Disentangled Cycle Consistency for Highly-realistic Virtual Try-On

Chongjian Ge[1] Yibing Song[2*] Yuying Ge[1] Han Yang[3] Wei Liu[4] Ping Luo[1]
[1]The University of Hong Kong [2]Tencent AI Lab
[3]ETH Zürich [4]Tencent Data Platform
{rhettgee, yuyingge}@connect.hku.hk yibingsong.cv@gmail.com hanyang@ethz.ch
wl2223@columbia.edu pluo@cs.hku.hk

Abstract

Image virtual try-on replaces the clothes on a person image with a desired in-shop clothes image. It is challenging because the person and the in-shop clothes are unpaired. Existing methods formulate virtual try-on as either in-painting or cycle consistency. Both of these two formulations encourage the generation networks to reconstruct the input image in a self-supervised manner. However, existing methods do not differentiate clothing and non-clothing regions. A straightforward generation impedes the virtual try-on quality because of the heavily coupled image contents. In this paper, we propose a Disentangled Cycle-consistency Try-On Network (DCTON). The DCTON is able to produce highly-realistic try-on images by disentangling important components of virtual try-on including clothes warping, skin synthesis, and image composition. Moreover, DCTON can be naturally trained in a self-supervised manner following cycle consistency learning. Extensive experiments on challenging benchmarks show that DCTON outperforms state-of-the-art approaches favorably.

1. Introduction

Virtual try-on of fashion images aims at changing the clothes of a person with other in-shop clothes. There are wide applications including costume matching, fashion image editing, and clothes retrieval for e-commerce. Existing methods mainly focus on a direct try-on based on 2D images because of the available person images and in-shop clothes images online. However, these images are unpaired since the collection of images with multiple models, of which each model wears different and pixel-wise aligned clothes is infeasible.

To handle unpaired images, existing methods such as VITON [15], CP-VTON [35], CP-VTON+ [24], and ACGPN [40] formulate virtual try-on as an inpainting problem. They first mask the clothes region of a person image, and then recover the clothes region by using the same in-shop clothes for self-supervised network training. The pipeline is shown in the top row of Fig. 1. It is regarded as a one-way reconstruction from the corrupted input image to its original image. Since these methods only use one clothes during training (i.e., clothes 1 is matched to input 1), they are not effective when the person image and the target

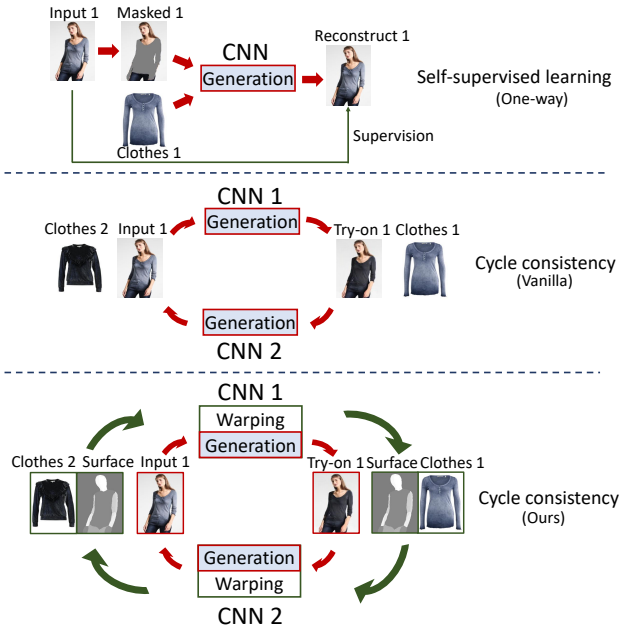

Figure 1. Comparison of virtual try-on pipelines. The inpainting methods (e.g., CP-VTON [35] and ACGPN [40]) shown in the top row use one in-shop clothes to replace the same input clothes. The vanilla CycleGAN [18] shown in the middle row introduces two in-shop clothes for cycle consistency at the expense of generating coupled image contents (i.e., clothes, skin, and human poses). In the last row, we propose DCTON to disentangle virtual try-on as clothes warping and non-clothes generation, which is built upon vanilla cycle consistency for self-supervised learning.

