This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Parser-Free Virtual Try-on via Distilling Appearance Flows

Yuying Ge[1] Yibing Song[2*] Ruimao Zhang[3,4] Chongjian Ge[1] Wei Liu[5] Ping Luo[1]
[1]The University of Hong Kong [2]Tencent AI Lab
[3]The Chinese University of Hong Kong (Shenzhen)
[4]Shenzhen Research Institute of Big Data [5]Tencent Data Platform
{yuyingge, rhettgee}@connect.hku.hk pluo@cs.hku.hk yibingsong.cv@gmail.com
ruimao.zhang@ieee.org wl2223@columbia.edu

Figure 1. Comparing our method with the recent state-of-the-art parser-based try-on methods (left) and an emerging parser-free method (right). On the left, we highlight the inaccurate segmentation regions in green boxes, which mislead existing parser-based methods such as CP-VTON [30], ClothFlow [8], CP-VTON+ [18], and ACGPN [32] to produce wrong results. On the right, the first parser-free method WUTON [13] was proposed recently, but its image quality is bounded by the fake images produced by the parser-based method, because [13] simply trained a "student" network to mimic the parser-based method using knowledge distillation. We see that our approach achieves significantly better image quality than previous state-of-the-art methods, without relying on human segmentation.

Abstract

Image virtual try-on aims to fit a garment image (target clothes) to a person image. Prior methods are heavily based on human parsing. However, slightly-wrong segmentation results would lead to unrealistic try-on images with large artifacts. A recent pioneering work employed knowledge distillation to reduce the dependency of human parsing, where the try-on images produced by a parser-based method are used as supervisions to train a "student" network without relying on segmentation, making the student mimic the try-on ability of the parser-based model. However, the image quality of the student is bounded by the parser-based model. To address this problem, we propose a novel approach, "teacher-tutor-student" knowledge distillation, which is able to produce highly photo-realistic images without human parsing, possessing several appealing advantages compared to prior arts. (1) Unlike existing work, our approach treats the fake images produced by the parser-based method as "tutor knowledge", where the artifacts can be corrected by real "teacher knowledge", which is extracted from the real person images in a self-supervised way. (2) Other than using real images as supervisions, we formulate knowledge distillation in the try-on problem as distilling the appearance flows between the person image and the garment image, enabling us to find accurate dense correspondences between them to produce high-quality results. (3) Extensive evaluations show large superiority of our method (see Fig. 1).

*Y. Song is the corresponding author. This work is done when Y. Ge is an intern in Tencent AI Lab. The code is available at https://github.com/geyuying/PF-AFN.

1. Introduction

Virtual try-on of fashion image is to fit an image of a clothing item (garment) onto an image of human body. This task has attracted a lot of attention in recent years because of its wide applications in e-commerce and fashion image editing. Most of the state-of-the-art methods such as VTON [9], CP-VTON [30], VTNFP [33], ClothFlow [8], ACGPN [32], and CP-VTON+ [18] were relied on human segmentation of different body parts such as upper body, lower body, arms, face, and hairs, in order to enable the learning procedure of virtual try-on. However, high-quality human parsing is typically required to train the try-on models, because slightly wrong segmentation would lead to highly-unrealistic try-on images, as shown in Fig.1.

To reduce the dependency of using accurate masks to guide the try-on models, a recent pioneering work WUTON [13] presented the first parser-free network without using human segmentation for virtual try-on. Unfortunately, [13] has an inevitable weakness in its model design. As shown in the bottom of Fig.2, WUTON employed a conventional knowledge distillation scheme by treating a parser-based model (*i.e*. a try-on network that requires human segmentation) as a "teacher" network, and distilling the try-on images (*i.e*. fake person images) produced by the teacher to a parser-free "student" network, which does not use segmentation as input. This is to make the parser-free student directly mimic the try-on ability of the parser-based teacher. However, the generated images of the parser-based teacher have large artifacts (Fig.1), thus using them as the teacher knowledge to supervise the student model produces unsatisfactory results since the image quality of the student is bounded by the parser-based model.

To address the above problems, this work proposes a new perspective to produce highly photo-realistic try-on images without human parsing, called Parser Free Appearance Flow Network (PF-AFN), which employs a novel "teacher-tutor-student" knowledge distillation scheme. As shown at the top of Fig.2, instead of treating the parser-based model as the teacher, PF-AFN only treats it as a "tutor" network that may produce unrealistic results (*i.e*. tutor knowledge), which need to be improved by a real teacher. The key is to design where the teacher knowledge comes from. To this end, PF-AFN treats the fake person image (tutor knowledge) as input of the parser-free student model, which is supervised by the original real person image (teacher knowledge), making the student mimic the original real images. This is similar to self-supervised learning, where the student network is trained by transferring the garment on the real person image to the fake person image produced by the parser-based model. In other words, the student is asked to change the clothes on the fake person image to the clothes on the real person image, enabling it to be self-supervised by the real person image that naturally has no artifacts. In this case, the images generated by our parser-free model significantly outperform its previous counterparts.

