CycleVTON: A Cycle Mapping Framework for Parser-Free Virtual Try-On

Chenghu Du[1,*], **Junyin Wang**[1,*], **Yi Rong**[1,2,3,†], **Shuqing Liu**[4], **Kai Liu**[1], **Shengwu Xiong**[1,2,3,5,†]

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, 430070
[2] Shanghai Artificial Intelligence Laboratory, Shanghai 200232
[3] Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya, 572000
[4] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, 430200
[5] School of Information Science and Technology, Qiongtai Normal University, Haikou, 571127
{duch, wjy199708, yrong, liukai356, xiongsw}@whut.edu.cn, liushuqingwtu@foxmail.com

Abstract

Image-based virtual try-on aims to transfer a target clothing onto a specific person. A significant challenge is arbitrarily matched clothing and person lack corresponding ground truth to supervised learning. A recent pioneering work leveraged an improved cycleGAN to enable one network to generate the desired image for another network during training. However, there is no difference in the result distribution before and after the clothing changes. Therefore, using two different networks is unnecessary and may even increase the difficulty of convergence. Furthermore, the introduced human parsing used to provide body structure information in the input also have a negative impact on the try-on result. How to employ a single network for supervised learning while eliminating human parsing? To tackle these issues, we present a **Cycle** mapping **V**irtual **T**ry-**O**n **N**etwork (CycleVTON), which can produce photo-realistic try-on results by using a cycle mapping framework without the parser. In particular, we introduce a flow constraint loss to achieve supervised learning of arbitrarily matched clothing and person as inputs to the deformer, thus naturally mimicking the interaction between clothing and the human body. Additionally, we design a skin generation strategy that can adapt to the shape of the target clothing by dynamically adjusting the skin region, i.e., by first removing and then filling skin areas. Extensive experiments conducted on challenging benchmarks demonstrate that our proposed method exhibits superior performance compared to state-of-the-art methods.

Introduction

Image-based virtual try-on aims to provide users with photo-realistic online try-on services. Recently, it has gained significant attention due to its immense practical and commercial value. However, two inherent challenges exist for this task. First, target clothing images are flat and must be deformed non-rigidly to fit the body's posture, thus mimicking the natural interaction between clothing and body. Second, unpaired input data lacks ground truth leading to the inability to conduct supervised learning (see Fig. 1).

In 2D space, non-rigidly deforming the clothing image to naturally fit the human pose is highly challenging. Earlier works (Han et al. 2018; Wang et al. 2018) employed a learnable network with Thin-plate Spline (TPS), which achieved clothing warping by fitting estimated control points to target points. However, it mainly focused on pixel variations around control points, leading to excessive distortion, especially in regions with complex changes such as sleeves. Although over-distortion is constrained by regularizing TPS (Yang et al. 2020), this global deformation approach can still result in significant misalignments between the clothing and the human body. An advantageous approach is the semi-rigid Moving Least Squares (MLS) (Schaefer, McPhail, and Warren 2006; Yang, Yu, and Liu 2022), which fits a local surface based on the neighborhood of control points, ensuring the best alignment with the surrounding data points. This method (Yang, Yu, and Liu 2022) effectively balances the deformation in both non-rigid and rigid regions of the clothing. However, when dealing with highly flexible non-rigid regions like sleeves, MLS struggles to achieve precise and effective alignment. Another effective method is the learnable appearance flow (AF) (Zhou et al. 2016). It estimates a dense pixel offset field by focusing on the variations of each pixel, enabling clothing deformation (Han et al. 2019). However, excessively flexible AF can still lead to excessive deformation. In addition, the above-mentioned methods take only semantic information as input and lack ground truth flow; therefore, they cannot perceive the spatial and depth

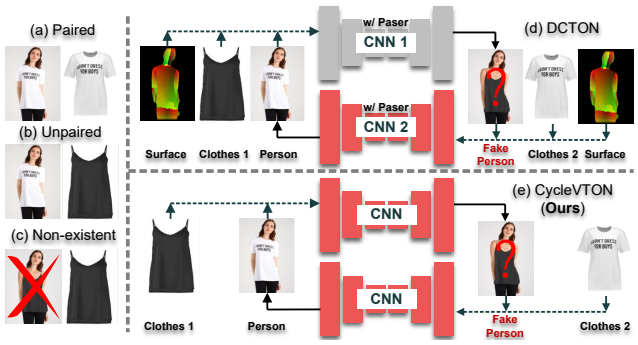

Figure 1: Comparison of (a, b) actual dataset cases and (c) non-existent dataset case. Comparison of (d) cycle consistency pipeline, DCTON, and (e) our cycle mapping pipeline.

