Style-Based Global Appearance Flow for Virtual Try-On

Sen He, Yi-Zhe Song, Tao Xiang
Center for Vision, Speech and Signal Processing, University of Surrey
iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
{sen.he,y.song,t.xiang}@surrey.ac.uk

Figure 1. Our global appearance flow based try-on model has a clear advantage over existing local flow based SOTA methods such as Cloth-flow [13] and PF-AFN [10], especially when there are large mis-alignment between reference and garment images (top row), and difficult poses/occlusions (bottom row).

Abstract

Image-based virtual try-on aims to fit an in-shop garment into a clothed person image. To achieve this, a key step is garment warping which spatially aligns the target garment with the corresponding body parts in the person image. Prior methods typically adopt a local appearance flow estimation model. They are thus intrinsically susceptible to difficult body poses/occlusions and large mis-alignments between person and garment images (see Fig. 1). To overcome this limitation, a novel global appearance flow estimation model is proposed in this work. For the first time, a StyleGAN based architecture is adopted for appearance flow estimation. This enables us to take advantage of a global style vector to encode a whole-image context to cope with the aforementioned challenges. To guide the StyleGAN flow generator to pay more attention to local garment deformation, a flow refinement module is introduced to add local context. Experiment results on a popular virtual try-on benchmark show that our method achieves new state-of-the-art performance. It is particularly effective in a 'in-the-wild' application scenario where the reference image is full-body resulting in a large mis-alignment with the garment image (Fig. 1 Top). Code is available at: https://github.com/SenHe/Flow-Style-VTON.

1. Introduction

The transition from offline in-shop retail to e-commerce has been accelerated by the recent pandemic caused lock downs. In 2020, retail e-commerce sales worldwide amounted to 4.28 trillion US dollars and e-retail revenues are projected to grow to 5.4 trillion US dollars in 2022. However, when it comes to fashion, one of key offline experiences missed by the on-line shoppers is the changing room where a garment item can be tried-on. To reduce the return cost for the online retailers and give shoppers the same offline experience online, image-based virtual try-on (VTON) has been studied intensively recently [9, 10, 13, 14, 19, 24, 38, 39, 42, 43].

A VTON model aims to fit an in-shop garment into a person image. A key objective of a VTON model is to align the in-shop garment with the corresponding body parts in the person image. This is due to the fact that the in-shop garment is usually not spatially aligned with the person image (see Fig. 1). Without the spatial alignment, directly applying advanced detail-preserving image to image translation models [18, 30] to fuse the texture in person image and garment image will result in unrealistic effect in the generated try-on image, especially in the occluded and misaligned regions.

Previous methods address this alignment problem through garment warping, i.e., they first warp the in-shop garment, which is then concatenated with the person image and fed into an image to image translation model for the final try-on image generation. Many of them [9,14,19,38,42,43] adopt a Thin Plate Spline (TPS) [7] based on the warping method, exploiting the correlation between features extracted from the person and garment images. However, as analyzed in previous works [5,13,42], TPS has limitations in handling complex warping, e.g., when different regions in the garment require different deformations. As a result, recent SOTA methods [10,13] estimate dense appearance flow [45] to warp the garment. This involves training a network to predict the dense appearance flow field representing the deformation required to align the garment with the corresponding body parts.

However, existing appearance flow estimation methods are limited in accurate garment warping due to the lack of global context. More specifically, all existing methods are based on local feature's correspondence, e.g., local feature concatenation or correlation[1], developed for optical flow estimation [6,17]. To estimate the appearance flow, they make the unrealistic assumption that the corresponding regions from the person image and the in-shop garment are located in the same local receptive filed of the feature extractor. When there is a large mis-alignment between the garment and corresponding body parts (Fig. 1 Top), current appearance flow based methods will deteriorate drastically and generate unsatisfactory results. Lacking a global context also make existing flow-based VTON methods vulnerable to difficult poses/occlusions (Fig. 1 Bottom) when correspondences have to be searched beyond a local neighborhood. This severely limits the use of these methods 'in-the-wild', whereby a user may have a full-body picture of herself/himself as the person image to try-on multiple garment items (e.g., top, bottom, and shoes).

To overcome this limitation, a novel global appearance flow estimation model is proposed in this work. Specifically, for the first time, a StyleGAN [21,22] architecture for dense appearance flow estimation. This differs fundamentally from existing methods [6,10,13,17] which employ a U-Net [30] architecture to preserve local spatial context. Using a global style vector extracted from the whole reference and garment images makes it easy for our model to capture global context. However, it also raises an important question: can it capture local spatial context crucial for local alignments? After all, a single style vector seemingly has lost local spatial context. To answer this question, we first note that StyleGAN has been successfully applied to local face image manipulation tasks, where different style vectors can generate the same face at different viewpoints [34] and different shapes [15,28]. This suggests that a global style vector does have local spatial context encoded. However, we also note that the vanilla StyleGAN architecture [21,22], though much more robust against large mis-alignment and difficult poses/occlusions compared to U-Net, is weaker when it comes to local deformation modeling. We therefore introduce a local flow refinement module in the existing StyleGAN generator to have the better of both worlds.