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

(a) Input (b) Target (c) ACGPN (d) CP-VTON (e) CA-GAN (f) DCTON

Figure 2. Virtual try-on comparisons. Inpainting based methods (ACGPN [40] and CP-VTON [35]) are not effective to establish an accurate correspondence in (c) and (d) when the target clothes are significantly different from that in input images. Meanwhile, a heavily coupled content generation (CA-GAN [18]) brings salient artifacts as shown in (e). Different from existing methods, our DCTON disentangles virtual try-on as clothes warping, skin synthesis, and image composition in a cycle consistency training configuration. The network is learned to produce highly-realistic try-on results as shown in (f).

in-shop clothes are significantly visually different. Examples are shown in Fig. 2(c) and (d), where clothes with long sleeves will be changed to those with short sleeves. The arm region is not accurately generated as shown in the first row. Meanwhile, there are large artifacts on the skirts in the second row. Besides these observations, these methods utilize separate modules for virtual try-on such as thin plate splines (TPS) [9] warping and semantic prediction. Their performance is limited due to a lack of end-to-end training for network potential exploitation.

Apart from the above inpainting-based methods, CA-GAN [18] incorporates cycle consistency for end-to-end network training. As shown in the middle row of Fig. 1, CA-GAN substitutes the clothes of an input person image (i.e., input 1) with an arbitrary target in-shop image (i.e., clothes 2). This network design improves correspondence matching between the person image and arbitrary target clothes. Nevertheless, it is still challenging to simultaneously generate the shape and the texture of clothes, the human skin, and the non-clothing contents in a cycle generative adversarial network (GAN). As shown in Fig. 2(e), artifacts appear around the arms and the logo region. This indicates a straightforward generation via cycle consistency training is insufficient for high quality virtual try-on.

In this paper, we address aforementioned limitations by proposing a disentangled cycle-consistency try-on network (DCTON). It disentangles virtual try-on into three submodules. The first one is clothes warping module that preserves clothes design (e.g., collar style, sleeve cutting, and logo). The second one is skin synthesis module for occluded human body part generation (e.g., the arm of the blouse and vest in Fig. 2). The third one is image composition module for output image generation. During training, DCTON disentangles these three components from input images to constitute a try-on cycle for self-supervised learning. Extensive experiments on the benchmark datasets show that DCTON performs favorably against state-of-the-art virtual try-on approaches.

2. Related Work

In this section, we review the literature of virtual try-on and cycle consistency for image generation.

2.1. Virtual Try-on

Studies on virtual try-on derive from fashion editing [28, 14, 48, 23] for efficient clothes substitution. The computer graphics model [46] and dimensionality reduction technique [10] are first developed for try-on generation. With the development of CNNs [32, 33], learning based methods evolve significantly. These methods can be categorized as 3D-based [12, 27, 25, 41] and 2D-based [18, 15, 35, 40] methods. Due to the lightweight data collection, 2D methods suite real-world scenarios and thus become popular.

However, training a 2D-based try-on model is still challenging due to a lack of paired triplet data [15, 7] (i.e., a reference person, a target in-shop clothes, and the person wearing this clothes). Inspired by self-supervised learning, prior arts address this issue either in a one-way reconstruction [15, 35, 24, 44, 40, 13] or a vanilla cycle consistency

generation [18]. For the one-way scheme, methods such as VITON [15], CP-VTON [35] and CP-VTON+ [24] first mask the region of both clothes and limbs, and then refill this region with either the same input clothes or the generated skin. These methods do not perform well when the target clothes is significantly different from that in the input images. Also, a lack of end-to-end training limit their generalization potential.

The cycle consistency structure is employed in CA-GAN [18] for virtual try-on. By feeding the generator with shuffled training samples (i.e., the reference person and an arbitrary clothes), CA-GAN improves clothes characteristics preserving while bringing undesirable artifacts in texture and body generation. This is because the generation of both clothes texture and occluded body parts is challenging for one network. To this end, our DCTON disentangles virtual try-on as clothes warping, skin synthesis, and image composition within a cycle consistency framework to produce highly-realistic try-on images.

2.2. Cycle Consistency for Image Generation

The self-supervised learning of cycle consistency introduces pixel-wise supervision for unpaired image-to-image generation [2, 20, 34, 45]. In [47], a CycleGAN framework is proposed for unpaired image synthesis. The DualGAN is proposed in [43] for image quality improvement. The relationships between different domains are explored in [20] based on the cycle consistency.

