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

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

Figure 1. Comparing SAL-VTON with the recent state-of-the-art methods on the VITON-HD testing dataset (left) and controlling the original try-on results of SAL-VTON via manually manipulating semantically associated landmarks (right).

Abstract

In this paper, a novel virtual try-on algorithm, dubbed SAL-VTON, is proposed, which links the garment with the person via semantically associated landmarks to alleviate misalignment. The semantically associated landmarks are a series of landmark pairs with the same local semantics on the in-shop garment image and the try-on image. Based on the semantically associated landmarks, SAL-VTON effectively models the local semantic association between garment and person, making up for the misalignment in the overall deformation of the garment. The outcome is achieved with a three-stage framework: 1) the semantically associated landmarks are estimated using the landmark localization model; 2) taking the landmarks as input, the warping model explicitly associates the corresponding parts of the garment and person for obtaining the local flow, thus refining the alignment in the global flow; 3) finally, a generator consumes the landmarks to better capture local semantics and control the try-on results. Moreover, we propose a new landmark dataset with a unified labelling rule of landmarks for diverse styles of garments. Extensive experimental results on popular datasets demonstrate that SAL-VTON can handle misalignment and outperform state-of-the-art methods both qualitatively and quantitatively. The dataset is available on https://modelscope.cn/datasets/damo/SAL-HG/summary.

1. Introduction

In recent years, with the rapid popularization of online shopping, virtual try-on [6, 9, 16, 32, 52] has attracted extensive attention for its potential applications. Image-based virtual try-on [4, 34, 49] aims to synthesize a photo-realistic try-on image by transferring a garment image onto the corresponding region of a person. Commonly, there are significant spatial geometric gaps between the in-shop garment image and the person image, leading to garments failing to align the corresponding body parts of person.

To address the above issue, prior arts take geometric deformation models to align the garment with the person's body. Early works [16, 22, 43] widely use the Thin-Plate Spline (TPS) deformation model [39], whereas the smoothness constraint of TPS transformation limits the warping capacity. Recently, the flow operation is applied, with a high dimension of freedom to warp garments [5, 11, 15, 18]. Nonetheless, the flow operation falls short on gar-

*Co-first authors contributed equally, † Corresponding author.

ment regions with large deformation. The aforementioned methods focus on modeling the overall deformation of the garment, but ignore the local semantic association between garment and person. Therefore, when there are large local deformations of garments, the try-on results usually occur misalignment, such as missing or mixing garments (see left part of Fig. 1). To address the local misalignment problem, Xie *et al.* [46] introduce the patch-routed disentanglement module to splice different parts of the garment. However, this method may result in significant blank spaces between spliced parts of the garment.

Fortunately, the landmarks in the garment image and the person image naturally have local semantic associations. As can be observed from Fig. 2, the pixels around landmark A on the try-on result should come from the landmark A' area on the garment. Such a pair of landmarks with the same local semantics are referred to as semantically associated landmarks. Based on this observation, this paper presents a novel virtual try-on algorithm named SAL-VTON, which links the garment with the person via semantically associated landmarks to help align the garment with the person. Notably, the proposed approach varies differently from the previous landmark-guided try-on methods [28,37]. LM-VTON [28] and LG-VTON [37] utilize landmarks to supervise the TPS transformation. However, the potential of local semantic association has not been fully explored, and the limited degrees of freedom of TPS transformation further hinder performance improvements. SAL-VTON, for the first time, introduces the local flow estimated via semantically associated landmarks to effectively model the local semantic association. In addition, a generator with Landmark-Aware Semantic Normalization Layer (LASNL) is carried out to better capture local semantics.

Specifically, the proposed SAL-VTON consists of three stages. Firstly, the semantically associated landmarks are estimated using the landmark localization model. Subsequently, the semantically associated landmarks are employed as a new representation for virtual try-on, and fed into the warping model. Based on the semantically associated landmarks and learnable deformable patches, the warping model explicitly associates the corresponding parts of the garment and person to obtain the local flow, which contributes significantly to refine the poor alignment in the global flow. Finally, conditioned on the landmarks, the LASNL generator can achieve improved alignment in virtual try-on images. The estimated landmarks on the try-on result assist the generator in determining if a specific region needs to generate corresponding garment parts. In this way, SAL-VTON can effectively model the local semantic association between the garment and the person, making up for the misalignment in the overall deformation of the garment. Moreover, the try-on results of SAL-VTON can be precisely controlled by manually manipulating the landmarks (see right part of Fig. 1).

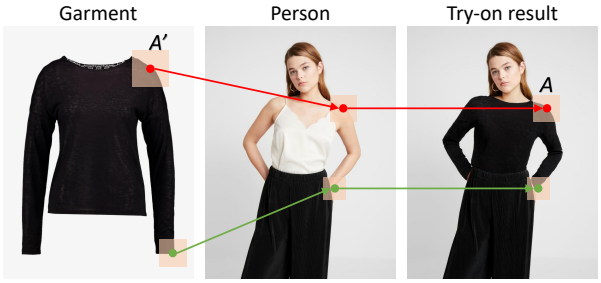

Figure 2. An example for the semantically associated landmarks on the in-shop garment image and the try-on image.