Concretely, our StyleGAN-based warping module ($\mathcal{W}$ in Fig. 2) consists of stacked warping blocks that takes as inputs a global style vector, garment features and person features. The global style vector is computed from the lowest resolution feature maps of the person image and the in-shop garment for global context modeling. In each warping block in the generator, the global style vector is used to modulate the feature channels which takes in the corresponding garment feature map to estimate the appearance flow. To enable our flow-estimator to model the fine-grained local appearance flow, e.g., the arm and hand regions in Fig. 5, in each warping block on top of the style based appearance flow estimation part, we introduce a refinement layer. This refinement layer first warps the garment feature map, which is subsequently concatenated with the person feature map at the same resolution and then used to predict the local detailed appearance flow.

The contributions of this work are as follow: (1) We propose a novel style-based appearance flow method to warp the garment in virtual try-on. This global flow estimation approach makes our VTON model much robust against large mis-alignments between person and garment images. This makes our method more applicable to 'in-the-wild' application where a full-body person image with natural poses is used (see in Fig. 1). (2) We conduct extensive experiments to validate our method, demonstrating clearly that it is superior to existing state-of-the-art alternatives.

2. Related Work

Image based virtual try-on Image based (2D) VTON can be categorized into parser-based methods and parser-free methods. Their main difference is whether an off-the-shelf human parser[2] is required in the inference stage.

Parser-based methods apply a human segmentation map to mask the garment region in the input person image for warping parameter estimation. The masked person image is concatenated with the warped garment and then fed into a generator for target try-on image generation. Most methods [9,13,14,38,42,43] apply a pre-trained human parser [11]

---

[1]It is worth noting that the tensor correlation methods [6,10,17] have the potential to reach global receptive field. However, its computation grows quadratically with respect to the input size. To make it tractable, its actual implementation is still based on limited local neighborhoods.

[2]Sometimes, pre-trained pose [3] and densePose [12] detection models are also used in a parser based model.

to parse the person image into several pre-defined semantic regions, e.g., head, top, and pants. For better try-on image generation, [42] also transforms the segmentation map to match the target garment. The transformed parsing result, together with the warped garment and the masked person image are used for final try-on image generation. The reliance on a parser make these methods sensitive to bad human parsing results [10, 19] which inevitably lead to inaccurate warping and try-on results.

In contrast, parser-free methods [10, 19], in the inference stage, only takes as inputs the person image the garment image. They are designed specifically to eliminate the negative effects induced by the bad parsing results. Those methods usually first train a parser-based teacher model and then distill a parser-free student model. [19] proposed a pipeline which distills the garment warping module and try-on generation network using paired triplets. [10] further improved [19] by introducing cycle-consistency for better distillation.

Our method is also a parser free method. However, our method focuses on the design of the garment warping part, where we propose a novel global appearance flow based garment warping module.

3D virtual try-on Compared to image based VTON, 3D VTON provides better try-on experience (e.g., allowing being viewed with arbitrary views and poses), yet is also more challenging. Most 3D VTON works [2, 27] rely on 3D parametric human body models [25] and need scanned 3D datasets for training. Collecting large scale 3D datasets is expensive and laborious, thus posing a constraint on the scalability of a 3D VTON model. To overcome this problem, recently [44] applied non-parametric dual human depth model [8] for monocular to 3D VTON. However, existing 3D VTON still generate inferior texture details compared to the 2D methods.

StyleGAN for image manipulation StyleGAN [21, 22] has revolutionized the research on image manipulation [28, 33, 41] lately. Its successful application on the image manipulation tasks often thanks to its suitability in learning a highly disentangled latent space. Recent efforts have been focused on unsupervised latent semantics discovery [4, 34, 37]. [24] applied pose conditioned StyleGAN for virtual try-on. However, their model cannot preserve garment details and is slow during inference.

The design of our garment warping network is inspired from StyleGAN in image manipulation, especially its super performance in shape deformation [28, 34]. Instead of using style modulation to generate the warped garment, we use style modulation to predict the implicit appearance flow which is then used to warp the garment via sampling. This design is much more suited to garment detail-preserving compared to [24].

Appearance flow In the context of VTON, appearance flow was first introduced by [13]. Since then, it has gained more attention and adopted by recent state-of-the-art VTON models [5, 10]. Fundamentally, appearance flow is used as a sampling grid for garment warping, it is thus information lossless and superior in detail preserving. Beyond VTON, appearance flow is also popular in other tasks. [45] applied it for novel view synthesis. [1, 29] also applied the idea of appearance flow to warp the feature map for person pose transfer. Different from all these existing appearance flow estimation methods, our method, via style modulation, applies a global style vector to estimate the appearance flow. Our method is thus intrinsically superior in its ability to coping with large mis-alignments.