*These authors contributed equally.
†Corresponding authors.

information of the human body, thereby hindering natural clothing deformation.

Regarding the second challenge, there are currently three coping architectures: inpainting-based, knowledge distillation-based, and cycle consistency-based pipelines. The inpainting-based methods (Han et al. 2018; Yang et al. 2020; Yang, Yu, and Liu 2022) achieve self-supervised training by reconstructing the removed regions (skin and clothing) of variation in the mannequin during the try-on process. These methods only repetitively learn to reconstruct the clothing corresponding to each mannequin. Therefore, they may fail when there are significant differences between the given clothing style and the mannequin's clothing style. Additionally, if there are errors in the human parsing used as input, it can also lead to failed try-on results. Then, knowledge distillation-based methods (Ge et al. 2021b; He, Song, and Xiang 2022) are proposed to eliminate this interference. They provide ground truth of unpaired input images to the student network from a teacher network pre-trained by the parser. However, their teacher-knowledge still is negatively impacted by the parser, which inevitably introduces irresponsible knowledge, leading to erroneous knowledge transfer to the student network. Another pioneering approach (Ge et al. 2021a; Du et al. 2023) is based on CycleGAN (Zhu et al. 2017) to achieve translations between different clothing, enabling one CNN to provide ground truth for another CNN (see Fig. 1 (d)). However, it was difficult for this dual-network structure to converge due to the lack of ground truth during training. Moreover, this paradigm is still subject to the negative impact of the parser, as it relies on human parsing as input.

To address the above challenges, a novel **Cycle** mapping **V**irtual **T**ry-**O**n **N**etwork (CycleVTON) is proposed in this work, which consists of two pipelines with shared weights, namely forward pipeline and reverse pipeline, as shown in Fig. 2. Recalling that the dataset only has three parts: i) an arbitrary *clothing 1*, ii) a certain *clothing 2*, and iii) a *person* wearing clothing 2, as shown in Fig. 1. In this case, unpaired *person* and *clothing 1* can serve as inputs in the forward pipeline but lack corresponding ground truth. Moreover, the *person* can be ground truth in the reverse pipeline, its input should be *clothing 2* and the *person* wearing *clothing 1*, but that is also non-existent. Therefore, we assume that generating the *person* wearing *clothing 1* in the forward pipeline provides the missing input for the reverse pipeline, thus realizing a cyclic structure; however, during training, this fails to converge.

To address this issue, we considered that when the person changes clothing, only the upper body's skin region and clothing region change, while the rest remain unchanged. Therefore, we only need to focus on generating the skin and clothing regions in the forward pipeline. Accordingly, we design: 1) A flow-constraint loss (FCL) is applied directly to supervise the flow field in the forward pipeline, which cascades two pipelines to enable the network to perceive the depth and spatial information of the human body. 2) A skin generation strategy (SGS) that cascades jointly two pipelines to supervise the generation of skin regions cleverly. Note that our framework is without the parser and does not take any person representations (e.g. parsing) as input, thereby eliminating the impact of flawed parsing.

The main contributions of our paper are as follows:

- We propose a parser-free cycle mapping framework, a novel perspective, for the virtual try-on task. It achieves translation end-to-end between different clothing using only a single network.
- We introduce a flow-constraint loss to achieve supervised learning of arbitrarily matched clothing and person as inputs to the deformer.
- We design an effective skin generation strategy to generate the missing skin region. It learns skin changes adaptively by generating skin in the reverse pipeline that is removed in the forward pipeline.

Related Work

Image-based Virtual Try-On Recently, image-based virtual try-ons have been widely discussed. Regarding the architecture's input, the person representation (pose points, human parsing, etc.) is often introduced to guide clothing deformation and try-on result synthesis. VITON (Han et al. 2018) and CP-VTON (Wang et al. 2018) utilize rough body shapes and pose point maps as person representations. ACGPN (Yang et al. 2020), DCTON (Ge et al. 2021a), and RT-VTON (Yang, Yu, and Liu 2022) introduce human parsing provided by a separate parser as person representations. WUTON (Issenhuth, Mary, and Calauzenes 2020), PF-AFN (Ge et al. 2021b), and Flow-Style (He, Song, and Xiang 2022) incorporate human parsing in their teacher network. USC-PFN (Du et al. 2023) utilizes human parsing as person representations for generating pseudo-labels of the human body. Although it has been proven that introducing additional person representations can eliminate partial artifacts and occlusions (Yang et al. 2020), flawed human parsing can bring undesirable outcomes (Ge et al. 2021b). For example, when white clothing is erroneously segmented as the white background, in the generated results, the clothing region will not change according to the appearance of the target garment but will continue to be generated as part of the background.