Cycle consistency learning has been applied to many applications including image style transfer [4, 31], object tracking [36, 37], and photo enhancement [5, 42]. However, cycle consistency learning is not effective when handling person image generation [22, 29], pose-guided animation [3], image restoration [21, 38], and virtual try-on [15]. Inspired by the cycle consistency scheme [47], we reformulate the try-on task as a conditional unpaired image-to-image generation problem. The try-on result is conditionally generated by the images of the reference person and the target clothes. A straightforward cycle consistency is not effective for try-on as the generation of both clothes texture and occluded human parts is challenging. In this work, we disentangle try-on to several sub-modules for high-quality results production.

3. Proposed Method

We disentangle virtual try-on as clothes warping, skin synthesis, and image composition within the cycle consistency framework. Three encoders are utilized for the disentanglement. Fig. 3 shows an overview of our pipeline. In the following, we first illustrate each component of the disentanglement in Sec. 3.1. Then, the cycle consistency training will be presented in Sec. 3.2 to empower networks for highly-realistic try-on generation.

3.1. Disentangled Virtual Try-on

We use the subscript 1 to illustrate the image contents related to the input clothes, and subscript 2 to denote the image contents related to the target clothes. Specifically, we denote the input image as I_1, the in-shop target clothes image as C_2, the skin region of the input image as S_1, respectively. On the other hand, the in-shop clothes of the input image is denoted as C_1, the skin region of the output image is denoted as S_2, and the output image is denoted as I_2. These notations will be used to present the process of disentanglement.

3.1.1 Clothes Warping

There are two sequentially-connected encoder-decoder networks and one encoder in the clothes warping module. We use the Densepose descriptor [1] to extract the human surface representation of the input image I_1, which is denoted as D. Then we send D and C_2 into an encoder-decoder network named as MPN (mask prediction network). The MPN will produce the mask of the clothes region (i.e., M_1^{clothes}) and skin region (i.e., M_1^{skin}) of the input image, which are used as the prior guidance for further warping and generation, respectively. We train MPN with the supervision from the parsing labels of I_1 via the pixel-wise L_1 loss on each corresponding mask region. Note that different from previous works, we adopt the Densepose descriptor for human representation since it provides both the key point positions and semantic parsing results (e.g., body and arm shape), while vanilla 2D pose estimators can only provide the key point positions. The semantic parsing results improve our model to become sensitive around the human shapes for clothes fitting and characteristics generation.

After obtaining M_1^{clothes}, we send it together with C_2 to the second encoder-decoder network, which is denoted as STN (spatial transformer network) [17]. The STN will warp C_2 according to the guidance from M_1^{clothes}. Specifically, STN first produces a transformation matrix T and guides this matrix via Thin-Plate Spline (TPS) [9] (i.e., $\mathcal{T}$) to warp the clothes image C_2. After obtaining the warped target clothes C_2^{warp} and the skin region M_1^{skin}, we use an encoder to extract their pyramid features to further concatenate with other encoded features for output generation. The parameters of STN are kept fixed during the cycle consistency training. We pretrain the STN by only using the in-shop clothes image C_1 and the input image I_1. The loss function of STN can be written as:

$$L_a = ||\mathcal{T}(C_1) - I_1 \odot M_1^{\text{clothes}}||_1, \tag{1}$$

where M_1^{clothes} is the mask region of the input image that is given by the parsing result, and $\odot$ is the element wise multiplication operation.