3. Methodology

3.1. Problem definition

Given a person image ($p \in \mathbb{R}^{3 \times H \times W}$) and an in-shop garment image ($g \in \mathbb{R}^{3 \times H \times W}$), the goal of virtual try-on is to generate a try-on image ($t \in \mathbb{R}^{3 \times H \times W}$) where the garment in g fits to the corresponding parts in p. In addition, in the generated t, both details from g and non-garment regions in p should be preserved. In other words, the same person in p should appear unchanged in t except now wearing g.

To eliminate the negative effect of inaccurate human parsing, our proposed model ($\mathcal{F}$ in Fig. 2) is designed to be a parser-free model. Following the strategy adopted by existing parser-free models [10, 19], we first pre-train a parser-based model ($\mathcal{F}^{PB}$). It is then used as a teacher for knowledge distillation to help train the final parser-free model $\mathcal{F}$. Both $\mathcal{F}$ and $\mathcal{F}^{PB}$ consist of three parts, i.e., two feature extractors ($\mathcal{E}_p^{PB}$, $\mathcal{E}_g^{PB}$ in $\mathcal{F}^{PB}$ and $\mathcal{E}_p$, $\mathcal{E}_g$ in $\mathcal{F}$), warping module ($\mathcal{W}^{PB}$ in $\mathcal{F}^{PB}$ and $\mathcal{W}$ in $\mathcal{F}$), and a generator ($\mathcal{G}^{PB}$ in $\mathcal{F}^{PB}$ and $\mathcal{G}$ in $\mathcal{F}$). Each of them will be detailed in the following sections.

3.2. Pre-training a parser-based model

As per standard in existing parser-free models [10, 19], a parser-based model $\mathcal{F}^{PB}$ is first trained. It is used in two ways in the subsequent training of the proposed parser-free model $\mathcal{F}$: (a) to generate person image (p) to be used by $\mathcal{F}$ as input and (b) to supervise the training of $\mathcal{F}$ via knowledge distillation.

Concretely, $\mathcal{F}_{PB}$ takes as inputs the semantic representation (segmentation map[3], keypoint pose and dense pose) of a real person image ($p_{gt} \in \mathbb{R}^{3 \times H \times W}$) in the training set and an unpaired garment ($g_{un} \in \mathbb{R}^{3 \times H \times W}$). The output of $\mathcal{F}_{PB}$ is the image p where the original person is wearing g_{un}. p will serve as the input for $\mathcal{F}$ during training. This design, according to [10], benefits from the fact that we now

---

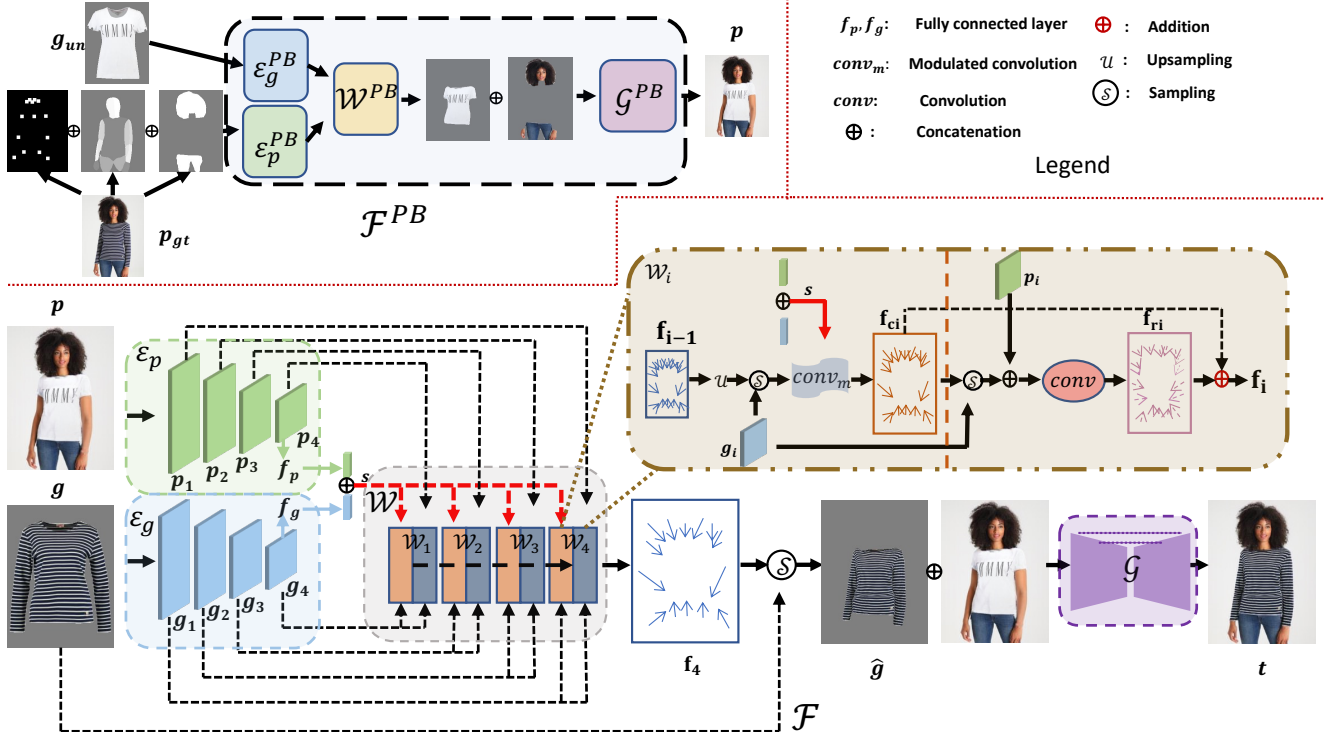[3]The garment region in the segmentation map is flipped as background region