Cycle Mapping Framework Starting from the initial proposal of CycleGAN (Zhu et al. 2017), which allows unpaired image-to-image translation between different style domains, this framework has undergone extensive research and development. In virtual try-on tasks, CycleGAN was adapted (DCTON (Ge et al. 2021a)), which generates the clothing and skin of the person separately, and then performs synthesis cyclically. However, this dual-model structure with the parser poses significant challenges in terms of convergence and computation cost (see Fig. 1 (d)). In virtual try-on tasks, there is no transformation between different style domains, meaning the input and output of the network maintain the same distribution. Therefore, USC-PFN (Du et al. 2023) proposed a shared-weight CycleGAN architecture, which utilizes the same network for the cyclic synthesis of different try-on results. However, its architecture is complex and dispersed.

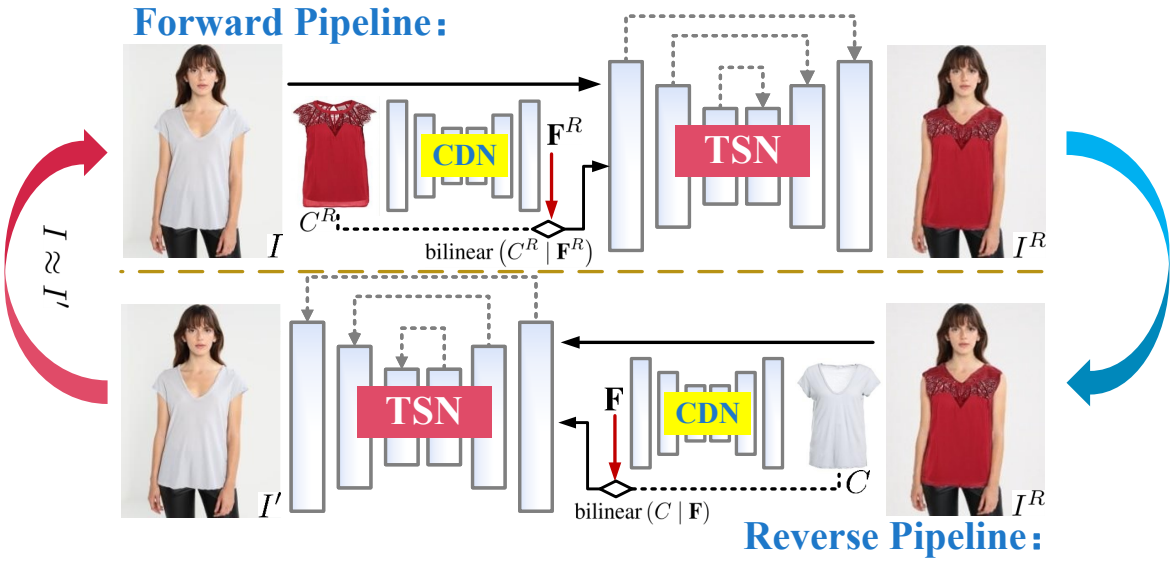

Figure 2: The overall pipeline of our CycleVTON. The person image and the clothing image are fed into CycleVTON to directly generate the try-on result.

Method

Overview

Given a reference person image ($I \in \mathbb{R}^{3\times H\times W}$) and an arbitrary clothing image ($C^R \in \mathbb{R}^{3\times H\times W}$), the virtual try-on task aims at synthesize a photo-realistic person ($I^R \in \mathbb{R}^{3\times H\times W}$) wearing the clothing C^R, *i.e.*, the desired try-on result I^R. In this paper, our work is formulated in general terms:

$$I^R = \underset{\langle C^R, I\rangle \in \mathcal{D}}{\text{CycleVTON}} < C^R, I >, \tag{1}$$

where $\mathcal{D}$ is training set. As illustrated in Fig. 2, our CycleVTON is composed of two pipelines: a forward pipeline and a reverse pipeline. Each pipeline is comprised of a clothing deformation network (**CDN**) and a try-on synthesis network (**TSN**). The CDN and TSN in the forward pipeline and those in the reverse pipeline share weights separately.

End-to-End Cycle Mapping Training

The CycleVTON is trained end-to-end. However, due to the lack of ground truth in the forward pipeline (see Fig. 1), some process differences between the forward and reverse pipelines exist during training.

Forward Pipeline of Framework In the forward pipeline, we take the person I and the arbitrary clothing C^R that exists in the dataset as inputs, to generate the try-on result I^R that the person I wearing the clothing C^R.

