Powering Image-based Virtual Try-on with Structural Information and Cascade Warping

Anonymous CVPR 2021 submission

Paper ID 3405

Abstract

*Image-based virtual try-on aims to transfer a garment image onto a person's image naturally. However, due to the lack of structure information, existing methods often fail to fit a garment seamlessly onto the human body while preserving its characteristics. To address this issue, we adopt new structural information, i.e. clothes parsing and DensePose, and further propose a novel framework powered by **S**tructural **I**nformation and **C**ascade **W**arping (SICW) with three components: 1) A target parsing prediction network guided by rich structural information; 2) A cascade clothes warping module built on top of TPS transformation and pixel flow; 3) A decomposed neural rendering module dealing with skin inpainting and shading refinement separately. We further construct a new dataset RAIV with **R**ich **A**nnotations for **I**mage-based **V**irtual try-on, which will be released to facilitate future research. Extensive experiments demonstrate the new structural information's effectiveness and our model's superiority over state-of-the-art methods.*

1. Introduction

Virtual try-on [12, 34, 14, 40] has attracted lots of interest due to its massive potential in e-commerce and video production. Given a reference person image and a target clothes image, virtual try-on aims at generating a new image for that person wearing the given clothes. In this way, customers could see themselves in their desired garments, which would significantly improve the online shopping experience. Traditional methods achieve this goal via graphics approaches and generate try-on images by manipulating 3D models. Nevertheless, it is expensive and impractical to craft decent models for all clothes with diverse textures and designs on sale. To this end, image-based virtual try-on has attracted lots of attention because of the availability of abundant clothing and try-on images on the Internet.

However, image-based virtual try-on remains a challenging task due to the gap between 2D information it acquires and non-rigid spatial manipulations it tries to simulate for clothes. As shown in Fig. 1, existing methods [37, 40] often fail to either (a) fit the clothing onto the human body, (b) preserve the garment's characteristics, or (c) generate natural shading. We address these issues by introducing rich structural priors and decomposing the problem into several sub-tasks to free up network capacity.

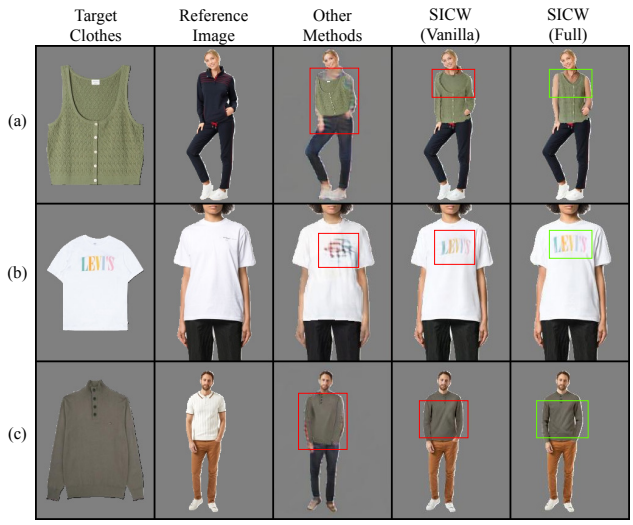

Figure 1. Representative comparisons of our model and other methods (a & c: CP-VTON [37], b: ACGPN [40]). SICW (Vanilla) indicates our method without using DensePose and clothes parsing, and SICW (full) is SICW with full structural information.

Specifically, we improve the clothing-agnostic person representation by introducing DensePose [30], which provides semantic 3D body information while eliminates the effects of clothes. We further integrate the target clothing image and its clothes parsing into a clothing representation. Moreover, we propose a new framework powered by **S**tructural **I**nformation and **C**ascade **W**arping, namely SICW. As shown in Fig. 2, it consists of three modules. Firstly, a target parsing prediction (TPP) network predicts the after-trying-on human parsing to guide subsequent

stages. Secondly, a cascade clothes warping (CCW) module warps the input clothing to fit the predicted shape. We notice that TPS transformation [1] is good at general shape matching but fails to conduct partial movements (e.g., lifting arms), while appearance flow [42] handles minor movement well but often causes distortion when dealing with large displacements. Thus, we combine them by first performing a TPS transformation for rough alignment and then using appearance flow for refinement. Finally, a decomposed neural rendering (DNR) module is used to in-paint the occluded skin area, refine the warped garment's shading, and render the final try-on result.