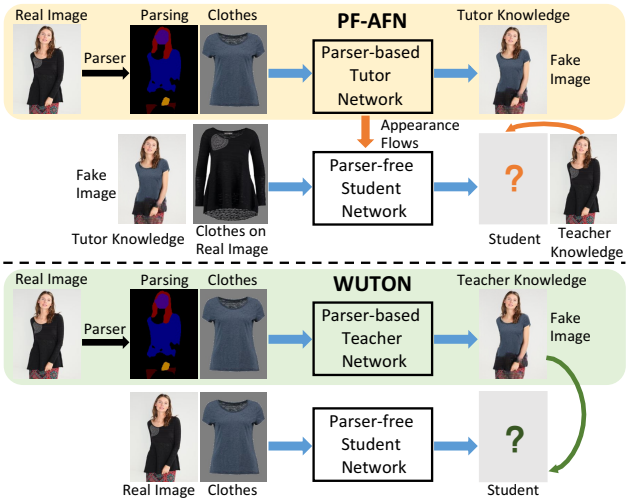

Figure 2. The comparison between WUTON [13] and PF-AFN. WUTON treats a parser-based network as a "teacher" network and distills the fake person image produced by the teacher to a parser-free "student" network, making the student directly mimic the try-on ability of the parser-based teacher. In comparison, our PF-AFN treats the fake person image as "tutor knowledge" and uses it as the input of the parser-free "student" network, which is supervised by the real person image (teacher knowledge). The parser-based "tutor" network further distills the appearance flows between the person image and the clothes image, facilitating the high-quality image generation in the "student" network.

To further improve image quality of the student, other than using real images as supervisions, we formulate knowledge distillation of the try-on problem as distilling the appearance flows between the person image and the garment image, facilitating to find dense correspondences between them to generate high-quality images.

Our work has three main **contributions**. First, we propose a "teacher-tutor-student" knowledge distillation scheme for the try-on problem, to produce highly photo-realistic results without using human segmentation as model input, completely removing human parsing. Second, we formulate knowledge distillation in the try-on problem as distilling appearance flows between the person image and the garment image, which is important to find accurate dense correspondences between pixels to generate high-quality images. Third, extensive experiments and evaluations on the popular datasets demonstrate that our proposed method has large superiority compared to the recent state-of-the-art approaches both qualitatively and quantitatively.

2. Related Work

Virtual Try-on. Existing deep learning based methods on virtual try-on can be classified as 3D model based approaches [2, 20, 24, 19] and 2D image based ones [9, 30,

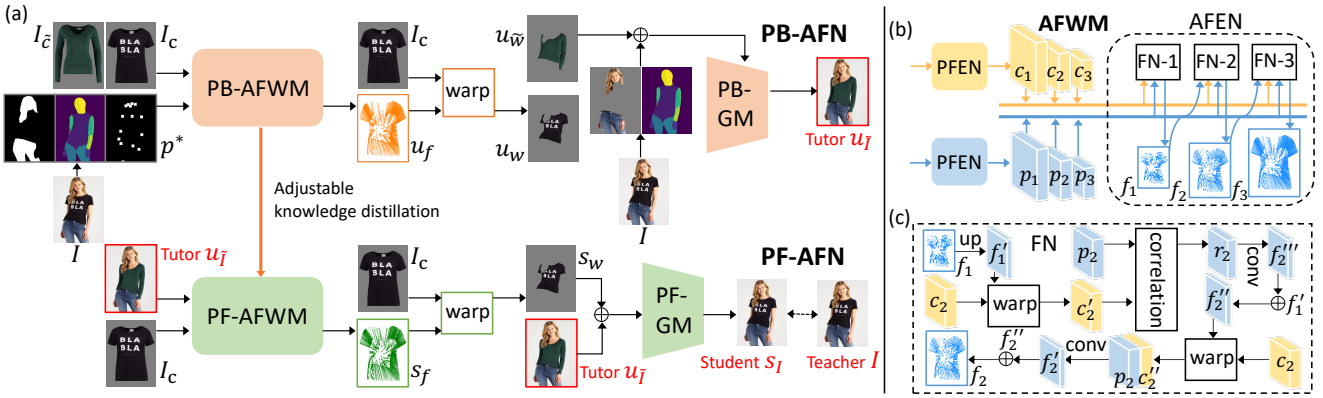

Figure 3. The training pipeline of PF-AFN. The training data is the clothes image I_c and the image I of a person wearing the clothes. We obtain p^* from the person image I as the parser-based inputs. Given p^*, the parser-based network PB-AFN randomly selects a different clothes image $I_{\tilde{c}}$ to synthesize the fake image $u_{\tilde{I}}$ as the tutor. We use the tutor $u_{\tilde{I}}$ and the clothes image I_c as inputs to train the parser-free network PF-AFN, where the generated student s_I is directly supervised by the real image I. Furthermore, PB-AFN estimates the appearance flows u_f between I_c and p^*, and distills the appearance flows to PF-AFN through the adjustable knowledge distillation. During inference, a target clothes image and a reference person image will be fed into PF-AFN to generate the try-on image, without the need of human parsing results or human pose estimations.