Figure 2. A schematic of our framework. The pre-trained parser based model $\mathcal{F}^{PB}$ generates an output image as the input of parser free model $\mathcal{F}$. The two feature extractors in $\mathcal{F}$ extract the feature of person image and garment image, respectively. A style vector is extracted from the lowest resolution feature maps from person image and the garment image. The warping module takes in the style vector and feature maps from the person image and garment image, and output an appearance flow map. The appearance flow is then used to warp the garment. Finally, the warped garment is concatenated with person image and fed into the generator to generate the target try-on image. Note that $\mathcal{F}_{PB}$ is only used during training.

have paired person image p_{gt} and garment image g in p_{gt} to train the parser-free model $\mathcal{F}$, that is:

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \|t - p_{gt}\|, \qquad (1)$$

where $t = \mathcal{F}(p, g)$ is the generated try-on image from $\mathcal{F}$. Note that $\mathcal{F}^{PB}$ is only used during the training of $\mathcal{F}$.

3.3. Feature extraction

We apply two convolutional encoders ($\mathcal{E}_p$ and $\mathcal{E}_g$) to extract the features of p and g. Both $\mathcal{E}_p$ and $\mathcal{E}_g$ share the same architecture, composed of stacked residual blocks. The extracted features from $\mathcal{E}_p$ and $\mathcal{E}_g$ can be represented as $\{p_i\}_1^N$ and $\{g_i\}_1^N$ ($N = 4$ in Fig. 2 for simplicity), where $p_i \in \mathbb{R}^{c_i \times h_i \times w_i}$ and $g_i \in \mathbb{R}^{c_i \times h_i \times w_i}$ are the feature maps extracted from the corresponding residual block in $\mathcal{E}_p$ and $\mathcal{E}_g$, respectively. The extracted feature maps will be used in $\mathcal{W}$ to predict the appearance flow.

3.4. Style based appearance flow estimation

The main novel component of the proposed model is a style-based global appearance flow estimation module. Different from previous methods that estimate appearance flow

based on local feature correspondence [10, 13], originally proposed in optical flow estimation [6, 17], our method, based on a global style vector, first estimates a coarse appearance flow via style modulation and then refine the predicted coarse appearance flow based on local feature correspondence.

As illustrated in Fig. 2, our warping module ($\mathcal{W}$) consists of N stacked warping blocks ($\{\mathcal{W}_i\}_1^N$), each block is composed of a style-based appearance flow prediction layer (orange rectangle) and a local correspondence based appearance flow refinement layer (blue rectangle). Concretely, we first extract a global style vector ($s \in \mathbb{R}^c$) using the features output from the Nth (final) blocks of $\mathcal{E}_p$ and $\mathcal{E}_g$, denoted as p_N and g_N, as:

$$s = [f_p(p_N), f_g(g_N)], \qquad (2)$$

where f_p and f_g are fully connected layers, and $[\cdot, \cdot]$ denotes concatenation. Intrinsically, the extracted global style vector s[4] contains the global information of the person and garment, e.g., position, structure, etc. Similar to style based image manipulation [15, 28, 33, 34], we expect the global

---

[4]Intuitively, $s = f_p(p_N)$ is enough to generate the appearance flow. But we empirically found that $s = [f_p(p_N), f_g(g_N)]$ yields better results.

style vector s capture the required deformation for warping g into p. It is thus used for style modulation in a StyleGAN style generator for estimating a appearance flow field.