However, the widely used VITON [14] dataset does not meet our needs. It does not provide DensePose and clothes parsing, and the annotations it has are of poor quality. We also notice that most (1,575 out of 2,032) images in VITON's testing set also appear in its training set, which would disturb the validation of a model's generalizing ability. Thus, to facilitate our research and the community, we collect a new dataset RAIV with **R**ich **A**nnotations for **I**mage-based **V**irtual try-on. RAIV provides 16,052 high-resolution image pairs ,as well as high-quality annotations.

Our contributions can be summarized as follows: (1) We propose three improvements to image-based virtual try-on, including new structural information (DensePose and clothes parsing), cascade clothes warping for better shape fitting and character preserving, and decomposed neural rendering for skin inpainting and clothes shading refinement. (2) We construct a high-resolution and high-quality dataset RAIV, which we believe could become a new benchmark for image-based virtual try-on. We will release the dataset to facilitate the research community. (3) We conduct conprehensive module-by-module experiments, which demonstrate the effectiveness of the new structural information and the superiority of our SICW framework.

2. Related Work

2.1. Conditional Image Synthesis

Generative Adversarial Networks (GANs) [11] have achieved great success in image generation, which revolutionized many computer vision tasks, such as style transfer [44, 38, 7], object detection [22], and image inpainting [26]. Various techniques have been proposed to optimize the vanilla GAN, among which conditional GAN [23] (cGAN) becomes the most promising and widely used method in image translation. cGAN adopt both the real images and the random noises as the input of the generator, making it possible to generate images while maintaining the content of that source image. Built on cGAN, a series of autoencoder-based image translation methods are proposed to generate the realistic looking while content preserving images [4, 5, 25]. Among them, PatchGAN [23] is a widely used deblurring module to make the generated images more photo-realistic, which is also adopted in our model to better assist the skin in-painting process.

2.2. Virtual try-on

Virtual try-on allows people to see their overall appearance without actually wearing the clothes. It is appearing to both the customer and the company because of the enjoyment and convenience when trying on numerous clothes.

Most traditional methods utilized 3D based models to solve this problem. Some works [43, 6, 36, 15, 34, 2, 12, 27, 31] rely on the reconstruction and simulation of 3D clothing models to render the images, which produced reasonable try-on results. However, such 3D-based methods are usually built on the high-quality 3D reconstruction model, which leads to slow rendering speed during the test period. What's more, such models require a large amount of high-resolution 3D data, which is not accessible to ordinary customers, making it further difficult to be employed in real-world applications.

Recently, due to the superior representation ability of the deep neural networks, a series of models are proposed to solve this problem by only utilizing the 2D images. The key challenge in such image-based methods is to model the highly non-rigid transformation of the clothes. Most of the existing methods [14, 37, 41, 20, 40, 28, 33] use thin-plate spline (TPS) [1] to warp clothes. For example, VITON [14] estimates TPS parameters by shape context matching, which utilizes hand-crafted descriptors to model the correspondences between the source and the target clothing shape. A composition mask is then utilized to fuse the warped clothing with the coarse try-on result. CP-VTON [37] instead proposes a two-stage pipeline which includes clothing warping and coarse-to-fine try-on image generation. Although this approach improves the preservation quality of clothing textures, it suffers from the distortion of the body parts and clothing items that should not be affected by the warped clothes' layout. Later on, VT-NFP [41] addressed this issue by estimating a human parsing map after clothing warping, which helps preserve clothing and body details by indicating the region that should not be affected by the warping stage. ACGPN [40] further improved the preservation of the body detail by restricting the estimated human parsing with the original image's semantic layout. Such models achieve reasonable results on a set of basic transformations, however, the clothes would be heavily distorted when the human pose becomes complicated.