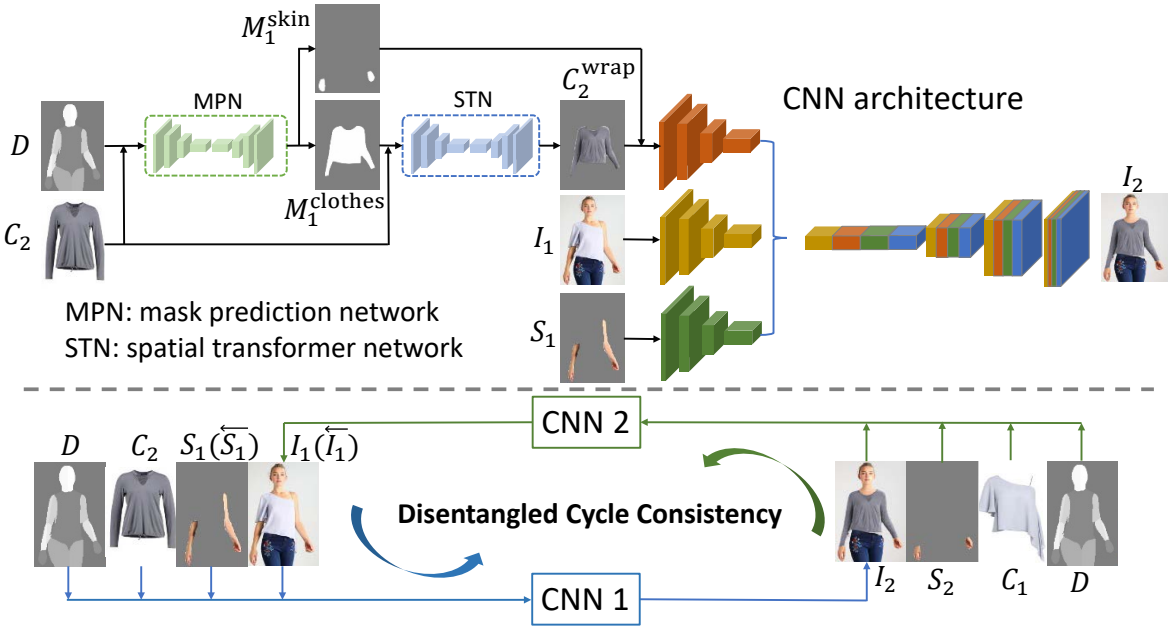

Figure 3. The pipeline of our disentangled cycle consistency framework. We show the CNN architecture above where there are clothes warping, skin synthesis, and image composition modules. The encoded features from these modules are concatenated to decode the output image. The cycle consistency is shown below where we use two CNNs with the same architecture. We send the output image from one CNN to another CNN as input to constitute self-supervision for end-to-end learning.

Due to the huge variation of poses in the real-world try-on scenario, the original transformation matrix T may not be effective enough to produce stable $\mathcal{T}$ during training. Simply adopting the STN is not capable of dealing with the large misalignment and deformation, thus bringing artifacts on the warped clothes C_2^{warp}. We further incorporate a regularization term to robustly produce T. In practice, we first introduce a homography matrix H to reduce the variations of T. For the n-th training iteration, we construct an objective function as:

$$R_b = ||H \times T^{n-1} - T^n||^2, \tag{2}$$

where T^{n-1} is from the $(n-1)$-th iteration. We can use SVD [11] to solve the Homogeneous Linear Least Squares problems as well as optimize H, and use the optimized H to compute Eq. (2) as a regularization term. As a result, the whole loss function to pretrain STN can be written as:

$$L_{\text{STN}} = L_a + R_b, \tag{3}$$

where R_b regularizes the transformation matrix T during STN training. To this end, we have successfully disentangled the clothes warping via a sequential network.

3.1.2 Skin Synthesis

The skin synthesis aims to recover the occluded human body regions during try-on. We extract the skin region of the input image (i.e., S^1) by using the input surface D. After obtaining S_1, another encoder branch is exploited to capture its pyramid feature representations. The encoder we use contains the same architecture as that in Sec. 3.1.1. The encoded features of S^1 are concatenated with other encoded features at each feature level to represent the output image I_2 in the CNN feature space.

3.1.3 Image Composition

After obtaining the encoded feature representations of warped clothes C_2^{warp} and skin image S_1, we send the input image I_1 into an encoder for global image representation. The encoder structure is the same as the other two encoders. We then concatenate the encoded features of C_2^{warp}, S_1, and I_1 sequentially and send them into the decoders for output image I_2 generation. To this end, we perform clothes warping, skin synthesis, and image composition in three independent modules and fuse their feature representations to produce the try-on result.