More specifically, in the style-based appearance flow prediction layer of each block $\mathcal{W}_i$, we apply style modulation to predict a coarse flow:

$$\mathbf{f_{ci}} = conv_m(\mathcal{S}(g_{N+1-i}, \mathcal{U}(\mathbf{f_{i-1}})), s), \qquad (3)$$

where $conv_m$ denotes modulated convolution [21], $\mathcal{S}(\cdot, \cdot)$ is the sampling operator, $\mathcal{U}$ is the upsampling operator, and $\mathbf{f_{i-1}} \in \mathbb{R}^{2 \times h_{i-1} \times w_{i-1}}$ is the predicted flow from last warping block. Note that the first block $\mathcal{W}_1$ in $\mathcal{W}$ only takes in the lowest resolution garment feature map and the style vector, i.e., $\mathbf{f_{c1}} = conv_m(g_N, s)$. As can be seen from Equation 3, the predicted $\mathbf{f_{ci}}$ depends on the garment feature map and the global style vector. It thus has a global receptive field and is capable to cope with large mis-alignments between the garment and person images. However, as the style vector s is a global representation, as a trade-off, it has a limited ability to accurately estimate the local fine-grained appearance flow (as shown in Fig. 5). The coarse flow is thus in need of a local refinement.

To refine $\mathbf{f_{ci}}$, we introduce a local correspondence based appearance flow refinement layer in each block $\mathcal{W}_i$. It aims to estimate a local fine-grained appearance flow:

$$\mathbf{f_{ri}} = conv([\mathcal{S}(g_{N+1-i}, \mathbf{f_{ci}}), p_{N+1-i}]), \qquad (4)$$

where $\mathbf{f_{ri}}$ is the predicted refinement flow, and $conv$ denotes convolution. Fundamentally, the refinement layer estimates the refinement flow through the local correspondence, i.e., the correspondence between warped person features and garment feature in the same receptive field. Note that after the warping by $\mathbf{f_{ci}}$, we can assume that the corresponding regions/features in g_{N+1-i} and p_{N+1-i} are now located in the same receptive field. Therefore, we can apply the local correspondence used in previous works [10, 13] to predict the local fine-grained appearance flow.

Finally, we add the coarse flow and the local fine-grained appearance flow together as the output of each warping block:

$$\mathbf{f_i} = \mathbf{f_{ci}} + \mathbf{f_{ri}}. \qquad (5)$$

The predicted appearance flow $\mathbf{f_N}$ from the last block in $\mathcal{W}$ is used to warp the garment:

$$\hat{g} = \mathcal{S}(g, \mathbf{f_N}). \qquad (6)$$

And the warped garment $\hat{g}$ is then concatenated with the person image and fed into a generator for target try-on image generation:

$$t = \mathcal{G}([\hat{g}, p]). \qquad (7)$$

The generator $\mathcal{G}$ has an encoder-decoder architecture with skip connections in between. We follow the designs in [18,

46] that have been proven to be effective in texture detail preservation.

3.5. Learning objectives

To train our model, we first apply a perceptual loss [20] between the output of $\mathcal{F}$ and the ground truth person image p_{gt}:

$$L_p = \sum_i \|\phi_i(t) - \phi_i(p_{gt})\|, \qquad (8)$$

where ϕ_i is the ith block of the pre-trained VGG network [35].

To supervise the training of the warping model $\mathcal{W}$, we apply a loss on the warped garment:

$$L_g = \|\hat{g} - m_g \cdot p_{gt}\|, \qquad (9)$$

where m_g is the garment mask of p_{gt} predicted by an off-the-shelf human parsing model.

As per standard in previous appearance flow methods [10, 13], we also apply a smoothness regularization on the predicted flow from each block in $\mathcal{W}$:

$$L_R = \sum_i \|\nabla \mathbf{f_i}\|, \qquad (10)$$

where $\|\nabla \mathbf{f_i}\|$ is the generalized charbonnier loss function [36].

As the inputs (segmentation map, keypoint pose and dense pose) to the parser-based person encoder ($\mathcal{E}^{PB}$) contain more semantic information than those of the parser-free model $\mathcal{F}$ (person image), we apply a distillation loss to guide the learning of person encoder $\mathcal{E}_p$ in $\mathcal{F}$:

$$L_D = \sum_i \|p_i^{PB} - p_i\|, \qquad (11)$$

where p_i^{PB} is the output feature map from i^{th} block in the person encoder $\mathcal{E}_p^{PB}$ in the pre-trained parser based model $\mathcal{F}_{PB}$.

The overall learning objective is:

$$L = \lambda_p L_p + \lambda_g L_g + \lambda_R L_R + \lambda_D L_D, \qquad (12)$$

where λ_p, λ_g, λ_R and λ_D denote the hyperparameters for balancing the four objectives.