On the other hand, ClothFlow [13] proposed a novel approach to tackle this problem, which predicts a dense pixel-level transformation to address complex clothing deformations. However, flow-based warping still cannot handle large deformations very well due to the lack of ground truth flow. After systematically comparing and analyzing these

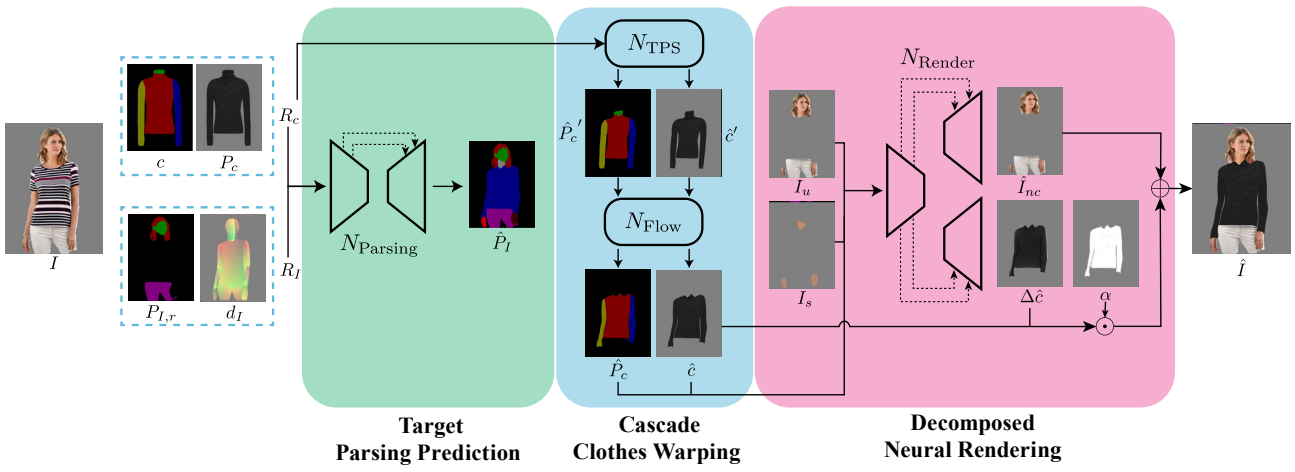

Figure 2. Model pipeline overview.

two clothing warping methods, we propose to combine the advantages of them, which leads to better shape fitting and texture preservation ability.

3. Our Approach

The task of image-based virtual try-on can be defined as follows: Given a clothing image (also known as target clothes) c and a person image (also known as reference image) I, we want to generate a photo-realistic image $\hat{I}$ of that person wearing the c without changing the pose. However, the ground truth data for supervised learning is hard to get as it requires a person to change clothes while maintaining the pose. To address this dilemma, we use obtainable *paired* I and c to construct two integrated representations (See Sec. 3.1): a clothing-agnostic person representation [14] R_I extracted from I and a clothing representation R_c extracted from c. We then let our model learn to reconstruct I from R_I and R_c.

As shown in Fig. 2, the proposed SICW consists three modules. First, the target parsing prediction (TPP) network predicts the after-trying-on human parsing from rich structural information (See Sec. 3.2). Second, the cascade clothes warping (CCW) module deforms the input clothing c to the target shape on the predicted human parsing in a coarse-to-fine manner (See Sec. 3.3). Finally, the decomposed neural rendering (DNR) module is designed to generate a photo-realistic try-on result via skin inpainting and clothes shading refinement, as described in Sec. 3.4.

3.1. Person and Clothing Representations

The clothing-agnostic person representation [14] is designed for integrating descriptive and structural information for the person in I while eliminating the effects of the clothes. It is composed initially of a pose heatmap, a blurred body shape, and reserved regions (face and hair). We improve this representation by adopting DensePose [30] and human parsing. Our enhanced clothing-agnostic representation R_I now consists of a 3-channel DensePose map and a 12-channel reserved human parsing that removes the tops and torso parsing label from the original human parsing of I.

Similarly, we construct a clothing representation R_c by integrating the 3-channel target clothing image c with its 5-channel clothes parsing P_c (explained later in Sec. 4).

3.2. Target Parsing Prediction

We first predict target human parsing, which is used to guide the clothes warping and body completion. Different from previous methods, our method takes cloting-agnositic representation R_I and clothing representation R_c as inputs. The DensePose in R_I helps to exclude the influence of original shape of clothes in the reference image, and the clothes parsing helps to distinguish different parts of the clothes. Both annotations make the prediction on tops more correct and realistic.