To this end, we re-annotate images on the popular virtual try-on benchmarks including VITON [16] and VITON-HD [4] datasets. Existing popular clothing landmark datasets [12, 54] adopt different landmark definitions for different categories of garments. In contrast to other datasets, we adopt a unified labelling rule of landmarks for diverse styles of garments, including both standard and non-standard varieties. In the proposed dataset [1], every image is annotated with 32 landmarks, each of which possesses three kinds of attributes: visible, occluded and absent. The landmarks with the same serial number have the same semantics, which enhances the universality of the dataset.

This work makes the following main **contributions**: (1) A novel virtual try-on algorithm, SAL-VTON, is proposed, which links the garment with the person via semantically associated landmarks. SAL-VTON, for the first time, introduces the local flow that can alleviate the misalignment and the LASNL generator for virtual try-on. (2) A new landmark dataset is proposed, providing a new representation for virtual try-on, with a unified labelling rule of landmarks for diverse styles of garments. (3) Extensive experiments over two popular datasets demonstrate that SAL-VTON is capable of handling misalignment and significantly outperforms other state-of-the-art methods. Furthermore, the extended experiments show that the virtual try-on results can be edited via the landmarks.

2. Related Work

2.1. Clothing Landmark Localization

Clothing landmark localization [2,3,20] aims at locating the functional key points defined on clothing, such as the corners of the collar, hemline, and cuff. In previous years, clothing landmark datasets, such as DeepFashion [29], FLD [30], and ULD [47], contain very sparse landmarks for upper-body items (only six). For the purpose of dataset enrichment, Ge *et al.* [12] propose DeepFashion2 with an average of 23 defined landmarks in each category. However,

[1]More details about the semantically associated landmark dataset are reported in the supplementary material.