First, I and C^R are fed into CDN (ψ_w) to estimate a deformation field ($\mathbf{F}^R \in \mathbb{R}^{2\times H\times W}$). Then, bilinear interpolation is utilized to deform C^R based on $\mathbf{F}^R$, thereby obtaining the deformed clothing $\hat{C}^R$, formulated as:

$$\hat{C}^R = \text{bilinear}\left(C^R \mid \mathbf{F}^R\right), \ \mathbf{F}^R = \psi_w\left(C^R, I\right). \tag{2}$$

Since (C^R, I) lacks corresponding ground truth in the dataset, $\mathbf{F}^R$ and $\hat{C}^R$ cannot be optimized directly through supervised learning (Zhou et al. 2016). Following (Du et al. 2023), we pre-generated a pseudo ground truth field $\bar{\mathbf{F}}^R$ to fit the real ground truth.

Flow-Constraint Loss (FCL). Different from (Du et al. 2023), which employs an auxiliary network to correct the pre-trained deformer. In order to simplify the structure to facilitate end-to-end training, we utilize two networks to construct the flow constraint loss to constitute the supervision learning for this pipeline.

First, we employ C (paired with I) and Densepose descriptor (Guler, Neverova, and DensePose 2018) I_D of I, which contain human spatial information as inputs to train a Res-UNet ω (Diakogiannis et al. 2020). Meanwhile, we take paired C and I as inputs to simultaneously train another Res-UNet τ. Due to their deformation fields are identical in an ideal state, they are constrained by the L1 norm during the training phase to minimize their differences as much as possible.

Here, due to the limited shape and structural information contained in I_D, ω relying on I_D as input is unable to establish dense correspondences between I and C. τ takes C and I as inputs, so the feature space of τ includes the rich appearance (color, shape, and texture), depth, lighting, and other information between I and C. However, the appearance information therein may have a negative impact, namely, τ may fail when encountering different clothing. Therefore, we utilize the aforementioned structure to supplement all information of τ except for appearance into ω, aiming to enhance the accuracy of deformation.

To supervise the flow $\mathbf{F}^R$ to optimize CDN based on pre-trained ω^*, following (Xie et al. 2023), we estimate three deformation fields to deform different garment parts sepa-

rately, represented as:

$$\mathcal{L}_{fc} = \sum_{i=1}^{M} \left\| \mathbf{F}_i^R - \bar{\mathbf{F}}_i^R \right\|, \tag{3}$$

where $\bar{\mathbf{F}}^R = \omega^*(C^R, I_D)$, M = 3. $\| \ \|$ denotes L1 norm.

Afterward, $\hat{C}^R$ and I are fed into TSN (ψ_t) to generate the desired try-on result I^R, which can be formulated as:

$$I^R = \psi_t \left(\hat{C}^R, I \right). \tag{4}$$

Similarly, ψ_t cannot be optimized directly through supervised learning. To address this problem, we considered that only the upper body region changes when the person changes clothes, while the rest remains unchanged. Therefore, we only need to focus on generating the skin and clothing regions in the forward pipeline. We first optimize ψ_t by minimizing the identity information (unchanged head, trousers, etc.) dissimilarity between I and I^R. Meanwhile, $\hat{C}^R$ is continuously optimized synchronously. It can be formulated as:

$$\mathcal{L}_{save} = \sum_{I_{id} \in I} Dis \left(I_{id}^R + \hat{C}^R, I_{id} + \hat{C}^R \right), \tag{5}$$

where $\_id$ represents the identity information. $Dis(,)$ is a distance function measuring the similarity.

Dynamic Data Interception. In order to achieve cyclic skin generation through self-supervision, we design a dynamic data interception strategy to pick out persons with more skin than the target try-on result as input during the training phase. Following (Ge et al. 2021b; Yang, Yu, and Liu 2022), we pretrain a generator capable of generating the semantic map S for the target try-on result I^R. Let P denote the semantic map of the original person I, the calculation method for the skin area ratio can be expressed as:

$$\Theta(S, P) = \frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(\frac{S}{P + \epsilon} \right)_{ij}, \tag{6}$$

where ϵ denotes a small positive constant to avoid numerical issues. H and W represent the height and width of the image, respectively. Thus, we can obtain an interception coefficient κ. When κ equals 0, the training is skipped to the next batch of data. When κ equals 1, training can proceed as usual:

$$\kappa = \begin{cases} 1, & \text{if } \Theta(S_a, P_a) < r \text{ and } \Theta(S_n, P_n) < r, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where S_a and P_a represent semantic maps of the arms, while S_n and P_n represent semantic maps of the neck. r is a controllable target skin area ratio. Due to possible errors in the semantic map, r is typically less than 1.0.