3.2. Cycle Consistency Training

Fig. 3 shows the cycle consistency construction. We generate a try-on result I_2 given an input image I_1 with its skin region S_1, target clothes C_2, and Densepose descriptor D. In return, we use the generated try-on results I_2, the skin region brought by the target clothes S_2 (i.e., $M_1^{\text{skin}} \odot I_2$), and the target clothes C_1 and D as the input to generate

an inversely predicted input image $\overleftarrow{I}_1$. Note that during the training process, the designed networks CNN_1 and CNN_2 in Fig. 3 share the same architectures. The cycle consistency will be established by enforcing $\overleftarrow{I}_1 \approx I_1$ to formulate self-supervision. We further illustrate the loss functions during the cycle consistency training as follows:

Adversarial Loss. We introduce two discriminators $\mathcal{D}_p$ and $\mathcal{D}_s$ during the adversarial loss $\mathcal{L}_{adv}$ computation stage. The learned generators will synthesize a target try-on image I_2, an inversely predicted input image $\overleftarrow{I}_1$, a target skin image S_2, and an inversely predicted input skin image $\overleftarrow{S}_1$. We expect the appearance of both $\overleftarrow{I}_1$ and I_2 is similar to that of I_1, and the appearance of both $\overleftarrow{S}_1$ and S_2 is similar to that of S_1. The loss function can be written as follows:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{I_2,\overleftarrow{I}_1}[log(D_p(I_2)\cdot D_p(\overleftarrow{I}_1))]+ \\ & \mathbb{E}_{S_2,\overleftarrow{S}_1}[log(D_s(S_2)\cdot D_s(\overleftarrow{S}_1))]+ \\ & \mathbb{E}_{I_1,S_1}[log((1-D_p(I_1))\cdot(1-D_s(S_1)))],\end{aligned} \tag{4}$$

where $\overleftarrow{S}_1$ indicates the generated skin of $\overleftarrow{I}_1$.

Cycle Consistency Loss. In addition to the adversarial loss that ensures similar appearance distributions between the try-on results and the target images, we propose the cycle consistency loss to improve the pixel-wise self supervision. The cycle consistency loss term is based on ℓ_1 on the synthesized try-on results and the corresponding skin regions, respectively. It can be written as follows:

$$\mathcal{L}_{cyc} = \big|\big|\overleftarrow{I}_1 - I_1\big|\big|_1 + \big|\big|\overleftarrow{S}_1 - S_1\big|\big|_1. \tag{5}$$

Content Preserving Loss. For the contents within the human region excluding the skin and clothes regions, we aim to identically preserve them in the output try-on results. To this end, we design a content preserving loss term which measures the similarities between I_1 and $\overleftarrow{I}_1$, and I_1 and I_2 within this region. The loss term can be written as follows:

$$\mathcal{L}_{pre} = \big|\big|M\odot(I_2 - I_1)\big|\big|_1 + \big|\big|M\odot(\overleftarrow{I}_1 - I_1)\big|\big|_1, \tag{6}$$

where $M = 1 - M_1^{\text{skin}} - M_1^{\text{clothes}}$ denotes the mask of the human body excluding the clothes and skin.

Perceptual Loss. We utilize the perceptual loss [30] to ensure similar CNN feature representations between the warped clothes. This improves the correspondence accuracy during clothes warping. The perceptual loss can be written as:

$$\begin{aligned}\mathcal{L}_{vgg} = \sum_{l=1}\frac{1}{W_l H_l C_l}(&\big|\big|\phi_l(C_2{}^{\text{wrap}} - M_1{}^{\text{clothes}}\odot I_2)\big|\big|_1 \\ &+\big|\big|\phi_l(C_1{}^{\text{wrap}} - M_2{}^{\text{clothes}}\odot \overleftarrow{I}_1)\big|\big|_1),\end{aligned} \tag{7}$$

where ϕ_l denotes the feature of the l-th layer in VGG19 [19], and W_l, H_l, C_l are the spatial parameters of the corresponding CNN features.

Objective Function. Our final objective function consists of all the aforementioned loss terms and can be written as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{pre}\mathcal{L}_{pre}, \tag{8}$$

where λ_{cyc}, λ_{vgg} and λ_{pre} are the constant scalars balancing the contributions from these loss terms.

4. Experiments

In this section, we illustrate the benchmark datasets, implementation details, evaluation results, and ablation studies. The datasets we use are VITON and VITON-HD.

VITON. There are 19k image groups in this dataset. Each image group contains a frontal view of a model and an in-shop clothes image. We follow [35] to exclude 2747 invalid image groups, and thus maintain a training set consisting of 14,221 groups and a testing set consisting of 2,032 groups.