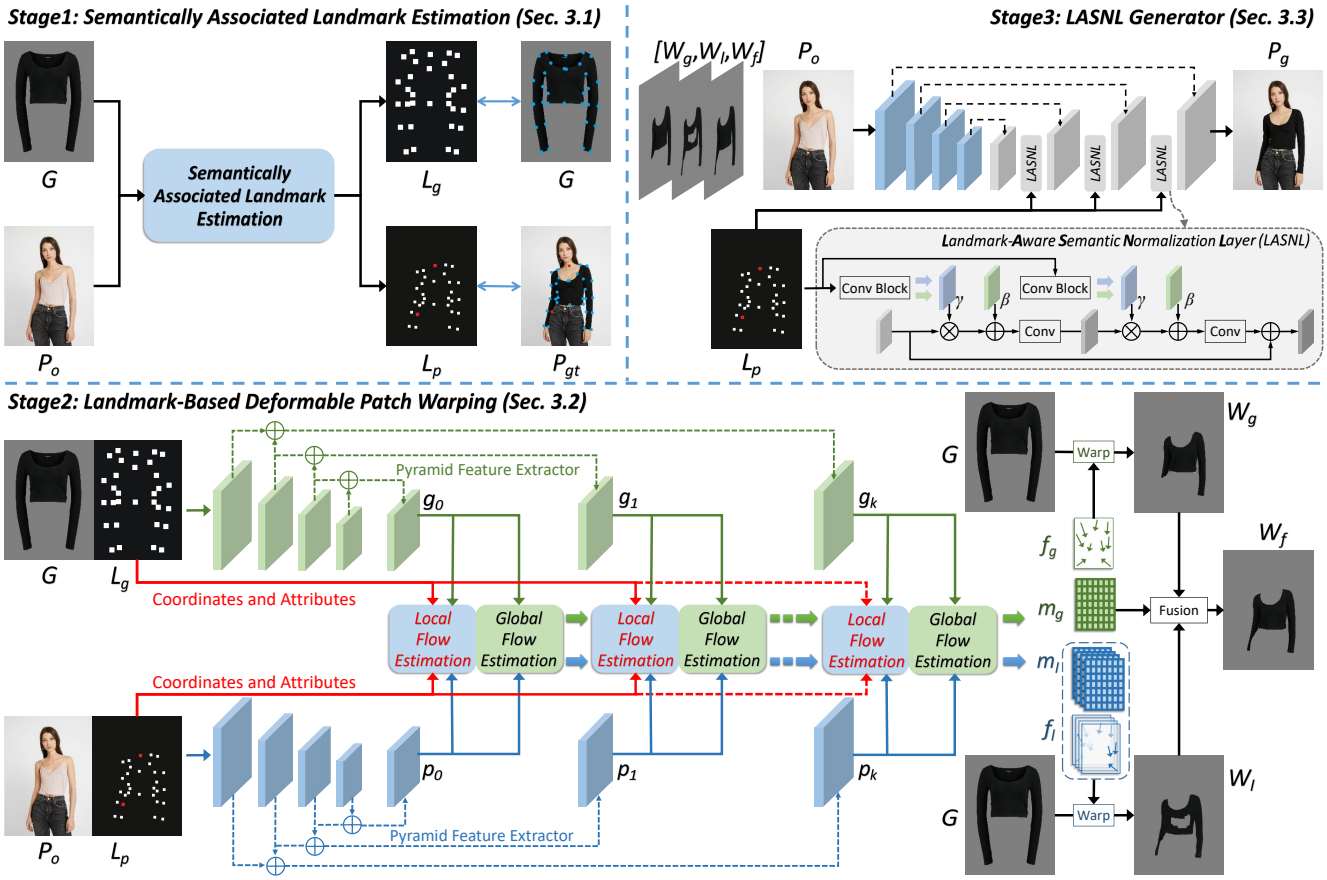

Figure 3. The detailed flowchart of our proposed SAL-VTON framework. SAL-VTON consists of three stages. Stage 1: Semantically associated landmarks are estimated by landmark estimation model. Stage 2: The warping model estimates the global flow and the local flow. The two complement each other and fuse to produce a refined result. Stage 3: LASNL generator synthesizes the final output image.

different categories of garments have different landmark definitions. The number of landmarks on 13 categories of garments also ranges from 8 to 39. In the virtual try-on datasets, there are garment categories that do not exist in DeepFashion2, thus limiting the role of landmarks in the virtual try-on task. In contrast to DeepFashion2, the present study employs a singular definition of landmarks to unify all garment types and subsequently re-annotate virtual try-on datasets. In our dataset, the number of landmarks on all types of garments is 32.

2.2. Virtual Try-on

According to whether the human parser [13, 14] is needed in the inference stage, image-based virtual try-on can be divided into parser-based methods [7, 25, 26, 44] and parser-free methods [11, 18, 21]. The existing available image-based virtual try-on datasets only contain garments and a person wearing the garments. To obtain trainable data pairs, previous methods mainly rely on masking the garment region of the person image, as well as use the human parser as the person representation to construct the trainable data pairs. Consequently, the parser-based methods necessitate the incorporation of the human parser during both training and inference stages. The human parser is usually derived from pre-trained human parser models. To mitigate the negative effects resulting from inaccurate parsing outcomes, the parser-free methods first train a parser-based model, by which the person wearing different garments can be obtained accordingly. Subsequently, the garment and the generated person image serve as inputs and the original person image acts as the supervision to train the parser-free model. Our method is based on the parser-free method and introduces a new representation named semantically associated landmarks.

The spatial transform has an extensive applications in virtual try-on. According to the type of spatial deformation, they can be roughly divided into TPS-based and flow-based methods. TPS-based methods [10, 33, 43, 48] adopt TPS transformation to warp the whole or parts of the garment. However, the smoothness constraint of TPS restricts the warping ability. Currently, a majority of leading algorithms are flow-based methods [8, 18, 26], which predict the global

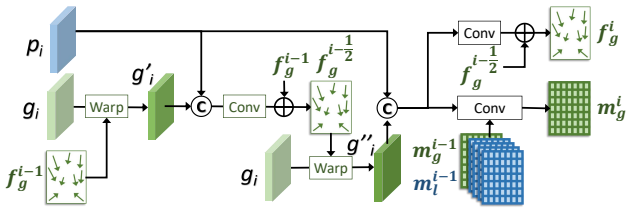

Figure 4. The detail of the global flow estimation module.

flow with a high dimension of freedom to warp the whole garment. Our method also belongs to the flow-based method. However, different from all these existing flow-based methods, our method focuses on designing the local flow via semantically associated landmarks.

3. Proposed Method

Framework Overview. As described in Fig. 3, given an in-shop garment image $G \in \mathbb{R}^{H\times W\times 3}$ and a person wearing other garments $P_o \in \mathbb{R}^{H\times W\times 3}$ (H and W denote the image height and width, respectively), the goal of virtual try-on is to generate a photo-realistic try-on image $P_g \in \mathbb{R}^{H\times W\times 3}$ of the same person P_o wearing the input garment G. Following the strategy adopted by existing parser-free methods [11, 18], parser-based models [18, 26] are first pre-trained on the VITON [16] and VITON-HD [4] datasets respectively, to obtain the rough P_o. Then, the (P_o, G, P_g) triplets are adopted to train the proposed SAL-VTON. The pipeline of SAL-VTON can be roughly divided into three stages. In the first stage, the landmark estimation model is applied to estimate the semantically associated landmarks L_g on the in-shop garment and L_p on the try-on person (Sec. 3.1). In the second stage, the warping model predicts the global flow and the local flow to deform the garment G and align it with the body of the person, in which the local flow is able to refine the poor alignment in the global flow (Sec. 3.2). In the third stage, the LASNL generator is used to synthesize the try-on image P_g (Sec. 3.3).

