions to obtain 24 landmarks in total. Based on the above, we further remove the landmarks (1,3,6,28,29,31) on the neck, waist and hemline regions to obtain 18 landmarks. In this way, "every other landmark removed" is achieved without breaking the semantic relationship of the local regions. The model is retrained on the reduced version of the landmarks. The results of ablation experiments using fewer landmarks are reported in Table 1.

In addition, the quantitative results with or without local flow are shown in Fig. 7. The proposed local flow properly handles the misaligned local regions.

Table 1. The results of ablation study using fewer landmarks.

Config	SSIM↑	PSNR↑	FID↓
32 landmarks	0.92	28.29	5.74
24 landmarks	0.91	27.93	5.89
18 landmarks	0.91	27.64	6.16

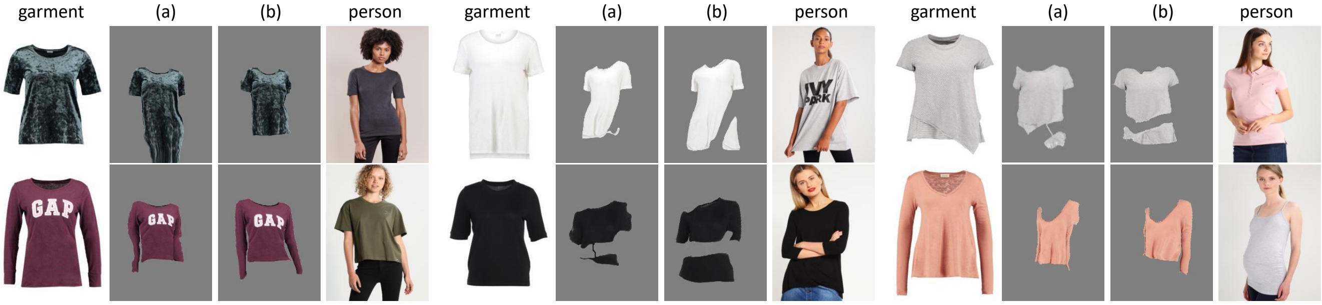

Figure 7. The visual comparison of the warping results. (a) without the local flow, (b) with the local flow.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 1, 2

[2] 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. 2

[3] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019. 1

[4] 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. 1

[5] 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. 2

[6] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 2

[7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 1

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2020. 2

[9] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 172–180, 2017. 1

[10] 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 Workshop*, pages 7850–7859, 2020. 2

[11] Xingxing Zou, Xiangheng Kong, Waikeung Wong, Congde Wang, Yuguang Liu, and Yang Cao. Fashionai: A hierarchical dataset for fashion understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 296–304, 2019. 1