As for network, an U-Net [32] $N_{Parsing}$ is used to predict target human parsing, i.e. $\hat{P}_I = N_{Parsing}(R_I, R_c)$. In training procedure, we use paired data as training set. Note that the parsing of parts to be predicted is not available in inputs, the generation ability on unpaired data can be guaranteed. OHEM [35] loss and consistency loss are used in training procedure. OHEM loss calculates CE loss between P_I and $\hat{P}_I$ on only the top-k most unconfident pixels, which improves the accuracy on details significantly. Consistency loss refers to the CE loss between results given transfromed inputs and ground truth applied the same transformation, formulated as

$$\mathcal{L}_{cons} = CE(t(P_I), N_{Parsing}(t(R_I), t(R_c)))$$

where t refers to image transformation, and we use mirror transformation and affine transformation when training.

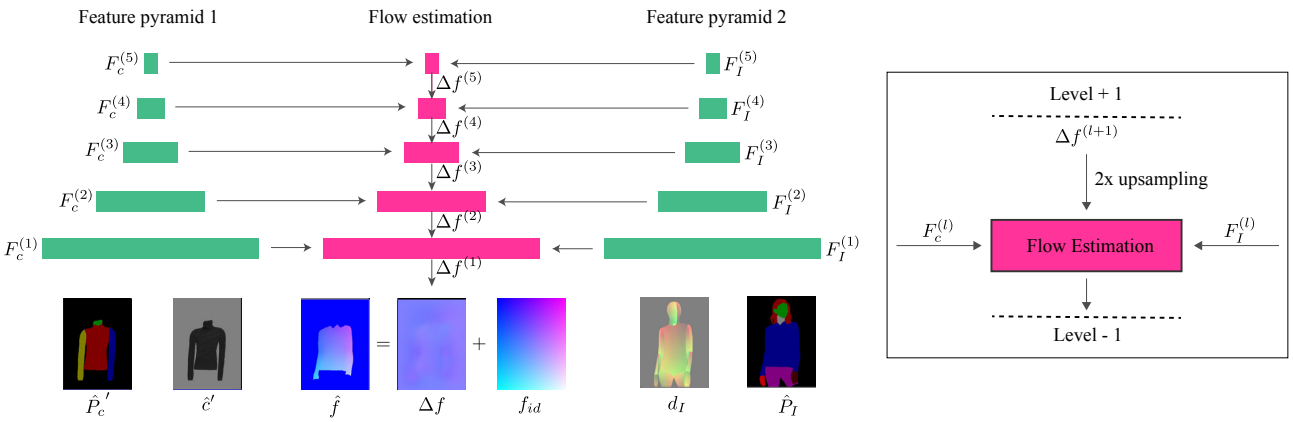

Figure 3. Flow model overview

The overall loss function can be formulated as

$$\mathcal{L} = \lambda_{ohem}\mathcal{L}_{ohem} + \lambda_{cons}\mathcal{L}_{cons} \tag{1}$$

where λs are weights of each term to control the final performance.

3.3. Cascade Clothes Warping

We then transform the clothing image c to fit the target shape $\hat{P}_{I,c}$, which is a 1-channel map extracted from the target human parsing $\hat{P}_I$. The clothes warping's ground truth I_c is cropped from the reference image I according to its human parsing.

Since clothes deformations are highly non-rigid, we adopt the cascade clothes warping (CCW) module to warp clothes in a coarse-to-fine manner. Specifically, we first use thin-plate spline (TPS) [1] transformation to roughly align the clothes c with the target shape $\hat{P}_{I,c}$. TPS transformation preserves the clothes details very well as it provides a smooth interpolation between a set of control points. However, it has problems (Fig. 7) fitting the target shape due to its limited degrees of freedom or handling occlusion since it interpolates a continuous surface. Thus, we further use a dense pixel flow (Appearance Flow [42]) to refine the warping result. Powered by DensePose and clothes parsing, the flow-based warping can infer the body orientation and automatically pick the non-occluded pixels. We do not use the flow-based warping alone, as what has been done in [13], because it produces blurred results when there is a large size difference between the input clothes and the clothing on the person.

Rough Alignment by TPS Transformation: Formally, given the clothing representation R_c, DensePose d_I, and the target shape $\hat{P}_{I,c}$, we train a spatial transformation network (STN) [19] N_{TPS} to estimate the coordinates of 25 control points $\hat{\theta} \in \mathbb{R}^{25\times 2}$. We then apply the TPS transformation on the clothing representation, getting the corresponding warped result by

$$\hat{R}_c{}' = \{\hat{c}', \hat{P}_c{}'\} = W_{\hat{\theta}}(R_c), \tag{2}$$

where $\mathcal{W}_{\hat{\theta}}(\cdot)$ is a bilinear sampling decided by $\hat{\theta}$, $\hat{c}'$ is the warped clothes, $\hat{P}_c{}'$ is the warped clothes parsing.