3.1. Semantically Associated Landmark Estimation

As illustrated in Fig. 3, the HRNet [41] is modified to estimate the semantically associated landmarks. Specifically, the inputs of the modified HRNet consist of two components: the garment G and the person P_o wearing other garments. The outputs are the semantically associated landmarks L_g on the in-shop garment and L_p on the try-on person. The L_p indicates the locations where the landmarks L_g will appear in the final try-on result when the person P_o puts on the garment G. In addition to predicting the locations, the attributes (visible, occluded and absent) of the landmarks are be estimated simultaneously. For training the modified HRNet, we utilize binary CrossEntropy for the landmark heatmap regression and CrossEntropy for the landmark attributes classification.

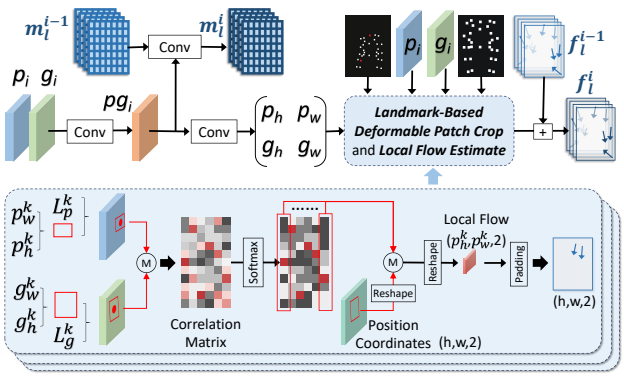

Figure 5. The illustration of the local flow estimation module.

3.2. Landmark-Based Deformable Patch Warping

The same as the previous methods [15, 18], two pyramid feature extractors are adopted to extract the features of two modalities. The garment pyramid feature extractor takes the concatenation of the garment G and landmarks L_g as input, where the landmarks L_g are in the form of a heatmap. The person P_o wearing another garment and the heatmap of landmarks L_p are concatenated to serve as input for the person pyramid feature extractor. The two feature extractors have the identical Feature Pyramid Network (FPN) [27] structure but do not share weights. The FPN network consists of N encoding layers where each layer has a downsampling convolution with a stride of 2, followed by the residual blocks [17]. This study sets $N = 5$, but for simplicity, the case of $N = 4$ is shown in Fig. 3.

Global Flow. The flow used in virtual try-on is a set of 2D coordinate vectors [15, 50, 53]. Each vector indicates which pixels in the garment image G ought to be utilized to fill the given pixel in the person image P_o. Taking Fig. 2 as an example, the pixels around landmark *A'* should be filled into the area around landmark *A*. Through the global flow $f_g \in \mathbb{R}^{h\times w\times 2}$, the garment G is warped to W_g to fit the shape of the person body. As illustrated in 4, the inputs of the global flow estimation module are two pyramid features (p_i, g_i), previous global flow f_g^{i-1}, as well as the global and local attention maps (m_g^{i-1}, m_l^{i-1}) from the previous level. When $i = 0$, there is no previous global flow f_g^{i-1}, thus $g_i' = g_0$. Similar to most recent flow-based methods [11, 18], we first warp the garment feature g_i using the previously obtained global flow f_g^{i-1}. Subsequently, the warped garment feature g_i' and the p_i are concatenated to refine the global flow f_g^{i-1}. The final warped feature g_i'' and the person feature p_i are fed into Conv block to predict f_g^i and the global attention map m_g^i, respectively. Before obtaining m_g^i, this study fuses $[m_g^{i-1}, m_l^{i-1}]$ into Conv block to refine the global attention map, where $[\cdot,\cdot]$ indicates the concatenation operation.

Local Flow. In our case, the local flow $f_l \in \mathbb{R}^{h\times w\times 64}$ contains 32 sets of 2D coordinates, each of which corresponds to the local region near a landmark. However, it is not sufficiently robust to apply a fixed-size region near the landmark due to the geometric gap between the garment and the person. Therefore, learnable deformable patches associated with the semantically associated landmarks are adopted. As shown in Fig. 5, the width and height of deformable patches are obtained as follows:

$$pg_i = C_0([p_i, g_i]), \tag{1}$$

$$p_h, p_w = \sigma(C_1(pg_i)) * I_h, \sigma(C_1(pg_i)) * I_w \tag{2}$$

$$g_h, g_w = \sigma(C_1(pg_i)) * I_h, \sigma(C_1(pg_i)) * I_w, \tag{3}$$

where C_0 and C_1 represent the convolution blocks, $\sigma(\cdot)$ denotes *Sigmoid* function, and I_h and I_w are initial height and width. We initialize I_h and I_w to a fifth of the size of the image. Let (p_w, p_h, L_p) and (g_w, g_h, L_g) represent the learned deformable patches on the person and garment, respectively. The centres of deformable patches are the semantically associated landmarks L_p and L_g.