VITON-HD. The images in this dataset are the same to those of VITON but with a higher resolution of 512×384. The VTION-HD dataset is more challenging since the results are in higher resolution where artifacts are more obvious on the try-on results.

4.1. Implementation Details

Architectures. Our network consists of four independent encoders, two decoders, and one pre-trained STN network. The architectures of the encoders and the decoders are from the Res-Unet [6], and the corresponding discriminators are from PatchGAN [16]. There are five convolutional layers with a stride number of 2 and two residual blocks in each encoder. The decoder in MPN and the decoder used to generate the try-on results both contain five deconvolutional layers. The number of filters for the convolutional layers is 64, 128, 256, 512, 512 in each encoder, and 1536, 2048, 1024, 512, 256 in the final decoder used to output the try-on results, respectively. The STN is an encoder-decoder where the encoder consists of 5 convolutional layers with a stride number of 2. Each convolutional layer is followed by a max-pooling layer.

Figure 4. Visual evaluation on the VITON dataset. Compared to existing methods [35, 24, 40, 18], our DCTON is effective to preserve human body characteristics, and clothes textures, and generate occluded human body parts. These advantages enable DCTON to generate highly-realistic try-on results.

Training and Testing. We pretrain an STN network with paired data (i.e., the clothes region of the in-shop clothes image and the try-on results) by using the objective function in 3.1. Then, we train DCTON end-to-end with the input of the model, the segmented skin, the Densepose descriptor, and the random in-shop clothes. DCTON is trained under 100 epochs. The parameter values of λ_{cyc}, λ_{vgg}, and λ_{pre} are all set as 10. The initial learning rate is set to be 0.0002 and the model is optimized by the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. During testing, we only use CNN_2 shown in Fig. 3 for online inference. The inputs to the network are the same as those during training.

4.2. Qualitative Evaluations

We compare DCTON to inpainting based one-way reconstruction methods CP-VTON [35], CP-VTON+ [24] and ACGPN [40], and the vanilla cycle consistency method CA-GAN [18]. Fig. 4 shows the evaluation results. In the first row, we aim to indicate the clothes characteristics preserving ability of these methods. The target in-shop clothes and the input image clothes are significantly different. Existing methods do not attend to the target clothes and fit this clothes to the clothes region of the input image. To this end, limitations occur around the collar, sleeves, and the clothes boundaries. These limitations are solved by our DCTON where the target in-shop clothes are arbitrary during training. We use various clothes to train DCTON with a high generalization ability.

In the second row, we aim to show the texture transfer ability of these methods. There are blur and distortions in the results generated by CP-VTON and CP-VTON+. Although these limitations are alleviated in the result of ACGPN, the whole clothes content is incorrectly generated. Compared with the vanilla cycle consistency method CA-GAN, DCTON is able to preserve the subtle embroiderer. Moreover, the subtle clothes texture is well preserved without distortion, due to the accurate clothes warping from STN.

In the third and last rows, we aim to indicate that whether existing methods maintain the non-clothes regions. The one-way inpainting methods are not effective for detail preservation (i.e., skirts in the third row). Moreover, there are limitations for these methods when generating occluded body parts including peculiar upper limbs, necks and hands. From these examples, we conclude that the one-way inpainting methods bring blur on human bodies and clothes boundaries. They are not effective to preserve the target clothes characteristics (e.g., the collars and sleeves). This limitation is partially alleviated in CP-VTON+ and ACGPN. However, without using arbitrary clothes during training, the incorrect content generation around occluded human bodies occurs. The CAGAN uses cycle consistency to attend to clothes characteristics while the subtle textures are ignored. In comparison, we use disentangled cycle consistency during training. The learned DCTON is able to generate highly-realistic try-on results. The challenging factors including clothes textures warping, characteristics preserving, and occluded human body generation are effectively solved.

Besides evaluations on VITON, we show visual results

Figure 5. Visual evaluations on the VITON-HD dataset. Our DCTON is effective to generate try-on results under a higher resolution. The results of existing methods are upsampled in this figure.

on VITON-HD in Fig. 5. The VITON-HD dataset is more challenging for virtual try-on because the details are more obvious and artifacts are more salient. Nevertheless, our DCTON is effective to generate highly-realistic try-on results. Compared to existing methods, DCTON preserves the target clothes characteristics as shown around the collar regions in the first row. Meanwhile, DCTON is advantageous to generate occluded body parts (i.e., the arm region in the second row). Overall, our DCTON is effective for virtual try-on under such resolution where existing methods do not attempt. The results from existing methods in this figure are upsampled for a direct view comparison.