4. Experiments

Datasets We experiment our model on the VITON dataset[5] [14]. It is the most popular dataset used in previous VTON works. VITON contains a training set containing $14,221$ image pairs[6] and a testing dataset of $2,032$

---

[5]The usage of the dataset has been permitted by the author in [14].

[6]Each pair means a person image and the image of garment on the person.

pairs. Both person and garment images are of the resolution 256×192.

We also create a testing dataset, denoted by augmented VITON, to evaluate model's robustness to the random positioned person image (see example in Fig. 4) with larger misalignments with the garment images in the original dataset. As most testing person images in VITON are well positioned such that the person image and the garment are well pre-aligned (e.g., most corresponding regions in the person image and garment image are roughly located in the same receptive field), it is not suited for this evaluation. Concretely, the augmented VITON dataset is created by randomly augmenting the testing person image in VITON via shifting and zooming in/out. In particular, we randomly augment 1/3 testing person images in VITON by shifting the person's position in the image and randomly augment another 1/3 test images in VITON by zooming in/out the person in the image and keep another 1/3 testing images unchanged. When evaluated on this dataset, all compared models are trained with person image augmentation.

Implementation details Our model is implemented in PyTorch. We train our model with a single Nvidia RTX 2080-Ti GPU. We set the batch size as 4 and train the model with 100 epochs. We train the model with Adam optimizer [23]. The initial learning rate is set to $5e-4$ which is linearly decayed after 50 epochs. Each residual block in $\mathcal{E}_p$ and $\mathcal{E}_g$ is followed by a pooling layer to reduce the spatial dimension. We set $N = 5$ and $c = 256$ in the implementation. We will release the code upon the acceptance of this work.

Evaluation metrics and baselines We evaluate our model both automatically and manually. In the automatic evaluation, as per standard in VTON, we evaluate model performance using structure similarity (SSIM) [40] and Fréchet Inception Distance (FID) [16]. According to [10, 31], inception score (IS) [32] is not suitable to evaluate VTON images, we thus do not adopt it in the evaluation. In the manual (subjective) evaluation, we run perceptual study on Amazon Mechanical Turk (AMT) to compare the quality of the generated try-on images from different models. Given an input person image, a garment image and the generated try-on image from two models, the AMT workers were asked to vote which generated try-on image is better. Each AMT worker was randomly allocated 100 images to compare two models. 15 AMT workers participated in the evaluation for all models comparison.

We compare our methods with other parser-based methods VTON [14], CP-VTON [38], Cloth-flow [13], CP-VTON++ [26], ACGPN [42], DCTON [9] and ZFlow [5]. We also compare with the SOTA parser-free method PF-AFN [10].

Methods	Warping	Parser	SSIM↑	FID↓
VTON [14]	TPS	Y	0.74	55.71
CP-VTON [38]	TPS	Y	0.72	24.45
CP-VTON++ [26]	TPS	Y	0.75	21.04
Cloth-flow [13]	AF	Y	0.84	14.43
ACGPN [42]	TPS	Y	0.84	16.64
DCTON [9]	TPS	Y	0.83	14.82
PF-AFN [10]	AF	N	0.89	10.09
Zflow [5]	AF	Y	0.88	15.17
Cloth-flow⋆ [13]	AF	N	0.89	10.73
Ours	AF	N	**0.91**	**8.89**

Table 1. Quantitative results of different models on VITON. Warping represents the warping methods used in different models. Parser indicates whether human parser is used in the model during inference. TPS: Thin Plate Spline. AF: Appearance Flow. ⋆: re-trained with parser free training paradigm.

Main results The quantitative results on VITON testing dataset are shown in Table 1. It can be seen that our model achieves new state-of-the-art performance. Importantly, given the already low FID score (10.09) achieved by prior SOTA method PF-AFN, our method can further decrease it by 11.9%. In the meanwhile, the following observations can be made from Table 1. (1) Appearance flow based warping methods generally perform better than TPS based warping methods. (2) Although it takes more training time, parser-free methods are much better than parser-based methods. Our model, benefiting from the proposed novel global appearance flow estimation method, outperforms the previous SOTA parser-free methods (PF-AFN [10] and Cloth-flow [13]) on all evaluation metrics. The human evaluation results are shown in Table 2. The result is consistent with that in Table 1. Our model outperforms all compared models with more than 10% preference rate. The qualitative results from different models are illustrated in Fig. 3. Overall, our method generates better try-on images. For example, the hard pose and occlusion in second and third rows.

The quantitative results on augmented testing dataset are shown in Table 3. As can be seen that our model again performs best on the augmented VITON testing dataset. Importantly, all other models' performance drops dramatically. And our model can still maintain the performance (SSIM score) compared to that on the original VITON testing dataset. The qualitative examples are illustrated in Fig. 4. Only our model can generate consistent (e.g., the garment's left sleeve) and high quality try-on images given the large mis-alignments.