In Fig. 5, one pair of deformable patches (p_w^k, p_h^k, L_p^k) and (g_w^k, g_h^k, L_g^k) having the same local semantics are taken as an instance, where $k \in \{1, \ldots, 32\}$. Based on the deformable patches, we crop the features of person and garment respectively. Then, the correlation matrix is obtained by matrix multiplication of the two cropped features, which performs pixel-by-pixel matching of features. Through the correlation matrix, we can search for the correlative position coordinates on the garment as the 2D coordinates of the local flow. The obtained local flow has more explicit constraints. In this way, the local region pixels of the garment will fill the same semantics local region of the person via the local flow $f_l^i \in \mathbb{R}^{h\times w\times 64}$. In addition, the local flow estimation module predicts the local attention map $m_l^i \in \mathbb{R}^{h\times w\times 32}$. The local flow f_l and the attention map m_l are simultaneously applied to the garment G:

$$W_l = \sum_{k=1}^{K} \frac{\exp(m_{lk})}{\sum_{j=1}^{K} \exp(m_{lj})} \mathcal{W}(G, f_{lk}), \tag{4}$$

where $\mathcal{W}(\cdot,\cdot)$ denotes the warping operation, W_l denotes the warped garment with the local flow and K is 32.

Fusion. The global flow deforms the garment as a whole, however, there exist certain local regions that exhibit misalignment, such as missing and mixing garments. The local flow focuses on the specific local association between the garment and the person. The two flow approaches complement each other and fuse to obtain a new result W_f:

$$W_f = (1 - \sigma(m_g)) * W_l + \sigma(m_g) * W_g. \tag{5}$$

As shown in Fig. 6, the proposed approach helps to alleviate the misalignment and improve the try-on result.

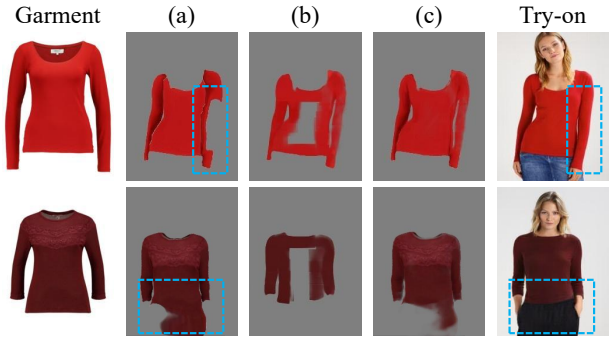

Figure 6. Fusing the warping results from the global flow and the local flow. (a) the warped garment from the global flow, (b) the local warping result from the local flow, and (c) the fusion result.

In the first row, the local flow fills in the missing garment at the elbow position in the global flow. In the second row, the local flow helps the global flow eliminate the overflow in the hemline to avoid garment mixing.

To train our warping model, we apply the perceptual loss [23] and L1 loss:

$$\mathcal{L}_p = \sum_i \|\phi_i(W_g) - \phi_i(W_{gt})\| + \sum_i \|\phi_i(W_f) - \phi_i(W_{gt})\|, \tag{6}$$

$$\mathcal{L}_{L1} = \|W_g - W_{gt}\| + \|W_f - W_{gt}\|, \tag{7}$$

where ϕ_i is the i-th block of the pre-trained VGG network [38] and W_{gt} is the garment on the ground truth P_{gt}. We also apply a smoothness regularization on the global flow from each global flow estimation module:

$$\mathcal{L}_R = \sum_i \|\nabla f_g\|, \tag{8}$$

where $\|\nabla f_g\|$ is the generalized charbonnier loss function [40]. To supervise the training of the local flow, we apply a loss on the warped garment W_l:

$$L_p = \|(W_l - W_{gt}) * M_l\|, \tag{9}$$

where M_l denotes the landmark mask with one channel, which is the union of 32 hard landmark masks. A hard landmark mask is similar to a landmark heatmap but can take 0 and 1 values only. The centre coordinates are the landmark coordinates on the ground truth person P_{gt}.

3.3. Try-On via Landmark Semantic Normalization

The generator aims to generate the final try-on image P_g based on the outputs from the previous stages. In general, we fuse the warped garments (W_g, W_l, W_f) with the person P_o, guided by the heatmap of landmark L_p. For L_p, we propose the Landmark-Aware Semantic Normalization Layer, LASNL for short. LASNL introduces prior knowledge to the generator, which enables the generator to determine whether the local regions of the person P_o need to wear the garment, and further alleviate the local misalignment.

Figure 7. Qualitative results from different models (VITON-HD, HR-VITON and ours) on the VITON-HD testing dataset.

Landmark-Aware Semantic Normalization Layer. Let $\mathbf{F}^i \in \mathbb{R}^{N\times C_i\times H_i\times W_i}$ be a i-th feature map in the network for a batch of N samples, where C_i is the number of channels. Here we use $\mathbf{L}^i \in \mathbb{R}^{N\times 32\times H_i\times W_i}$ to denote the heatmap of landmark L_p. First, we obtain the modulation parameters $\gamma^i \in \mathbb{R}^{N\times C_i\times H_i\times W_i}$ and $\beta^i \in \mathbb{R}^{N\times C_i\times H_i\times W_i}$:

$$\mathbf{L}^i_s = C_{com}(\mathbf{L}^i), \tag{10}$$

$$\gamma^i, \beta^i = C_\gamma(\mathbf{L}^i_s), C_\beta(\mathbf{L}^i_s), \tag{11}$$

where C_{com}, C_γ and C_β represent the convolution layers. Then, the process of normalization can be expressed as:

$$\gamma^i_{k,y,x} * \frac{\mathbf{F}^i_{n,k,y,x} - \mu^i_{n,k}}{\sigma^i_{n,k}} + \beta^i_{k,y,x}, \tag{12}$$

where $n \in \{1,\ldots,N\}$, $k \in \{1,\ldots,C_i\}$ and (y,x) is the pixel index. $\mu^i_{n,k}$ and $\sigma^i_{n,k}$ are the mean and standard deviation of the sample n and channel k.