8, 33, 18, 32, 13]. As the former require additional 3D measurements and more computing power, 2D image based approaches are more broadly applicable. Since available datasets [9, 5] for 2D image try-on only contain unpaired data (clothes and a person wearing the clothes), previous methods [9, 30, 8, 33, 18, 32] mainly mask the clothing region of the person image and reconstruct the person image with the corresponding clothes image, which require accurate human parsing. When parsing results are inaccurate, such parser-based methods generate visually terrible try-on images with noticeable artifacts. WUTON [13] recently proposes a pioneering parser-free approach, but makes the quality of the generated image from a parser-free network bounded by fake images from a parser-based network.

Appearance Flow. Appearance flow refers to 2D coordinate vectors indicating which pixels in the source can be used to synthesize the target. It motivates visual tracking [26], image restorations [17, 31] and face hallucination [27]. Appearance flow is first introduced by [35] to synthesize images of the same object observed from arbitrary viewpoints. The flow estimation is limited on the non-rigid clothing regions with large deformation. [15] uses 3D appearance flows to synthesize a person image with a target pose, via fitting a 3D model to compute the appearance flows as supervision, which are not available in 2D try-on.

Knowledge Distillation. Knowledge distillation leverages the intrinsic information of a teacher network to train a student network, which was first introduced in [12] for model compression. As introduced in [34], knowledge distillation has also been extended as cross-modality knowledge transfer, where one model trained with superior modalities (*i.e.* multi-modalities) as inputs intermediately supervises another model taking weak modalities (*i.e.* single-modality) as inputs, and the two models can use the same network.

3. Proposed Approach

We propose an approach to produce highly photo-realistic try-on images without human parsing, called Parser Free Appearance Flow Network (PF-AFN), which employs a novel "teacher-tutor-student" knowledge distillation scheme. We further formulate knowledge distillation of the try-on problem as distilling the appearance flows between the person image and the clothes image. We first clarify the overall training scheme with the "teacher-tutor-student" knowledge distillation in Sec. 3.1. We use an appearance flow warping module (AFWM) to establish accurate dense correspondences between the person image and the clothes image, and a generative module (GM) to synthesize the try-on image, which are introduced in detail in Sec. 3.2 and Sec. 3.3. At last, we describe how we distill the appearance flows to generate high-quality images in Sec. 3.4.

3.1. Network Training

As shown in Fig. 3, our method contains a parser-based network PB-AFN and a parser-free network PF-AFN. We first train PB-AFN with data (I_c, I) following the existing methods [30, 8, 32], where I_c and I indicate the image of the clothes and the image of the person wearing this clothes. We concatenate a mask containing hair, face, and the lower-body clothes region, the human body segmentation result, and the human pose estimation result as the person representations p^* to infer the appearance flows u_f between p^* and the clothes image I_c. Then the appearance flows u_f are used to generate the warped clothes u_w with I_c. Concatenating this warped clothes, the preserved regions on the

person image and human pose estimation along channels as inputs, we could train a generative module to synthesize the person image with the ground-truth supervision I.

After training PB-AFN, we randomly select a different clothes image $I_{\widetilde{c}}$ and generate the try-on result $u_{\widetilde{I}}$, that is the image of person in I changing a clothes. Intuitively, the generated fake image $u_{\widetilde{I}}$ is regarded as the input to train the student network PF-AFN with the clothes image I_c. We treat the parser-based network as the "tutor" network and its generated fake image as "tutor knowledge" to enable the training of the student network. In PF-AFN, a warping module is adopted to predict the appearance flows s_f between the tutor $u_{\widetilde{I}}$ and the clothes image I_c and warp I_c to s_w. A generative module further synthesizes the student s_I with the warped clothes and the tutor. We treat the real image I as the "teacher knowledge" to correct the student s_I, making the student mimic the original real image. Furthermore, the tutor network PB-AFN distills the appearance flows u_f to the student network PF-AFN though adjustable knowledge distillation, which will be explained in Sec. 3.4.

3.2. Appearance Flow Warping Module (AFWM).

Both PB-AFN and PF-AFN contain the warping module AFWM, to predict the dense correspondences between the clothes image and the person image for warping the clothes. As shown in Fig. 3, the output of the warping module is the appearance flows (*e.g.* u_f), which are a set of 2D coordinate vectors. Each vector indicates which pixels in the clothes image should be used to fill the given pixel in the person image. The warping module consists of dual pyramid feature extraction network (PFEN) and a progressive appearance flow estimation network (AFEN). PFEN extracts two-branch pyramid deep feature representations from two inputs. Then at each pyramid level, AFEN learns to generate coarse appearance flows, which are refined in the next level. The second-order smooth constraint is also adopted when learning the appearance flows, to further preserve clothes characteristics, *e.g.* logo and stripe. The parser-based warping module (PB-AFWM) and the parser-free warping module (PF-AFWM) have the identical architecture except for the difference in the inputs.

