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

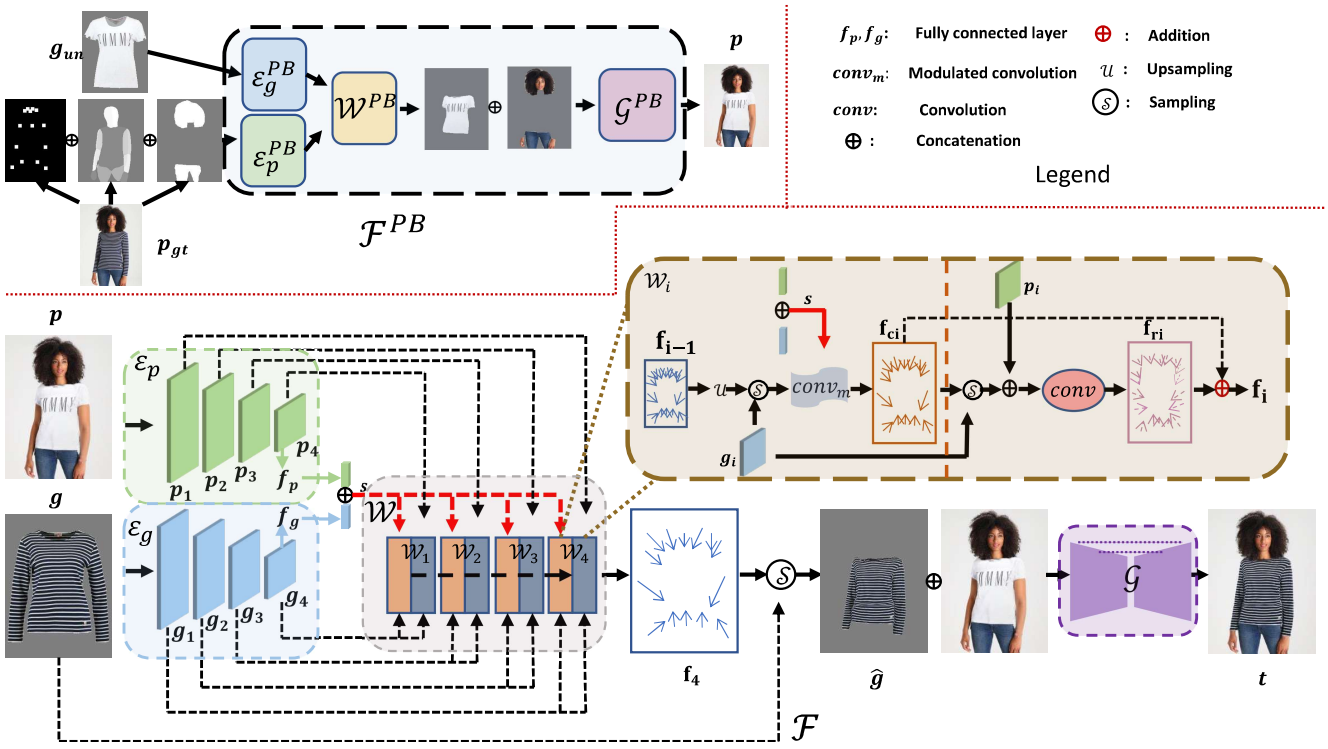

Figure 2. A schematic of our framework. The pre-trained parser based model $\mathcal{F}^{PB}$ generates an output image as the input of parser free model $\mathcal{F}$. The two feature extractors in $\mathcal{F}$ extract the feature of person image and garment image, respectively. A style vector is extracted 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:

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

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 N^{th} (final) blocks of $\mathcal{E}_p$ and $\mathcal{E}_g$, denoted as p_N and g_N, as:

$$s = [f_p(p_N), f_g(g_N)], \tag{2}$$

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.