LASNL Generator. Following the designs in prior works [11, 18], the LASNL generator adopt an encoder-decoder architecture [36] with skip connections. In the decoder, we replace the residual block with the proposed landmark-aware semantic normalization layer. Experiments demonstrate that the performance of try-on can be improved in this manner. To train the LASNL generator, we follow [26] to calculate the perceptual loss, feature matching loss and adversarial loss for the output P_g and the ground truth P_{gt}.

4. Experiments

Datasets. Experiments are conducted on the most popular datasets: VITON [16] and VITON-HD [4]. The VITON dataset contains a training set of 14,221 image pairs and a testing dataset of 2,032 pairs. All images in the VITON dataset have a resolution of 256×192. The VITON-HD is a high-resolution virtual try-on dataset. The resolution of the images in VITON-HD dataset is 1024×768. The VITON-HD dataset contains a training set of 11,647 image pairs and a testing dataset of 2,032 pairs.

Implementation details. All the experiments are conducted using Pytorch [35]. Adam [24] is adopted as the optimizer with $\beta 1 = 0.5$, $\beta 2 = 0.999$. We set the batch size to 8 and train the models with 100 epochs. The initial learning rate of the semantically associated landmark estimation model and the warping model is set to 3e-4 and decays linearly after 50 epochs. The initial learning rates of the generator and the discriminator of the try-on image generator are set to 1e-4 and 4e-4, respectively.

Evaluation metrics and baselines. In order to make a fair comparison, we follow the settings of recent works [18, 26] to compare the performance of SAL-VTON with other baseline models. Specifically, our method is evaluated on the VITON-HD dataset using five widely used metrics, that is, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [45], Learned Perceptual Image Patch Similarity (LPIPS) [51], Fréchet Inception Distance (FID) [19, 42] and Kernel Inception Distance(KID) [1]. For the VITON dataset, three widely used metrics, namely PSNR, SSIM and FID, are adopted.

To verify the effectiveness of our proposed method, we report the performance comparisons between ours and representative virtual try-on methods, covering CP-VTON [43], ClothFlow [15], CP-VTON+ [31], ACGPN [48], LM-VTON [28], PF-AFN [11], ZFlow [5], StyleFlow [18], VITON-HD [4] and HR-VITON [26], with the last two being implemented on the high-resolution dataset.

Main results. The quantitative results on the VITON-HD dataset are shown in Table 1. Following previous studies, SSIM, PSNR and LPIPS are applied to evaluate our method for paired setting. FID and KID are uesd for the unpaired setting. It is clearly found that our method surpasses other comparative algorithms in all evaluation metrics. Specifically, our method achieves a 1.24 dB improvement in PSNR and decreases LPIPS, FID and KID by 26.2%, 12.7% and 43.0% than the second-best results. We show the qualitative comparison of the visual results in Fig. 7 to

Figure 8. Qualitative results from different models (CP-VTON+, ACGPN, PF-AFN, StyleFlow and ours) on the VITON testing dataset.

Table 1. The quantitative results on VITON-HD test datasets. The best values are highlighted by the black **bold**. The ↑ or ↓ indicates higher or lower metric corresponds to better results.

Methods	LPIPS↓	SSIM↑	PSNR↑	FID↓	KID↓
CP-VTON	0.158	0.786	_	43.28	3.762
ACGPN	0.112	0.850	_	43.29	3.730
VITON-HD	0.077	0.873	20.82	11.59	0.247
HR-VITON	0.065	0.892	21.90	10.91	0.179
Ours	**0.048**	**0.907**	**23.14**	**9.52**	**0.102**

verify the effectiveness of our method over the VITON-HD dataset. In the first row of Fig. 7, the images generated by other competing methods show apparent misalignments such as long black sleeves mixed with skirts and missing garments on the arm and waist. Even in the case of physical occlusion, SAL-VTON can generate more photo-realistic images compared to other methods.

The quantitative results on VITON testing dataset are shown in Table 2. As can be seen clearly, our proposed SAL-VTON achieves the best overall results than other comparative try-on methods. The qualitative results from different models are illustrated in Fig. 8. Overall, our method generates better try-on images with less misalignment, especially in the local regions with large geometric distortions[2]. For example, in the hard poses on the right side of Fig. 8, the other methods failed to align the garment to the elbow and forearm. Benefiting from semantically associated landmarks, our approach explicitly links the sleeve of the garment with the arm of the person, addressing the misalignment.