Pyramid Feature Extraction Network (PFEN) As shown in Fig. 3 (b), PFEN contains two feature pyramid networks (FPN) [16] to extract two-branch pyramid features from N levels. For the parser-based warping module, the inputs are the clothes image I_c and the person representations p^*, while the inputs of the parser-free warping module are the clothes image I_c and the generated fake image $u_{\widetilde{I}}$. Here we use $\{c_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ to indicate two-branch pyramid features respectively. In practice, each FPN contains N stages. It is worth note that we set $N = 5$ in our model but show the case $N = 3$ in Fig. 3 for simplicity.

Appearance Flow Estimation Network (AFEN). AFEN consists of N Flow Networks (FN) to estimate the appearance flows from N levels' pyramid features. The extracted pyramid features (c_N, p_N) at the highest level N are first fed into FN-1 to estimate the initial appearance flows f_1. Then f_1 and the pyramid features at the $N-1$ level are fed into FN-2 for a finer flow f_2. The above process continues until the finest flow f_N is obtained, and the target clothes is warped according to f_N.

As illustrated in Fig. 3 (c), we carefully design the FN module, which performs pixel-by-pixel matching of features to yield the coarse flow estimation with a subsequent refinement at each pyramid level. Take the FN-2 as an example, the inputs are two-branch pyramid features (c_2, p_2), as well as the estimated appearance flow f_1 from previous pyramid level. The operations in FN can be roughly divided into four stages. **In the first stage**, we upsample f_1 to obtain f_1', and then c_2 is warped to c_2' through sampling the vectors in c_2 where the sampling location is specified by f_1'. **In the second stage**, the correlation maps r_2 is calculated based on c_2' and p_2. In practice, the j-th point in r_2 is a vector representation, which indicates the result of vector-matrix product between the j-th point in c_2' and the local displacement region centered on the j-th point in p_2. In such case, the number of channels of r_2 equals to the number of points in the above local displacement region. **In the third stage**, once r_2 is obtained, we then feed it into a ConvNet to predict the residual flow f_2''', which is added to f_1' as the coarse flow estimation f_2''. **In the fourth stage**, c_2 is warped to c_2'' according to the newly generated f_2''. Then c_2'' and p_2 are concatenated and fed into a ConvNet to compute the residual flow f_2'. By adding f_2' to f_2'', we obtain the final flow f_2 at pyramid level 2.

Intuitively, FN performs matching between two-branch high-level features and a further refinement. AFEN progressively refines the estimated appearance flows through cascading N FN, to capture the long-range correspondence between the clothes image and the person image, thus it is able to deal with large misalignment and deformation.

Second-order Smooth Constraint. According to Fig. 4, the target clothes usually contain tightly arranged text and the repeating pattern (*e.g.* stripes appear). The appearance flows between the person image and the clothes image need to be predicted accurately, or the minor mistakes should result in very unnatural warping results. To better preserve the clothes characteristics, we introduce a second-order smooth constraint to encourage the co-linearity of neighbouring appearance flows. The constraint is defined as follows:

$$\mathcal{L}_{sec} = \sum_{i=1}^{N} \sum_{t} \sum_{\pi \in \mathcal{N}_t} \mathcal{P}(f_i^{t-\pi} + f_i^{t+\pi} - 2f_i^t) \qquad (1)$$

where f_i^t denotes the t-th point on the flow maps of i-th scale (*i.e.* corresponding to the $\{f_i\}_{i=1}^N$ in Fig. 3 (b)). $\mathcal{N}_t$ indicates the set of horizontal, vertical, and both diagonal neighborhoods around the t-th point. The $\mathcal{P}$ is the gen-

Figure 4. Comparison between (a) with and without the second-order smooth constraint on appearance flows; (b) when PB-AFN generates accurate warping, our PF-AFN and the model without knowledge distillation; when PB-AFN generates wrong warping, our PF-AFN and the model with fixed knowledge distillation.

eralized charbonnier loss function [29]. As illustrated in Fig. 4 (a), adding $\mathcal{L}_{sec}$ helps maintain the details of the target clothes (*i.e.* the stripes and the characters on the clothes are retained without being distorted).

3.3. Generative Module (GM)

Both PB-AFN and PF-AFN contain the generative module to synthesize the try-on image. The parser-based generative module (PB-GM) concatenates the warped clothes, human pose estimation, and the preserved region on the human body as inputs, while the parser-free generative module (PF-GM) concatenates the warped clothes and the tutor image $u_{\tilde{I}}$ as inputs. Both modules adopt the Res-UNet, which is built upon a UNet [21] architecture, in combination with residual connections, which can preserve the details of the warped clothes and generate realistic try-on results.