Note that, due to the random pairing of clothing and persons during the training process, ideally, there are a total of $n \times n$ combinations (n is paired sample sizes). Therefore, under this strategy, the sample size is abundant.

Skin Generation Strategy (SGS). After the data is filtered, we observed that when translating from I to I^R and back to I, ($I \rightarrow I^R \rightarrow I$), the removed excess skin regions of I during $I \rightarrow I^R$ will be replenished during $I^R \rightarrow I$. In this way, self-supervised training can be conducted on the skin regions in the forward pipeline. The skin loss $\mathcal{L}_{skin}$ at this stage is then defined as:

$$\mathcal{L}_{skin} = \sum_{I \in \mathcal{D}} Dis \left(I \odot S_a \odot P_a, \ I^R \odot S_a \odot P_a \right), \tag{8}$$

where $\odot$ denotes entry-wise multiplication. $Dis(,)$ is a distance function measuring the similarity.

Reverse Pipeline of Framework In this pipeline, the inputs of CDN and TSN have been generated in the forward pipeline, $i.e.$ I^R. Therefore, we take I^R and the clothing C in pairs with I as inputs, to reconstruct the person I, $i.e.$ $I \approx I'$. It can be formulated as:

$$I' = \psi_t \left(\hat{C}, \ I^R \right), \tag{9}$$

where

$$\hat{C} = \text{bilinear}\left(C \mid \mathbf{F} \right), \ \mathbf{F} = \psi_w \left(C, I^R \right). \tag{10}$$

Since $\hat{C}$ and I' have the ground truth in this pipeline, ψ_t and ψ_w can be directly optimized through self-supervised learning. They can be formulated as:

$$\mathcal{L}_{clt} = \sum_{\bar{C} \in I} Dis\left(\hat{C}, \bar{C} \right), \quad \mathcal{L}_{try} = \sum_{I \in \mathcal{D}} Dis\left(I', I \right), \tag{11}$$

where $\bar{C}$ is the clothing region segmented from I. $Dis(,)$ is a distance function measuring the similarity.

In the forward pipeline, similar to (Du et al. 2023), the deformation effect of clothing is constrained by the performance of ω (FCL), meaning that the effect of CDN will never exceed ω because ω provides prior knowledge to CDN. Unlike (Du et al. 2023), in the reverse pipeline, we introduce $\mathcal{L}_{clt}$ to further optimize CDN, which can break through the limitation imposed by ω, as there is ground truth available in the reverse pipeline to optimize CDN. As shown in Table 1, we experimentally verify the above statement, which also indirectly validates the effectiveness of our proposed cycle mapping framework in optimizing clothing deformation performance.

Learning objectives Corresponding to notation $Dis(,)$, the pixel-wise $\mathcal{L}_1$ loss and VGG perceptual loss $\mathcal{L}_{per}$ (Johnson, Alahi, and Fei-Fei 2016) are combined to measure the similarity.

Clothing Deformation Network (CDN). The flow-constraint loss is only used to optimize CDN. The loss for the forward pipeline can be represented as: $\mathcal{L}_{FC} = \lambda_{fc}\mathcal{L}_{fc}$. Corresponding to Equation (11), $\mathcal{L}_{fc}$, $\mathcal{L}_{clt}$, and additional flow regularization loss $\mathcal{L}_{reg}$ (Minar et al. 2020) are adopted to optimize CDN in the reverse pipeline, which can be represented as: $\mathcal{L}_{RC} = \lambda_{clt}\mathcal{L}_{clt} + \lambda_{fc}\mathcal{L}_{fc} + \lambda_{reg}\mathcal{L}_{reg}$. λ denotes the hyperparameter used to balance the individual sublosses and defaults to 1.

	Methods	CDP	Paired		Unpaired	
			FID↓	KID↓	FID↓	KID↓
VITON	CP-VTON	TPS	43.95	2.23	42.13	2.11
	ACGPN	TPS	42.10	2.01	41.48	2.05
	DCTON	TPS	42.80	2.17	42.19	2.13
	PFAFN	AF	22.81	0.79	23.90	0.86
	Flow-Style	AF	20.07	0.55	20.38	0.48
	FCL ($\mathcal{L}_{fc}$)	AF	19.52	0.44	20.14	0.42
	Ours	AF	**18.44**	**0.36**	**19.64**	**0.37**
HD	VITION-HD	TPS	32.97	1.41	32.93	1.35
	HR-VITON	AF	23.51	0.77	24.83	0.76
	Ours	AF	**19.15**	**0.42**	**22.51**	**0.55**

Table 1: Quantitative results of the clothing deformation module between baselines and ours (256×192). CDP represents the different clothing deformation pipelines. Note values of KID are multiplied by 10^2 for readability. The up-/down arrow next to metric indicates that the higher/lower the better. The best result is in **bold**.