User study. Image metrics may have limitations in depicting the try-on quality, so a user study is conducted by inviting 50 participants from professional data annotation companies. Compared with ordinary volunteers, the

Table 2. The quantitative results on VITON test datasets. The best values are highlighted by the black **bold**. The ↑ or ↓ indicates higher or lower metric corresponds to better results.

Methods	SSIM↑	PSNR↑	FID↓
CP-VTON	0.78	21.01	30.50
ClothFlow	0.84	23.60	23.68
CP-VTON+	0.82	21.79	21.08
ACGPN	0.85	23.08	16.46
LM-VTON	0.85	21.55	17.18
PF-AFN	0.86	22.65	10.09
ZFlow	0.89	25.64	15.17
StyleFlow	0.91	25.87	8.89
Ours	**0.92**	**28.29**	**5.74**

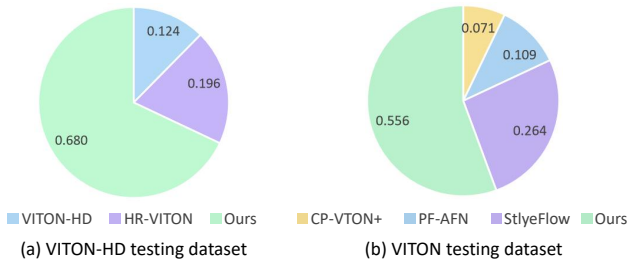

Figure 9. The user study on the VITON-HD and VITON datasets.

expertise background and experience of these participants could further assurance the reliability of the assessment results. The number of randomly selected image groups are increased to 1,000 for more objective and accurate evaluation results. Meanwhile, each image group are observed and evaluated by three different participants. From user study results in Fig. 9, SAL-VTON outperforms the existing state-of-the-art methods compellingly in both VITON dataset and VITON-HD dataset. The misaligned regions become noticeable as the resolution of the image

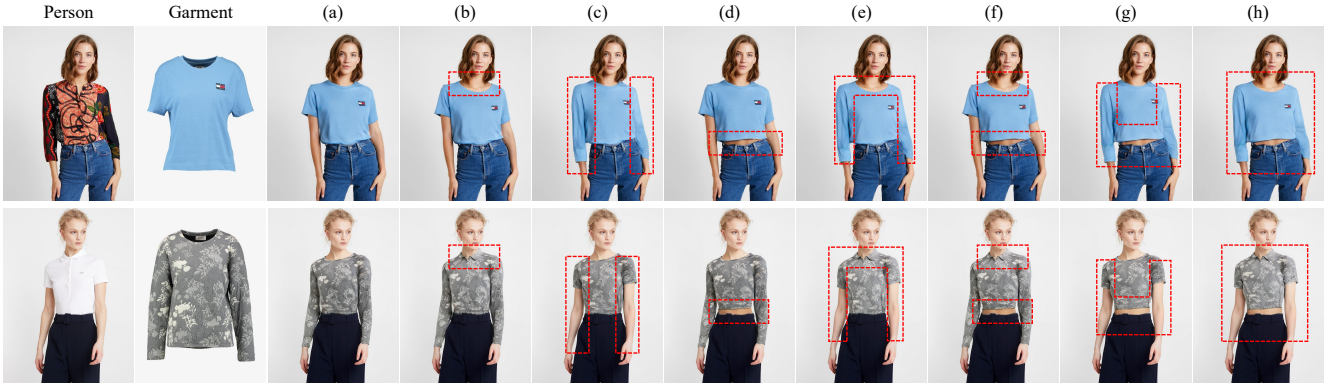

Figure 10. Controlling the try-on effect of SAL-VTON by manipulating the landmarks. (a) the original try-on result, (b) changing the shape of the collar, (c) controlling the sleeve length, (d) adjusting the hemline, (e) controlling both collar and sleeve, (f) controlling both collar and hemline, (g) controlling both sleeve and hemline, (h) controlling sleeve, collar and hemline.

Table 3. The results of ablation experiments over VITON dataset.

Config	SSIM↑	PSNR↑	FID↓
w/ Fixed Patch	0.90	28.03	5.91
w/ Deformable Patch	**0.92**	**28.29**	**5.74**
w/o LASNL	0.90	27.66	5.86
w/o Local Flow and LASNL	0.88	25.81	6.99

increases, as revealed by the study of Choi *et al.* [4]. From the results of the user study, our approach is more advantageous on the high-resolution dataset, which means our approach can handle misalignment.

Ablation study. To investigate the contribution of the devised modules in our proposed algorithm, an ablation study is carried out on the VITON dataset. The corresponding quantitative comparison is reported in Table 3. As can be observed, the use of deformable patches yields better results than fixed patches. Without local flow refinement and landmark-aware semantic normalization layer, there is a significant performance degradation for our approach. Overall, the best performance is reported when all modules are integrated together. Additionally, to better understand the deformable patches based on the semantically associated landmarks, we visualize a portion of them in Fig. 11. It is clear that SAL-VTON can adaptively adjust the size of patches for different parts of the garment. Compared with the fixed patches, the proposed deformable patches are more robust to spatial geometric changes.