4.3. Quantitative Evaluations

We use the Fréchet Inception Distance (FID) [8] and Structural SIMilarity (SSIM) [39] metrics to measure the similarity of data distributions between the generated try-on results and the reference image (i.e., the reference person image). For a comprehensive comparison, Inception Score (IS) [26] is also utilized to measure the perceptual quality of synthesized images. To make the fair comparison, the quantitative results generated by different methods are evaluated under the same configurations.

Table 3 shows the SSIM, IS and FID scores by CA-GAN [18], VITON [15], CP-VTON [35], CP-VTON+ [24], and ACGPN [40]. The IS results indicate that our DCTON outperforms CA-GAN, VITON, CP-VTON, CP-VTON+, and ACGPN by 0.29, 0.56, 0.26, 0.10, and 0.16, respectively. In the SSIM metric, our DCTON surpasses these methods by 0.09, 0.09, 0.11, 0.08, and 0.02, respectively. The lower FID score usually brings higher quality of the synthesized images. As such, our DCTON performs favorably against other methods. Note that even in the challenging VITON-HD dataset, our DCTON also brings considerable improvement. These results show the effectiveness and robustness of our method.

Table 1. The comparison of different methods under IS [26], SSIM [39] and FID [8] metrics. For IS and SSIM, the higher is the better. For FID, the lower is the better. DCTON* denotes the DCTON without the skin synthesis encoder. And we use DCTON$^\diamond$ to indicate the DCTON without the regularization term in STN.

Methods	Dataset	IS [26]↑	SSIM [39]↑	FID [8]↓
CA-GAN [18]	VITON	2.56 ± 0.09	0.74	47.34
VITON [15]	VITON	2.29 ± 0.07	0.74	55.71
CP-VTON [35]	VITON	2.59 ± 0.13	0.72	24.45
CP-VTON+ [24]	VITON	2.75 ± 0.14	0.75	21.04
ACGPN [40]	VITON	2.69 ± 0.12	0.81	16.64
DCTON*	VITON	2.81 ± 0.14	0.74	18.12
DCTON$^\diamond$	VITON	2.80 ± 0.23	0.79	15.70
DCTON	VITON	$\mathbf{2.85 \pm 0.15}$	**0.83**	**14.82**
DCTON	VITON-HD	$\mathbf{2.84 \pm 0.10}$	**0.81**	**15.55**

Besides the high-quality visual performance, DCTON is also advantageous by using less computational resource. We show the computational costs of ACPGN [40] and DCTON in Table 3. Under the same dataset (VITON) and hard-

Figure 6. Ablation study on the effect of the skin synthesis encoder. S-e denotes the skin encoder. Without the prior features guidance provided by the skin synthesis encoder, DCTON* is not capable of generating the realistic human skin.

Table 2. User study on the VITON test set. The ratio values indicate the percentages of subjects preferring DCTON.

Methods	CA [18]	VI [15]	CP [35]	CP+ [24]	AC [40]
DCTON	87.68%	80.32%	85.84%	79.82%	79.29%

Table 3. Time cost and computational complexity analysis.

Methods	Dataset	Training Time	#Params	FLOPS	FPS
ACGPN [40]	VITON	$\sim 40h$	139M	206G	10
DCTON	VITON	$\sim 44h$	153M	**194G**	**19**

ware configurations (8 Nvidia Telsa V100 GPUs), the training time of DCTON is similar to that of ACGPN. Under only 1 V100 GPU, the online inference speed of DCTON is almost twice faster than that of ACPGN. We also analyze the model parameters and FLOPs in Table 3. DCTON contains more parameters while taking less FLOPs. The nearly real-time generation speed of DCTON (i.e., 19 FPS on 1 V100 GPU) is suitable for the online cloud service.