We adopt the second-order difference constraint $\mathcal{L}_{reg}$ from ACGPN [40], which makes the warping more stable by pushing the control points to be regularly arranged. We also calculate the $\mathcal{L}_1$ Loss and perceptual loss [21] between $\hat{c}'$ and the ground truth I_c, presented as $\mathcal{L}'_1(\hat{c}', I_c)$ and $\mathcal{L}'_{perc}(\hat{c}', I_c)$, respectively.

Thus, the overall loss function for the STN is:

$$\mathcal{L}_{STN} = \lambda_{reg}\mathcal{L}_{reg} + \lambda'_1\mathcal{L}'_1 + \lambda'_{perc}\mathcal{L}'_{perc}, \tag{3}$$

where λs are balancing weights.

Shape Refinement by Appearance Flow: An appearance flow [42] $f \in \mathbb{R}^{H\times W\times 2}$ specifies, for each pixel $p^{(h,w)}$ in the output image, the coordinate at the input image where the pixel value is sampled to reconstruct $p^{(h,w)}$. Instead of estimating the flow directly [13], we let the network output an offset Δf which denotes the pixel movements and compute the final flow by $\hat{f} = f_{id} + \Delta f$. Here, f_{id} denotes an identify flow that copies every input pixel to the same place in the output image. Since adjacent pixels tend to move together, i.e., they have the same offset, letting the network learn to output the offset is easier than having it output equally-spaced pixel coordinates.

As shown in Fig. 3, we feed the warped clothing representation $\hat{R}_c{}'$ and human body information $(d_I, \hat{P}_I)$ into two separate encoders to get two feature pyramids $\{F_c^{(l)}\}, \{F_I^{(l)}\}$ $(l = 1, 2, ..., 5)$. As for flow estimation, we first feed the lowest-level features $F_c^{(5)}$ and $F_I^{(5)}$ into a residual block [16] to predict the initial offset flow Δf^5. Furthermore, at each level l $(l = 1, 2, 3, 4)$, the estimated flow from the level above is upscaled by 2 and used for

estimating the current-level flow together with the features from pyramid 1 and features from pyramid 2:

$$\Delta f^{(l)} = \text{ResBlock}^{(l)}(\mathcal{U}(\Delta f^{(l+1)}), F_c^{(l)}, F_I^{(l)})), \quad (4)$$

where $\mathcal{U}(\cdot)$ stands for the two-times bilinear upsampling.

The final flow is computed by $\hat{f} = f_{id} + \mathcal{U}(\Delta f^{(1)})$, and is used for producing the final warping result, formulated as $\hat{c} = \mathcal{W}_{\hat{f}}(\hat{c}')$.

Similar to TPS transformation, we add the $\mathcal{L}_1$ Loss and perceptual loss on the final warping result, denoted as $\mathcal{L}_1(\hat{c}, I_c)$ and $\mathcal{L}_{perc}(\hat{c}, I_c)$, respectively. Besides, we adopt the flow regularization term and intermediate flow supervisions from [29] to make the training process stable and encourage smooth flow fields.

3.4. Decomposed Neural Rendering

After getting target human parsing $\hat{P}_I$ and the warped clothing representation $\hat{R}_c$, we finally use the DNR module to generate the final try-on result by fusing the warped clothes with the reference person, where we further decompose the task into skin inpainting and clothes shading refinement.

Following ACGPN [40], we first compute the nonaffected body parts I_u by comparing the reference human parsing P_I with the predicted target parsing $\hat{P}_I$. Intuitively, I_u denotes the body parts on I that will not be affected by the clothes' change, thus should be preserved through the rendering. Moreover, we compute the average color for each skin region (i.e. left arm, right arm, and torso) and fills the corresponding area $\hat{P}_I$ with it. The computed region-specific skin color map I_s provides crude guidance for the skin inpainting process.