Extended experiment. In addition to experimenting with the regular virtual try-on, an interesting exploration is also conducted on whether the try-on results can be directly controlled by manually manipulating the landmarks, such as by changing the coordinate value of some landmarks or their attributes. It can be observed from Fig. 10 that manipulating landmarks can achieve garment editing effects, such as changing the shape of the collar, the sleeve length of garments, etc. All the manipulations (a)-(c) about the landmarks can be combined to obtain the combined results (e)-(f) in Fig. 10. This interesting finding demonstrates the validity of the semantically associated landmark, which plays an essential role in our model[2].

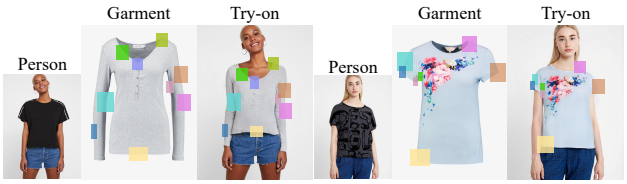

Figure 11. Visualization of the deformable patches predicted by SAL-VTON. The corresponding deformable patches on the garment and the try-on result are shown in the same color.

5. Conclusion

In this paper, we present a novel virtual try-on algorithm, named SAL-VTON, that links the garment with the person via semantically associated landmarks to alleviate misalignment. The proposed local flow properly handles the misaligned local regions. The LASNL generator can better capture local semantics to generate photo-realistic try-on images. Extensive experiments show that our proposed method not only achieves superior performance but can also be extended to try-on image editing. The encouraging results of SAL-VTON will inspire computer vision and computer graphics researchers to explore more effective methods for modelling the local semantic association for virtual try-on.

Acknowledgment. This work was supported by Alibaba Group through Alibaba Research Intern Program.

[2]More examples can be found in the supplemental materials.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[2] Ming Chen, Yingjie Qin, Lizhe Qi, and Yunquan Sun. Improving fashion landmark detection by dual attention feature enhancement. In *IEEE/CVF International Conference on Computer Vision Workshop*, pages 3101–3104, 2019. 2

[3] Ming Chen, Yingjie Qin, Lizhe Qi, and Yunquan Sun. Improving fashion landmark detection by dual attention feature enhancement. In *IEEE/CVF International Conference on Computer Vision Workshop*, pages 3101–3104, 2019. 2

[4] 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, 4, 6, 8

[5] 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. 1, 6

[6] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14638–14647, 2021. 1

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035, 2019. 3

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1161–1170, 2019. 3

[9] Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu, Xiaobo Li, Kairi Ou, Deli Zhao, and Zheng-Jun Zha. Weakly supervised high-fidelity clothing model generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3440–3449, 2022. 1

[10] 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. 3

[11] 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. 1, 3, 4, 6

[12] 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. 2

[13] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 3

[15] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 1, 4, 6

[16] 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, 2, 4, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[18] 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. 1, 3, 4, 6

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[20] Chang-Qin Huang, Ji-Kai Chen, Yan Pan, Han-Jiang Lai, Jian Yin, and Qiong-Hao Huang. Clothing landmark detection using deep networks with prior of key point associations. *IEEE Transactions on Cybernetics*, 49(10):3744–3754, 2019. 2

[21] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision*, pages 619–635. Springer, 2020. 3

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015. 1

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Hyug Jae Lee, Rokkyu Lee, Minseok Kang, Myounghoon Cho, and Gunhan Park. La-viton: A network for looking-attractive virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[26] 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. 3, 4, 6

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[28] Guoqiang Liu, Dan Song, Ruofeng Tong, and Min Tang. Toward realistic virtual try-on through landmark guided shape matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2118–2126, 2021. 2, 6

[29] 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. 2

[30] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016. 2

[31] 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. 6

[32] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020. 1

[33] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. 3

[34] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192, 2020. 1

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 6

[37] Debapriya Roy, Sanchayan Santra, and Bhabatosh Chanda. Lgvton: a landmark guided approach for model to person virtual try-on. *Multimedia Tools and Applications*, 81(4):5051–5087, 2022. 2

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[39] R. Sprengel, K. Rohr, and H.S. Stiehl. Thin-plate spline approximation for image registration. In *Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 1190–1191 vol.3, 1996. 1

[40] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014. 5

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019. 4

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 6

[43] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *European Conference on Computer Vision*, pages 589–604, 2018. 1, 3, 6

[44] Jiahang Wang, Tong Sha, Wei Zhang, Zhoujun Li, and Tao Mei. Down to the last detail: Virtual try-on with fine-grained details. In *Proceedings of International Conference on Multimedia*, pages 466–474, 2020. 3

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[46] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021. 2

[47] 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. 2

[48] 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. 3, 6

[49] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10510–10519, 2019. 1

[50] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow.

In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 4

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[52] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021. 1

[53] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301. Springer, 2016. 4

[54] 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. 2