Figure 3: Qualitative results of clothing deformation between Flow-Style, HR-VITON, and our method in the unpaired setting.

Try-on Synthesis Network (TSN). We introduce the adversarial loss $\mathcal{L}_{adv}$ to prevent the generation of abnormal results. Additionally, corresponding to Equation (5), the loss for the forward pipeline can be represented as: $\mathcal{L}_{FT} = \lambda_{save}\mathcal{L}_{save} + \lambda_{skin}\mathcal{L}_{skin} + \lambda_{adv}\mathcal{L}_{adv}$. Corresponding to Equation (11), $\mathcal{L}_{try}$ is used to optimize TSN in the reverse pipeline, which can be represented as: $\mathcal{L}_{RT} = \lambda_{try}\mathcal{L}_{try}$.

Overall Objectives. We employ two optimizers to separately optimize CDN and TSN. Therefore, we design two total loss functions, $\mathcal{L}_C$ and $\mathcal{L}_T$, represented as:

$$\mathcal{L}_C = \mathcal{L}_{FC} + \mathcal{L}_{RC}, \quad \mathcal{L}_T = \mathcal{L}_{FT} + \mathcal{L}_{RT}. \tag{12}$$

Experiments

Dataset

VITON We use VITON dataset (Han et al. 2018), which consists of 16,253 image groups with the resolution of 256×192. Each group includes a frontal-view woman image I, a top clothing image C paired with I, a semantic map, and a pose heatmap. The dataset is split into a training set with 14,221 groups and a testing set with 2,032 groups.

Methods	Publication	CDP	Mode	P.R.	SSIM↑	FID↓
VITON	CVPR'18	TPS	IP	Y	0.74	55.71
CP-VTON	ECCV'18	TPS	IP	Y	0.72	24.45
Cloth-flow	ICCV'19	AF	IP	Y	0.84	14.43
CP-VTON+	CVPRW'20	TPS	IP	Y	0.75	21.04
ACGPN	CVPR'20	TPS	IP	Y	0.84	16.64
DCTON	CVPR'21	TPS	CC	Y	0.83	14.82
PF-AFN	CVPR'21	AF	KD	N	0.89	10.21
ZFlow	ICCV 21	AF	IP	Y	0.88	15.17
RT-VTON	CVPR'22	MLS	IP	Y	-	11.66
SDAFN	ECCV'22	AF	IP	N	0.88	12.05
Ours	This Work	AF	**CM**	N	**0.92**	**9.41**

– : official code or data are not provided.

Table 2: Quantitative results of try-on synthesis between baselines and ours on **VITON**. Mode represents the type of pipelines. P.R. indicates whether the person representation is used during inference. The up/down arrow next to metric indicates that the higher/lower the better. The best result is in **bold**.

VITON-HD We also use VITON-HD dataset collected by (Choi et al. 2021) to demonstrate the generalization of handling high-resolution images, which comprises 13,679 image groups with the resolution of 512×384. It is also downsampled to a low resolution 256×192. All components are the same as VITON, and are split into a training set with 11,647 groups and a testing set with 2,032 groups.

Implementation Details

Our framework is implemented using PyTorch and trained on 1 Nvidia Tesla V100 GPU running Ubuntu 16.04. During training, we use the AdamW optimizer ($\beta_1 = 0.5$ and $\beta_2 = 0.999$) (Loshchilov and Hutter 2017) with a batch size of 1 and an initial learning rate of $1e^{-4}$. Our framework is iteratively optimized for 200 epochs, the learning rate is linearly reduced to 0 in the last 100 epochs. Our CycleVTON consists of CDN and TSN, which are implemented by Res-UNet (Diakogiannis et al. 2020).

Baselines

To objectively compare and evaluate the performance of our model, we leverage thirteen publicly available state-of-the-art methods as baselines. It consists of specialized high-resolution methods, including VITON-HD (Choi et al. 2021) and HR-VITON (Lee et al. 2022), and popular low-resolution methods, including VITON (Han et al. 2018), CP-VTON (Wang et al. 2018), Cloth-flow (Han et al. 2019), CP-VTON+ (Minar et al. 2020), ACGPN (Yang et al. 2020), DCTON (Ge et al. 2021a), PF-AFN (Ge et al. 2021b), ZFlow (Chopra et al. 2021), Flow-Style (He, Song, and Xiang 2022), SDAFN (Bai et al. 2022), and RT-VTON (Yang, Yu, and Liu 2022).