In the training phase, the parameters of the generative module GM and the warping module AFWM are optimized together by minimizing $\mathcal{L}$, as follows:

$$\mathcal{L} = \lambda_l \mathcal{L}_l + \lambda_p \mathcal{L}_p + \lambda_{sec} \mathcal{L}_{sec} \tag{2}$$

where $\mathcal{L}_l$ is the pixel-wise L1 loss and $\mathcal{L}_p$ is the perceptual loss [14] to encourage the visual similarity between the try-on image (*i.e.* the output s_I of the student network) and the real image I as below:

$$\mathcal{L}_l = ||s_I - I||_1 \tag{3}$$

$$\mathcal{L}_p = \sum_m || \phi_m(s_I) - \phi_m(I) ||_1 \tag{4}$$

where ϕ_m indicates the m-th feature map in a VGG-19 [25] network pre-trained on ImageNet [3].

3.4. Adjustable Knowledge Distillation

Other than supervising the parser-free student network PF-AFN with the real image I, we further distill the appearance flows between the person image and the clothes image, facilitating to find dense correspondences between them. As shown in Fig. 3 (a), the inputs of the parser-based tutor network PB-AFN include human parsing results, densepose estimations [1] and pose estimations of the input person. In contrast, the input of the student network PF-AFN is only the fake image and the clothes image. Thus, in most cases, the extracted features from PB-AFN usually capture richer semantic information and the estimated appearance flows are more accurate, thus can be used to guide PF-AFN. However, as mentioned before, if the parsing results are not accurate, the parser-based PB-AFN would provide totally wrong guidance, making its semantic information and predicted flows irresponsible. To address the above issue, we introduce a novel adjustable distillation loss to ensure only accurate representations and predictions are maintained. The definition is as follows:

$$\mathcal{L}_{hint} = \psi \sum_{i=1}^{N} ||u_{p_i} - s_{p_i}||_2 \tag{5}$$

$$\mathcal{L}_{pred} = \psi \sum_{i=1}^{N} ||\sqrt{(u_{f_i} - s_{f_i})^2}||_1 \tag{6}$$

$$\psi = \begin{cases} 1, \text{if } ||u_I - I||_1 < ||s_I - I||_1 \\ 0, \text{otherwise} \end{cases} \tag{7}$$

$$\mathcal{L}_{kd} = \lambda_{hint} \mathcal{L}_{hint} + \lambda_{pred} \mathcal{L}_{pred} \tag{8}$$

where u_I and s_I are the generated try-on image from PB-AFN and PF-AFN respectively, I is the real person image. u_{p_i} and s_{p_i} are features extracted from the person representation p^* and the fake image $u_{\tilde{I}}$ at the i-th scale (*i.e.* corresponding to the $\{p_i\}_{i=1}^N$ in Fig. 3 (b)). u_{f_i} and s_{f_i} are predicted appearance flows from PB-AFN and PF-AFN at the i-th scale (*i.e.* corresponding to the $\{f_i\}_{i=1}^N$ in Fig. 3 (b)). Specifically, ψ is the adjustable factor to decide whether the distillation is enabled by utilizing the teacher to assess the quality of the generated image. If the quality of the generated image u_I from the parser-based tutor network does not exceed that of s_I from the parser-free student network (*i.e.* the L1 loss between u_I and I is larger than that between s_I and I), the distillation will be disabled.

We compare the warped clothes in Fig. 4 (b) and visualize the activations using the guided prorogation algorithm [28]. When PB-AFN achieves pleasant performance as shown in the first row, the model without distillation fail to generate accurate warping for the sleeve when it is not activated by the arm. When PB-AFN performs poorly as shown in the second row, the model with the fixed distillation (not adjustable distillation) inherits the defects of PB-AFN with erroneous warping to lower-body region when it is activated by the lower-body. In both cases, PF-AFN warps the target clothes accurately, which demonstrates the efficiency of the adjustable knowledge distillation.