Ablation Study In this experiment, we validate the design of our appearance flow estimation blocks ($\mathcal{W}_i$). Specifically, we first experiment our method with only global style modulation (SM) based appearance flow estimation, that
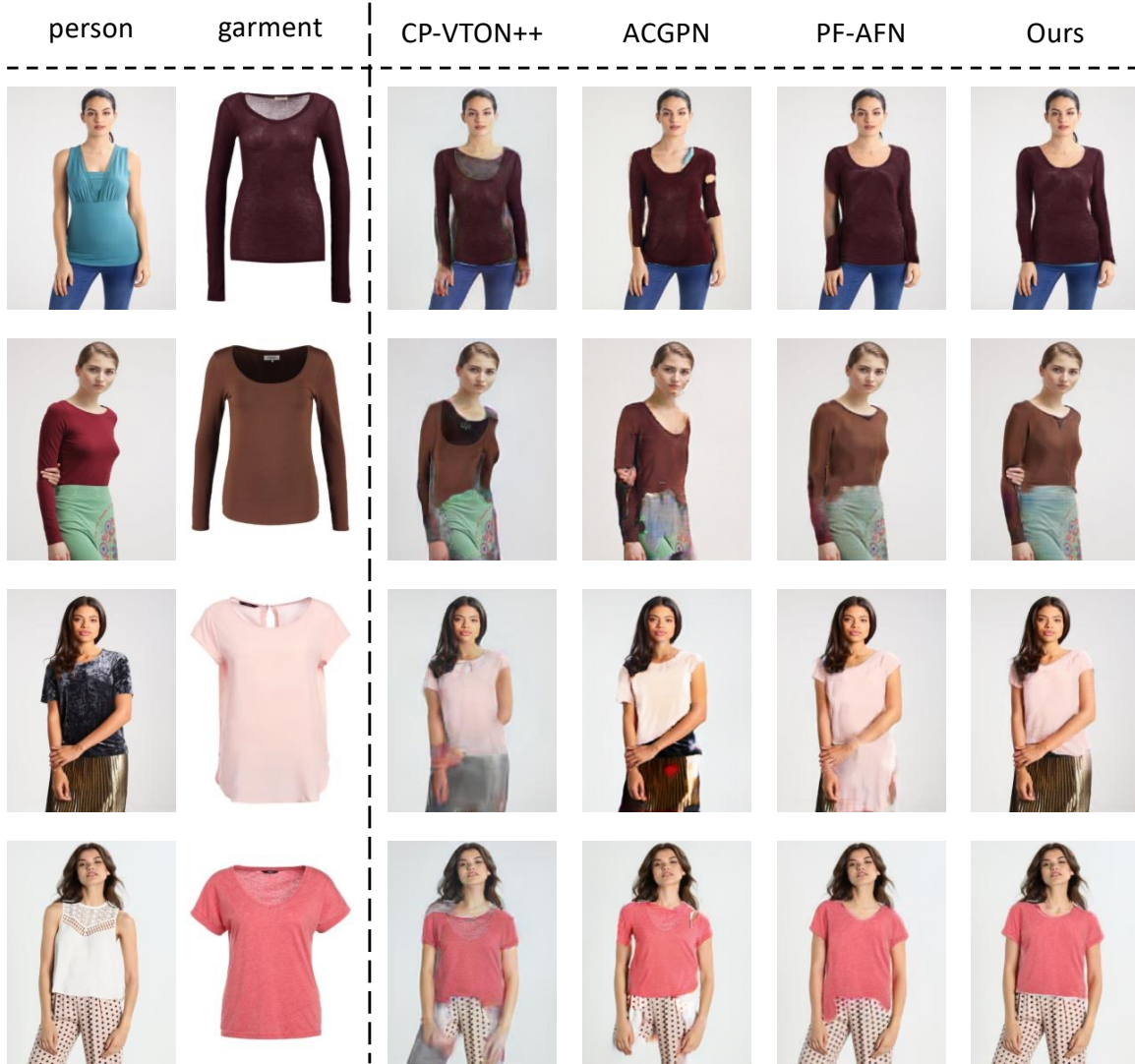
Figure 3. Qualitative results from different models (CP-VTON++ [26], ACGPN [42], PF-AFN [9] and ours) on VITON testing dataset.

Compared methods	preference rate
CP-VTON++ [26]	12.7% / 87.3%
ACGPN [42]	20.2% / 79.8%
Cloth-flow* [13]	38.5% / 61.5%
AF-PFN [10]	43.2% / 56.8%

Table 2. The preference rate comparing other models against our model (other models/our model) in human evaluation.