4.4. User Study

The quantitative evaluation metrics are not sufficient to reflect the visual quality of the images as they measure the overall distributions of two image sets. To further evaluate the performance of existing methods, we conduct a user study where there are over 50 subjects. To make a fair comparison, 200 images from the VITON test set have been randomly selected for each method. A total of 1000 groups of generated images are provided for the user study on five comparing methods. The evaluation guidance is to consider the overall perceptual quality as well as fine-grained texture details. Each subject is randomly assigned with 100 image groups to select which result is better. Each image group contains a reference person, a target clothes, the generated results from DCTON, and another method for comparison. The results in Table 2 show that our DCTON achieves both higher perceptual quality and better texture details.

Figure 7. Ablation study on the effect of the proposed regularization term in STN. We denote R as the abbreviation of the regularization term. Without the regularization, STN will fail in warping the detailed textures.

4.5. Ablation Study

We validate two components of DCTON (i.e., the generating module and warping module) in the ablation study. We use DCTON* to indicate DCTON without the skin synthesis encoder, and DCTON$^\diamond$ to indicate DCTON without the regularization term in STN. We first assess the effects of the skin synthesis encoder. Quantitative results in Table 3 show that after removing the skin encoder, the performance of DCTON* will decrease but is still better than other methods by a margin. The visual comparison presented in Fig. 6 shows that DCTON* tends to generate the skin either with peculiar colors or blurring. An experiment is also performed to validate the proposed regularization term in STN. As shown in Fig. 7, clothes with an obvious logos or embroiderer are presented as examples. From the first row in Fig. 7, the STN module without the extra assistance of the proposed regularization term is prone to output the clothes with obvious distortion on the clothing texture. The second row in Fig. 7 shows that the regularization term facilitates STN to warp the target clothes in a proper manner.

5. Concluding Remarks

Virtual try-on methods typically consist of either a one-way reconstruction scheme or a vanilla cycle consistency configuration. However, limitations still exist when these methods generate photo-realistic try-on results. The one-way reconstruction scheme hinders existing methods from sufficient training, while the vanilla cycle consistency methods lack the texture preservation ability. In this paper, we proposed DCTON to disentangle virtual try-on as clothes warping, skin synthesis, and image composition. These modules are integrated within one framework for end-to-end cycle consistent training. Extensive experimental results validate that our DCTON achieves favorable performance compared to state-of-the-art virtual try-on approaches.

Acknowledgement. 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *IEEE/CVF International Conference on Computer Vision*, 2018.

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[4] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

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

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

[8] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 1982.

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

[10] Jun Ehara and Hideo Saito. Texture overlay for virtual clothing based on pca of silhouettes. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2006.

[11] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear algebra*. 1971.

[12] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics*, 2012.

[13] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[14] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[15] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Neural Information Processing Systems*, 2015.

[18] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2017.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 2017.

[21] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, 2020.

[22] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Neural Information Processing Systems*, 2017.

[23] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[24] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[25] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 2017.

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Neural Information Processing Systems*, 2016.

[27] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *International Conference on 3d Body Scanning Technologies*, 2014.

[28] Wu Shi, Tak-Wai Hui, Ziwei Liu, Dahua Lin, and Chen Change Loy. Learning to synthesize fashion textures. *arXiv preprint arXiv:1911.07472*, 2019.

[29] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

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

[31] Yibing Song, Linchao Bao, Shengfeng He, Qingxiong Yang, and Ming-Hsuan Yang. Stylizing face images via multi-

ple exemplars. *Computer Vision and Image Understanding*, 2017.

[32] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[33] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[34] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[35] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *European Conference on Computer Vision*, 2018.

[36] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[37] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, Wei Liu, and Houqiang Li. Unsupervised deep representation learning for real-time tracking. *International Journal of Computer Vision*, 2021.

[38] Yinglong Wang, Yibing Song, Chao Ma, and Bing Zeng. Rethinking image deraining via rain streaks and vapors. In *European Conference on Computer Vision*, 2020.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[40] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[41] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016.

[42] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating hdr transformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[43] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[44] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[45] Shuchang Zhou, Taihong Xiao, Yi Yang, Dieqiao Feng, Qinyao He, and Weiran He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. In *British Machine Vision Conference*, 2017.

[46] Zhenglong Zhou, Bo Shu, Shaojie Zhuo, Xiaoming Deng, Ping Tan, and Stephen Lin. Image-based clothes animation for virtual fitting. In *SIGGRAPH Asia 2012 Technical Briefs*. 2012.

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[48] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *IEEE/CVF International Conference on Computer Vision*, 2017.