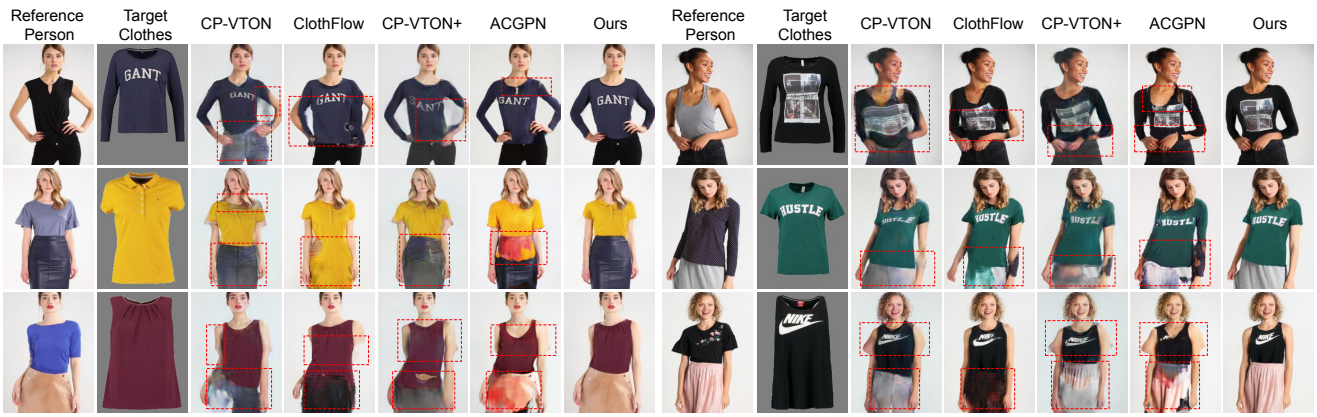

Figure 5. Visual comparison on VITON dataset. Compared with the recent parser-based methods [30, 8, 18, 32], our model generates more highly-realistic try-on images without relying on human parsing, which simultaneously handles large misalignment between the clothes and the person, preserves the characteristics of both the target clothes and the non-target clothes (*i.e.* skirt), and retains clear body parts.

4. Experiments

4.1. Datasets

We conduct experiments on VITON [9], VITON-HD [9] and MPV [5], respectively. VITON contains a training set of $14,221$ image pairs and a testing set of $2,032$ image pairs, each of which has a front-view woman photo and a top clothing image with the resolution 256×192. Most of previous work in virtual try-on apply this dataset for training and validation. VITON-HD is the same as VITON, except that the image resolution is 512×384. It hasn't been tackled before, since it is critically challenging to generate photo-realistic try-on results by giving inputs with high resolutions. As a recent constructed virtual try-on dataset, MPV contains $35,687$ / $13,524$ person / clothes images at 256×192 resolution and a test set of 4175 image pairs are split out. Since there are multiple images of a person wearing the target clothes from different views in MPV, following [13], we remove images tagged as back ones since the target clothes is only from the front. WUTON [13] is the only work that conducts experiments on this dataset.

4.2. Implementation Details

Architecture. Both PB-AFN and PF-AFN consist of the warping module (AFWM) and the generative module (GM), where the warping module includes dual pyramid feature extraction network (PFEN) and an appearance flow estimation network (AFEN). PFEN adopts the FPN [16] with five layers in practice, and each layer is composed of a convolution with a stride of 2, followed by two residual blocks [10]. AFEN comprises five flow network (FN) blocks, and each FN contains two ConvNets with four convolution layers. The generative module has the same structure of Res-UNet [4] in an encoder-decoder style.

Training. The training process on three datasets are same. we first train PB-AFN with the clothes image and the image of the person wearing the clothes. The parsing results and human pose estimations [1] are also applied in this phase. PB-AFN is optimized for 200 epochs with the initial learning rate 3×10^{-5} and we have $\lambda_l = 1.0$, $\lambda_p = 0.2$, and $\lambda_{sec} = 6.0$. PF-AFN adopts the same training schedule as PB-AFN and uses the same hyper-parameters setting, where $\lambda_{hint} = 0.04$ and $\lambda_{pred} = 1.0$.

Testing. During test, the reference person image and the target clothes image are given as the input of PF-AFN to generate the image. Additional inputs such as human parsing results and human pose estimations are removed.

4.3. Qualitative Results

Results of VITON. We mainly perform visual comparison of our method with recent proposed parser-based methods in Fig. 5, including CP-VTON [30], ClothFlow [8], CP-VTON+ [18], and ACGPN [32] . As shown in the first row of Fig. 5, when the reference person strikes a complex posture like standing with arms akimbo or two hands blocking in front of the body, large misalignment occurs between the target clothes and the person. In such case, baseline models all fail to warp the long-sleeve shirt to the corresponding body region, leading to broken sleeves, sleeves not attached to the arms and distorted embroideries. Actually, these methods cannot model the highly non-rigid deformation due to the deficiency of the warping methods, *i.e.* limited degrees of freedom in TPS [7].

In the second and the third rows of Fig. 5, images generated by baseline methods exist the clear artifacts, such as messy lower-body clothes and top clothes being warped to lower-body region. These parser-based models are delicate to segmentation errors because they heavily rely on parsing results to drive the image generation. Furthermore, when there exists huge discrepancy between the target clothes and the original clothes on the person (*e.g.* the person wears a low-collar blouse while the target clothes is high-necked),

Figure 6. Visual comparison on MPV dataset with parser-free inputs. Compared with WUTON [13], our model generates far more satisfactory results, which warps the target clothes to the person accurately even when the person strikes a complex posture (*i.e*. occlusions and cross-arms), and preserves the characteristics of both the target clothes and the non-target clothes (*i.e*. skirt).