Finally, our rendering network N_{render} takes the target human parsing $\hat{P}_I$, the warped clothing representation $\hat{R}_c$, and the additional guidance computed above and get three outputs, formulated as:

$$\hat{I}_{nc}, \Delta\hat{c}, \alpha = N_{\text{render}}(\hat{P}_I, \hat{R}_c, I_u, I_s), \quad (5)$$

where $\hat{I}^{nc}$ refers to generated non-clothing image, , $\Delta\hat{c}$ refers to shading to be added onto warped clothes, and α is an alpha mask for fusion. To free up network capacity, we use a shared encoder and two separate decoder, each predicting $\hat{I}^{nc}$ and $\Delta\hat{c}$ with the alpha mask.

The final try-on result $\hat{I}$ is then calculated by:

$$\hat{I} = \hat{I}_{nc} + \Delta\hat{c} \odot \alpha + \hat{c} \odot (1 - \alpha) \quad (6)$$

The training objective consists of the $\mathcal{L}_1$ Loss, perceptual loss, and adversarial loss. We use three PatchGAN [18] discriminators to improve the generation photo-realism, including D that works on the whole image, D_{Skin} that focuses on the skin region, and D_{Clothes} that focuses on the clothes region.

		VITON [14]	RAIV
	Dataset Size	16,253	16,052
	Resolution	256×192	≥ 512×384
Person	Person Image	✓	✓
	Human Silhouette[2]	✓	✓
	Human Parsing[1]	✓	✓+
	OpenPose [3]	✓	✓
	DensePose [30]	×	✓
Clothes	Clothing Image	✓	✓
	Clothing Silhouette[2]	✓	✓
	Clothing Parsing	×	✓

Table 1. A comparison of VITON and our RAIVdataset. Note that: 1) The human parsing annotation [9] in RAIVinvolves full-body parts, while the one [10] used in VITON lacks the torso region. 2) Human silhouette and clothing silhouette can be extracted from human parsing and clothing parsing, respectively.

4. RAIV Dataset

We collect a large-scale image-based virtual try-on dataset named RAIV, containing 16,052 pairs of high-resolution upper clothing and person images together with rich annotations. Table 1 presents an overview comparison between RAIVand the widely used VITON [14] dataset. Specifically, RAIVhas two appealing properties:

- **Rich annotations.** RAIVprovides human parsing, OpenPose [3], and DensePose [30] for person images, and clothing parsing for clothing images. We use a state-of-the-art human parser [8] to compute the human parsing, which follows the CIHP [9] annotation to have 19 semantic labels. OpenPose stands for 25 human body keypoints. DensePose provides some 3D information for human body and is clothes-agnostic. As for clothing parsing, we first manually annotate 6,000 clothes with five labels ('background', 'outer surface', 'inner surface', 'left sleeve', and 'right sleeve', shown in Fig. 4), and then train a parsing model [8] to get the rest.
- **High quality.** As shown in Fig. 5, RAIVprovides high-quality full body human parsing comapred to VITON. Moreover, we adopt a strict reviewing process to eliminate flawed images and annotations, such as incomplete Densepose and wrong clothing parsing. Besides, the image resolution in RAIVis at least 512×384, which is two times better than that in VITON.

5. Experiments

We conduct stage-by-stage ablation studies and comparative experiments with the state-of-the-art models CP-VTON [37] and ACGPN [40]. First, we evaluate the importance of introducing structural information (clothing parsing and DensePose) into the target human parsing predic-

Figure 4. Clothing parsing labels (black: background, red: outer surface, green: inner surface, blue: left sleeve, yellow: right sleeve).

Figure 5. A representative comparison of the human parsing quality between VITON [14] and RAIV.

tion stage. Next, we evaluate how structural information and cascade warping can benefit clothes warping. Finally, we link three stages as a whole model and evaluate its ability to generate photo-realistic try-on images. While all the above methods use paired data for training, their performances are evaluated under the both paired and unpaired testing settings. We also comprehensively compare and analyze our models under both settings. All the experiments are conduct on the widely used VITON [14] dataset and our proposed RAIVdataset. More details of the experiments are shown in the supplementary materials.

5.1. Evaluating Target Parsing Prediction

We compare the target human parsing prediction performance with ACGPN. Both methods are trained on our RAIV dataset for 20 epochs. Table 2 lists the IoU measurements on paired data. We take the prediction of four semantic parts ('Tops', 'Torso', 'Left Arm', and 'Right Arm') that have the most influence on the final result. It shows that our full model (F) outperforms ACGPN (A) on all these parts.