Evaluation Metrics

To perform the quantitative evaluation, we employ the following metrics in paired and unpaired settings:

Figure 4: Qualitative results between baseline methods and our method in the unpaired setting. Red boxes denote defects.

Unpaired Setting The unpaired setting is to generate the desired try-on result with the **unpaired** clothing-person images, $i.e.$ trying on arbitrary clothing. We take widely used Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018) as the evaluation metrics to evaluate the distribution similarity between the generated image and the real image. FID calculates the distance between two data distributions. KID measures the similarity between two sets of samples based on their kernel embeddings. It provides a more nuanced comparison of the generated samples against the real samples. A lower score for FID and KID indicates a higher quality of the result.

Paired Setting The paired setting is to reconstruct the desired try-on result with the **paired** clothing-person images, $i.e.$ trying on original clothing. We employ widely used Structure Similarity (SSIM) (Seshadrinathan and Bovik 2008) as the evaluation metric to evaluate the structure similarity between the generated image and the real image. It quantifies the degradation of structural information, color information, and luminance changes between the real and the generated images.

Qualitative Results

Clothing Deformation Comparisons To demonstrate the superior clothing deformation performance of our method compared to traditional AF-based approaches, we select the most state-of-the-art Flow-Style and HR-VITON for qualitative experiments, as shown in Fig. 3. Flow-Style may encounter occlusions due to inadequate handling capability for pixel continuity (1st and 2nd columns), and excessive distortions caused by weak space and depth perception (3rd and 4th columns). HR-VITON may encounter inadequate deformation caused by weak space and depth perception (5th columns), and inaccurate alignment due to flawed human parsing. However, our results demonstrate that the problems above are effectively mitigated through the cycle mapping strategy and flow-constraint loss. Our method can naturally warp the clothing to align with the body pose while preserving its fine details.

Try-on Synthesis Comparisons To demonstrate the effectiveness of our cycle mapping framework, we compare our method with seven SOTA baseline methods, as shown in Fig. 4. In the 3rd column, CP-VTON+ encounters severe artifacts due to arm detail loss, and excessive deformation caused by TPS. ACGPN addresses these by introducing human parsing, but the misalignment caused by TPS still remains significant. DCTON exhibits similar artifacts due to its disentangled strategy that has not considered optimizing clothing deformation. Recent SDAFN handles clothing deformation and try-on synthesis through a single network. However, its unique structure lacks the precise extraction of semantic information about the body, resulting in excessive clothing deformation and less accurate try-on synthesis, thus generating some artifacts. Both PF-AFN and Flow-Style seem to have the same issue of excessive clothing deforma-

	Methods	Mode	#Params	FLOPs	FPS
VITON	ACGPN	IP	139M	206G	10
	DCTON	CC	153M	194G	19
	PF-AFN	KD	99M	69G	34
	Ours	**CM**	**87M**	**29G**	**39**
HD	VITON-HD	IP	154M	1690G	3
	HR-VITON	IP	148M	1555G	4
	Ours	**CM**	**87M**	**468G**	**23**

Table 3: Computational complexity analysis. Mode represents the type of pipelines. The best result is in **bold**.

tion due to inadequate handling of irresponsible knowledge (inappropriate guidance and constraint) in their teacher models. RT-VTON reintroduces human parsing as input, where flawed parsing hinders the generation of high-quality try-on results, and its proposed semi-rigid clothing deformation network also struggles to effectively handle clothing deformation for complex poses. Overall, these baselines generally struggle to effectively handle the challenges of clothing deformation, and in the try-on synthesis phase, the use of irresponsible teacher knowledge or flawed human parsing results in significant artifacts and occlusions. In contrast, our designed cycle mapping framework enables mutual learning, constraint, and supervision between the two stages, resulting in highly realistic try-on results that effectively improve the above challenging factors.

Quantitative Results

Clothing Deformation Comparisons Table 1 lists the FID and KID scores between baselines and our method. In the paired setting, the FID and KID metrics outperform the best TPS-based methods (DCTON and VITON-HD) on VITON and VITON-HD datasets by 24.36, 1.81, 13.82, and 0.99, respectively. Outperforming the best AF-based methods (Flow-Style and HR-VITON) by 1.63, 0.19, 4.36, and 0.35, respectively. In the unpaired setting, surpassing the best TPS-based methods by 22.55, 1.76, 10.42, and 0.8, respectively. Outperforming the best AF-based methods by 0.74, 0.11, 2.32, and 0.21, respectively. This demonstrates that our CycleVTON significantly outperforms the baseline methods in terms of clothing warping, validating the superiority of our clothing deformation network and strategies.