Method	Dataset	FID	Human
CP-VTON[30]	VITON	24.43	11.15% / 88.85%
ClothFlow[8]	VITON	14.43	22.38% / 77.62%
CP-VTON+[18]	VITON	21.08	12.62% / 87.38%
ACGPN[32]	VITON	15.67	16.54% / 83.46%
PF-AFN (ours)	VITON	**10.09**	reference
WUTON[13]	MPV	7.927	28.38% / 71.62%
PF-AFN (ours)	MPV	**6.429**	reference

Table 1. Quantitative evaluation results FID [11] and user study results. For FID, the lower is the better. For Human result **"a / b"**, a is the percentage where the compared method is considered better over our PF-AFN, and **b is the percentage where our PF-AFN is considered better** over the compared method.

CP-VTON [30] and ACGPN [32] fail to preserve the characteristics of the target clothes, since they excessively focus on the silhouette of the original clothes during training. Moreover, these baseline models are also weak in generating non-target body parts, where obviously fake arms, blurring hands and finger gaps appear on the generated images.

In comparison, the proposed PF-AFN generates highly-realistic try-on results, which simultaneously handles large misalignment between the clothes and the person, preserves the characteristics of both the target clothes and the non-target clothes, and retains clear body parts. Besides the above advantages, benefited from the second-order smooth constraint on the appearance flows, PF-AFN is able to model long-range correspondences between the clothes and the person, avoiding the distortion in logo and embroideries. Since we do not mask any information such as clothes or body parts for the input person image during training, PF-AFN can adaptively preserve or generate the body parts, such that the body details can be retained.

Results of VITON-HD The results on VITON-HD are provided in the supplement material.

Results of MPV. The comparison with WUTON [13], which is a pioneer parser-free method, on MPV are shown in Fig. 6. WUTON produces visually unpleasant results with clear artifacts. For example, it cannot distinguish the boundary between the top and bottom clothes, making the target top clothes be warped to low-body region. In addition, when complicated poses appear in the person images such as occlusions and cross-arms, WUTON generates unnatural results with erroneous warping. Since WUTON is supervised by the fake images from a parser-based model that can be misleading, it inevitably achieves unsatisfying performance. In comparison, our PF-AFN can warp the clothes to the target person accurately even in the case of complicated poses and generate high-quality images, which preserves the characteristics of both the target clothes and the non-target clothes (*i.e*. skirt). PF-AFN benefits from being supervised by real images as well as finding accurate dense correspondences between the clothes and the person through distilling the appearance flows.

4.4. Quantitative Results

For virtual try-on, a target clothes and a reference person image are given to generate the try-on results during the test. Since we do not have the ground-truth images (*i.e*. reference person wearing the target clothes), we adopt the Fréchet Inception Distance (FID) [11] as the evaluation metric following [13], which captures the similarity of generated images to real images (*i.e*. reference person images). Lower score of FID indicates higher quality of the results. We do not use the Inception Score (IS) [23] since Rosca *et.al* [22] have pointed out that applying the IS to the models trained on datasets other than ImageNet will give misleading results.

Table. 1 lists the FID scores of CP-VTON [30], ClothFlow [8], CP-VTON+ [18], ACGPN [32], WUTON [13] and proposed PF-AFN on the VITON and MPV dataset. Compared with parser-based methods on VITON dataset, our PF-AFN outperforms them by a large margin, showing its great advantage in generating high-quality try-on images without being interfered by parsing results. PF-AFN also surpasses WUTON with parser-free inputs on MPV dataset, which demonstrates the superiority of our 'teacher-tutor-student' knowledge distillation scheme.

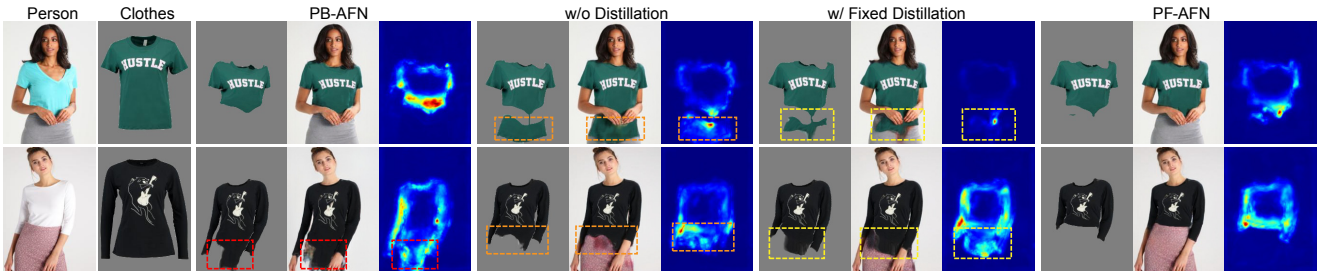

Figure 7. Ablation studies on the effects of the "adjustable knowledge distillation". Given a reference person image and a target clothes image, we show the warped clothes, the try-on image and the visualization of activations using [28] for each model.