Besides, an ablation study (C-F) is conducted to prove the structural information's effectiveness. We can conclude that adding clothing parsing (C to D) improves all selected parts' prediction accuracy. The unpaired testing results in Fig. 6 (a) verifies this improvement. We also notice that adding DensePose (C to E) improves all but the 'Tops' part, which is caused by the paired testing setting, where human silhouettes used in (C) indicate the tops' boundary while DensePose in (E) is clothing agnostic. However, when testing under the real-world scenario, i.e., testing with unpaired data, human silhouettes will mislead the prediction while DensePose will not, as shown in Fig. 6 (b). Moreover, adding the new information to ACGPN (B) leads to significant improvements to its original implementation (A) and shows comparable results with ours (F), which further validates the importance of the rich guiding information.

IDX	Method	IoU ↑			
		Tops	Torso	Left Arm	Right Arm
(A)	ACGPN [40]	0.799	0.664	0.731	0.767
(B)	ACGPN [40]♯	0.919	0.767	**0.799**	0.824
(C)	Ours	0.948	0.672	0.627	0.670
(D)	Ours^{P_c}	**0.956**	0.746	0.753	0.781
(E)	Ours^{d_I}	0.929	0.725	0.792	0.825
(F)	Ours♯	0.933	**0.768**	0.797	**0.832**

Table 2. Evaluation of Stage1 on RAIV dataset. ↑ means bigger value is better; P_c denotes using clothing parsing; d_I denotes using DensePose; ♯ denotes using both information.

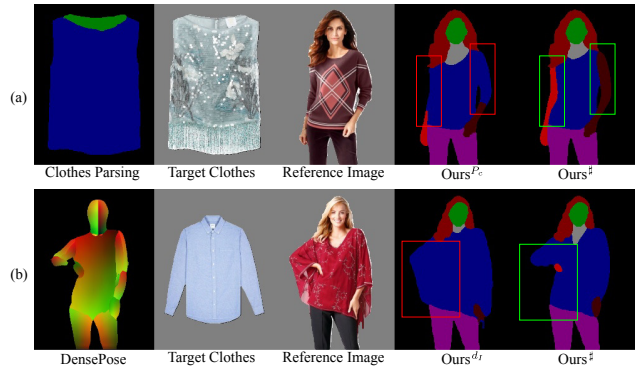

Figure 6. Abalation study on the use of clothes parsing and DensePose as inputs. (a) Clothes parsing helps improve the prediction performance when different clothes parts are hard to distinguish. (b) DensePose helps to exclude the influence of the origin shape of the reference image.

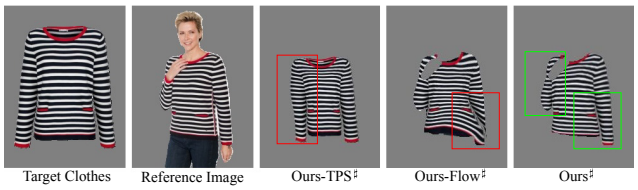

Figure 7. Abalation study on the effect of TPS and Flow. TPS transformation fails to fit the pose, and appearance flow causes distortion on texture. The combination of these two methods avoids both drawbacks.

5.2. Evaluating Cascade Clothes Warping

We evaluate different clothing warping methods with the paired testing setting. For the sake of fairness, we test each clothes warping methods independently by fetching inputs

	IDX	Method	SSIM ↑	MS-SSIM ↑
VITON	(A)	CP-VTON [37]	0.825	0.830
	(B)	ACGPN [40]	0.801	0.797
	(C)	Ours	**0.873**	**0.890**
RAIV	(D)	CP-VTON [37]	0.828	0.832
	(E)	CP-VTON [37]♯	0.831	0.837
	(F)	ACGPN [40]	0.824	0.815
	(G)	ACGPN [40]♯	0.828	0.821
	(H)	Ours	0.861	0.876
	(I)	Ours-TPS♯	0.821	0.825
	(J)	Ours-Flow♯	0.871	0.870
	(K)	Ours♯	**0.880**	**0.889**
	(L)	Ours-512♯	0.892	0.887

Table 3. Quantitative comparisons and ablation study of the clothing warping module. **Bolds** are the best results; ↑ means bigger value is better; ♯ denotes using our structural information; Ours-TPS and Ours-Flow are for the ablation study, which use only TPS or pixel flow for warping; Ours-512 denotes our model trained on 512×384 resolution. Note that all inputs are taken from the corresponding dataset, so warping results are not affected by the previous stage, if exists.