Try-on Synthesis Comparisons Tables 2 lists SSIM and FID scores of baselines and our method. In the paired setting, SSIM results indicate that our CycleVTON outperforms SOTA method, SDAFN by 0.04 on VITON. In the unpaired setting, FID results indicate that our CycleVTON surpasses RT-VTON and SDAFN by 2.25 and 2.64, respectively. These results demonstrate the effectiveness and robustness of the proposed CycleVTON.

Computational Complexity Analysis In addition, to demonstrate that our method not only achieves superior visual results and performance but also has lower computational complexity and cost, we measure the parameters

Figure 5: Ablation study of the proposed CycleVTON.

(#Params), floating point operations (FLOPs), and frames per second (FPS) of the baseline methods under the identical configuration (a Tesla V100 GPU). The results show in Table 3 that the FLOPs of our method is lower than half of PFAFN, while achieving FPS with low parameters on VITON. On VITON-HD dataset, the FLOPs of our method is one-third of HR-VITON, while achieving six times the FPS of HR-VITON with nearly half of #Params. This demonstrates that our method can be applied to real-time services.

Ablation Study

Fig. 5 shows that our framework without FCL and the skin generation strategy (G_{skin}) is not working.

Effectiveness of $\mathcal{L}_{fc}$ (FCL) When we add FCL to supervise the generation of clothing in the forward pipeline, the network learns to deform any clothing onto anybody well, demonstrating the positive contribution of FCL to CycleVTON (Table 1). However, the skin suffers from neglect.

Effectiveness of G_{skin} (SGS) When we introduce the skin generation strategy, the generated body's skin corresponds to the deformed clothing, enabling the network to adapt to the body layout corresponding to any given clothing.

Conclusion

In this paper, we propose a new architecture for virtual try-on tasks, called the cycle mapping framework (CycleVTON), to generate highly photo-realistic try-on results without human parsing. We present a shared-weight framework that takes only clothing and person images as inputs to mitigate occlusions and artifacts caused by irresponsible prior knowledge and flawed human parsing. To ensure its effective convergence, we introduce a skin generation strategy for variable skin regions and a flow constraint loss to establish spatial correlations between unpaired clothing and person images. Objective experiments on popular benchmarks show the superiority of our proposed network. In the future, we plan to extend this framework to more vision tasks.

Acknowledgments

This work was in part supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160604) and NSFC (Grant No. 62176194), and the Key Research and Development Program of Hubei Province (Grant No. 2023BAB083), the Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ-JYRC-2022-76, SKJC-2022-PTDX-031), and the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031).

References

Bai, S.; Zhou, H.; Li, Z.; Zhou, C.; and Yang, H. 2022. Single stage virtual try-on via deformable attention flows. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 409–425. Springer.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Chopra, A.; Jain, R.; Hemani, M.; and Krishnamurthy, B. 2021. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5433–5442.

Diakogiannis, F. I.; Waldner, F.; Caccetta, P.; and Wu, C. 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 94–114.

Du, C.; Liu, S.; Xiong, S.; et al. 2023. Greatness in Simplicity: Unified Self-Cycle Consistency for Parser-Free Virtual Try-On. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ge, C.; Song, Y.; Ge, Y.; Yang, H.; Liu, W.; and Luo, P. 2021a. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16928–16937.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021b. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

Guler, R.; Neverova, N.; and DensePose, I. 2018. Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 18–23.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10471–10480.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7543–7552.

He, S.; Song, Y.-Z.; and Xiang, T. 2022. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3470–3479.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Issenhuth, T.; Mary, J.; and Calauzenes, C. 2020. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 619–635. Springer.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Lee, S.; Gu, G.; Park, S.; Choi, S.; and Choo, J. 2022. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 204–219. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Minar, M. R.; Tuan, T. T.; Ahn, H.; Rosin, P.; and Lai, Y.-K. 2020. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, volume 3, 10–14.

Schaefer, S.; McPhail, T.; and Warren, J. 2006. Image deformation using moving least squares. In *ACM SIGGRAPH 2006 Papers*, 533–540.

Seshadrinathan, K.; and Bovik, A. C. 2008. Unifying analysis of full reference image quality assessment. In *2008 15th IEEE International Conference on Image Processing*, 1200–1203. IEEE.

Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L.; and Yang, M. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, 589–604.

Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; and Liang, X. 2023. GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23550–23559.

Yang, H.; Yu, X.; and Liu, Z. 2022. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3460–3469.

Yang, H.; Zhang, R.; Guo, X.; Liu, W.; Zuo, W.; and Luo, P. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7850–7859.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 286–301. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.