Different configurations of the AFEN	FID
An encoder-decoder architecture following [6].	12.95
FN w/o a refinement from f_2'' to f_2, only predicting f_2''.	10.79
FN w/o "correlation" and a refinement from f_2'' to f_2, only predicting f_2'' by forwarding p_2 and c_2' to "conv".	11.38
A single FN w/o cascading N FN modules.	11.90
PF-AFN with cascading N FN modules.	**10.09**

Table 2. Ablation studies of the appearance flow estimation network (AFEN), which consists of flow networks (FN), on VITON. Lower FID indicates better results.

4.5. User Study

Although FID can be used as an indicator of the image synthesis quality, it cannot reflect whether the the target clothes are naturally warped with details preserved or the body of the person are retained, so we further conduct a user study by recruiting 50 volunteers in an A / B manner. Specially, 300 pairs from the VITON test set are randomly selected, and CP-VTON [30], ClothFlow [8], CP-VTON+ [18], ACGPN [32], PF-AFN each generates 300 images. 300 pairs from the MPV test set are also randomly selected, and WUTON [13], PF-AFN each generates 300 images. For each compared method, we have 300 image groups, where each group contains four images, *i.e.* a target clothes, a reference person image, two try-on images from the compared method and our PF-AFN, respectively. Each volunteer is asked to choose the one with better visual quality. As shown in Table. 1, our PF-AFN is always rated better than the other methods with much higher percentage. In the A/B test conducted between WUTON and PF-AFN, 71.62% of the images generated by PF-AFN were chosen by the volunteers to have a better quality.

4.6. Ablation Study

Adjustable Knowledge Distillation. We show the ablation studies on the effects of the "adjustable knowledge distillation". (1) As shown in Fig. 7, when PB-AFN generates comparatively accurate warping in the first row, the model without knowledge distillation is activated by the low-body and mistakenly warps the top clothes to the low-body region since it does not receive parsing guidance during training. (2) In the second row, when PB-AFN generates erroneous warping caused by the parsing errors, the model with fixed distillation (not adjustable distillation) also generates the failure case because it receives misleading guidance from PB-AFN during training. (3) In contrast, our PF-AFN could generate satisfactory results in both cases. (4) FID on the results predicted by the student network without distillation is 11.40, with fixed distillation is 10.86, and with adjustable distillation is 10.09. Since lower FID indicates better results, the effectiveness of the adjusted knowledge distillation scheme is verified, where only accurate feature representations and predicted flows from a parser-based network will guide the parser-free student network during training.

Appearance Flow Estimation Network (AFEN). We show the ablation studies of the AFEN, which consists of Flow Networks (FN), in Table 2. (1) We use a simple encoder-decoder following [6]. The results are unsatisfying, which indicates that this architecture does not produce accurate appearance flows for clothes warping. (2) We remove refinement, correlation, and cascaded modules of FN, respectively, and get worse results. (3) With all of the components, PF-AFN achieves the best performance, which demonstrates the effectiveness of our AFEN.

5. Conclusion

In this work, we propose a novel approach, "teacher-tutor-student" knowledge distillation, to generate highly photo-realistic try-on images without human parsing. Our approach treats the fake images produced by the parser-based network (tutor knowledge) as input of the parser-free student network, which is supervised by the original real person image (teacher knowledge) in a self-supervised way. Besides using real images as supervisions, we further distill the appearance flows between the person image and the clothing image, to find accurate dense correspondence between them to for high-quality image generation. Extensive evaluations clearly show the great superiority of our approach over the state-of-the-art methods.

Acknowledgment This work is supported by CCF-Tencent Open Fund and General Research Fund of HK No.27208720.

References

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

[2] Remi Brouet, Alla Sheffer, Laurence Boissieux, and Marie-Paule Cani. Design preserving garment transfer. 2012.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS*, 162, 2020.

[5] 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 *ICCV*, 2019.

[6] Alexey Dosovitskiy, etc. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015.

[7] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*. 1977.

[8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019.

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] 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.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Thibaut Issenhuth, J. Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 2020.

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

[15] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[17] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, 2020.

[18] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, June 2020.

[19] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *TOG*, 36, 2017.

[20] Damien Rohmer, Tiberiu Popa, Marie-Paule Cani, Stefanie Hahmann, and Alla Sheffer. Animation wrinkling: augmenting coarse cloth simulations with realistic-looking wrinkles. *TOG*, 29, 2010.

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

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

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

[24] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, 2019.

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

[26] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *ICCV*, 2017.

[27] Yibing Song, Jiawei Zhang, Lijun Gong, Shengfeng He, Linchao Bao, Jinshan Pan, Qingxiong Yang, and Ming-Hsuan Yang. Joint face hallucination and deblurring via structure generation and detail enhancement. *IJCV*, 2019.

[28] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[29] 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, 2014.

[30] 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.

[31] Yinglong Wang, Yibing Song, Chao Ma, and Bing Zeng. Rethinking image deraining via rain streaks and vapors. In *ECCV*, 2020.

[32] 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.

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

[34] Long Zhao, etc. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *CVPR*, 2020.

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