Methods	SSIM ↑	FID↓	$\nabla_{\text{SSIM}}/\nabla_{\text{FID}}$
ACGPN	0.81	20.75	0.003/4.11
Cloth-flow* [13]	0.86	13.05	0.003/2.96
AF-PFN [10]	0.87	12.19	0.002/2.10
Ours	**0.91**	**9.91**	**0/1.02**

Table 3. Quantitative results of different models on augmented VI-TON and their relative performance drop ($\nabla_{\text{SSIM}}/\nabla_{\text{FID}}$) compared to the standard VITON testing dataset.

is, only using $\mathbf{f_{ci}}$ in Equation 3 in each $\mathcal{W}_i$. We then experiment our method with only refinement flow (RF) estimation, that is, only using $\mathbf{f_{ri}}$ in Equation 4 in each $\mathcal{W}_i$. Finally, we experiment with our combined method (SM + RF) which first estimates the appearance flow globally via style modulation and then refines the appearance flow locally through local correspondence. The quantitative results

are shown in Table 4. Our proposed global style modulation (SM) based appearance flow method outperforms the local correspondence based method. When they were combined, the performance is further boosted. As illustrated in Fig. 5, without local refinement, our method (global style modu-

Figure 4. Illustrating different VTON models' robustness to the randomly positioned person image. First row uses original person image as input. And second row uses vertically shifted person image as input. ACGPN [42], Cloth-flow [13], PF-AFN [10].

lation only) sometimes cannot accurately predict the local fine-grained appearance flow, e.g., the sleeve regions, and thus generates unsatisfactory try-on image. However, with only local correspondence based appearance flow estimation, e.g., only using $\mathbf{f_{ri}}$ in $\mathcal{W}_i$, the method suffers when the corresponding regions are not located in the same receptive field. As illustrated in Fig. 6, $\mathbf{f_{ri}}$ cannot accurately estimate the appearance flow when there exists a large misalignment between the input person images and garment images. Once $\mathbf{f_{ci}}$ was first used to reduce the misalignment, our model can successfully overcome the problem.

Methods	SSIM ↑	FID↓
RF	0.89	10.73
SM	0.89	9.84
SM + RF	**0.91**	**8.89**

Table 4. Results on VTON testing dataset when different appearance flow estimation methods were used in $\mathcal{W}_i$. RF: local correspondence based flow estimation. SM: style modulation based flow estimation.

5. Conclusion

In this paper, we have proposed a style based global appearance flow estimation method to warp the garment for virtual try-on. Our method via style modulation first estimates the appearance flow globally and then refines the appearance flow locally. Our method achieves state-of-the-art performance on the VITON benchmark and it is more robust against large mis-alignment between person and garment images, as well as difficult poses/occlusions. We conducted extensive experiments to show the superiority of our method and validated our architecture design.

Figure 5. Comparing results with only $\mathbf{f_{ci}}$ used in $\mathcal{W}_i$ and $\mathbf{f_{ci}} + \mathbf{f_{ri}}$ used in $\mathcal{W}_i$.

Figure 6. Comparing results with only $\mathbf{f_{ri}}$ used in $\mathcal{W}_i$ and $\mathbf{f_{ci}} + \mathbf{f_{ri}}$ used in $\mathcal{W}_i$ in the case of large misalignment between the input person image and garment image.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. In *SIGGRAPH Asia*, 2021. 3

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, 2019. 3

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[4] Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. In *CVPR*, 2021. 3

[5] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *ICCV*, 2021. 2, 3, 6

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *CVPR*, 2015. 2, 4

[7] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 2

[8] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *ICCV*, 2019. 3

[9] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *CVPR*, 2021. 1, 2, 6, 7

[10] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[11] 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 *CVPR*, 2017. 2

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[13] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[14] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1, 2, 5, 6

[15] Sen He, Wentong Liao, Michael Ying Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Disentangled lifespan face synthesis. In *ICCV*, 2021. 2, 4

[16] 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. In *NeurIPS*, 2017. 6

[17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 2, 4

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 5

[19] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 2020. 1, 2, 3

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[24] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: body-aware try-on via layered interpolation. *TOG*, 40(4):1–10, 2021. 1, 3

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[26] 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 *CVPRW*, 2020. 6, 7

[27] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *CVPR*, 2020. 3

[28] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, pages 739–755, 2020. 2, 3, 4

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, 2020. 3

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 2

[31] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017. 6

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6

[33] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 3, 4

[34] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2, 3, 4

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

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

[37] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. Warpedganspace: Finding non-linear rbf paths in gan latent space. In *ICCV*, 2021. 3

[38] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 6

[39] Jiahang Wang, Tong Sha, Wei Zhang, Zhoujun Li, and Tao Mei. Down to the last detail: Virtual try-on with fine-grained details. In *ACM MM*, 2020. 1

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

[41] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *IJCV*, 129(5):1451–1466, 2021. 3

[42] 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 *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[43] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *ICCV*, 2019. 1, 2

[44] 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 *ICCV*, 2021. 3

[45] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2, 3

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 5