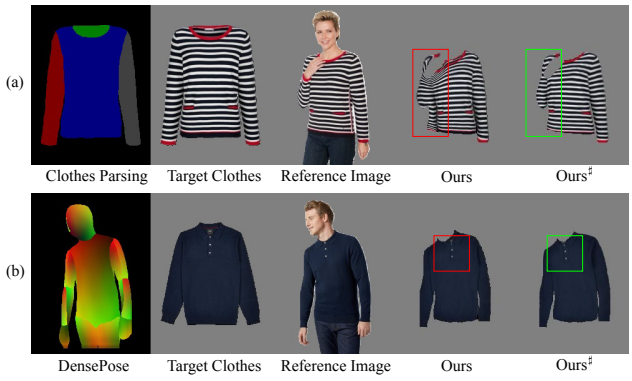

Figure 8. Abalation study of DensePose and clothes parsing. The clothes parsing prevents CCM from mixing sleeves and body together, and the DensePose helps CCM to predict the correct direction of reference person.

from the dataset.

For quantitative evaluation, we adopt SSIM and MS-SSIM [39] to measure the similarity between the warped clothing and the cropped clothing from the reference image). Table 3 lists the results of CP-VTON [37], ACGPN [40], and our SICW on both VITON [24] and RAIV datasets. It shows that our method (C, L) substantially outperforms the other methods on both datasets.

Besides, the ablation study (I-K) shows that cascade warping indeed improves the clothes warping quality. The qualitative comparison in Fig. 7 further proves this conclusion, where we see that TPS fails to fit the target shape while flow leads to distortions on the bottom of the warped garment.

Moreover, we study the effects of the new structural information. We see that all methods gain improvements (D to E, F to G, and H to K) by using the information, while our method is the most beneficiary one. The representative comparison in Fig. 8 further validates our claim.

	IDX	Method	SSIM ↑	MS-SSIM ↑	FID ↓	
					Paired	Unaired
VITON	(A)	CP-VTON [37]	0.819	0.899	18.75	39.76
	(B)	ACGPN [40]	0.867	0.916	11.63	16.29
	(C)	Ours	**0.892**	**0.949**	**10.25**	**15.80**
RAIV	(D)	CP-VTON [37]	0.613	0.802	35.77	40.23
	(E)	CP-VTON [37]♯	0.661	0.805	35.42	39.26
	(F)	ACGPN [40]	0.843	0.893	18.70	22.41
	(G)	ACGPN [40]♯	0.846	0.894	18.03	21.55
	(H)	Ours	**0.865**	0.902	15.61	21.11
	(I)	Ours♯	0.864	**0.915**	**14.89**	**19.63**
	(J)	Ours-512♯	0.879	0.906	13.54	18.27

Table 4. Quantitative comparisons on final try-on results. **Bolds** are the best results; ↑ means bigger value is better; ↓ means smaller value is better; ♯ denotes using our structural information; Ours-512 denotes our model trained on 512×384 resolution.

We also train our model on the 512×384 resolution (L) and get better results to the one trained on the 256×192 resolution, which shows our model's ability to generalize to higher resolutions.

5.3. Evaluating Final Try-on Results

We compare the final try-on results with VITON and ACGPN models on both VITON and RAIV datasets. For testing on paired data, we adopt SSIM and MS-SSIM as evaluation metrics. We also use FID [17] to measure the visual quality of synthesized images on both paired-setting (denoted as 'Paired') and unpaired-setting (denoted as 'Unpaired'). Table 4 lists the quantitative comparisons. Our model clearly outperforms the other methods and becomes the state-of-the-art one in both datasets. We also see that each model gains some improvements by adopting the new structural information (D to E, F to G, and H to J). These improvements are not as significant as what we get in previous stages (our model even drops 0.001 on SSIM). Like what we mentioned in Sec. 5.1, this phenomenon is due to the limitation of the paired testing setting. When testing on unpaired data, as shown in Fig. 10, using the new structural information improves the try-on results.

We compare the final try-on results with VITON and ACGPN on both VITON and RAIV datasets. For testing on paired data, we adopt SSIM and MS-SSIM as evaluation metrics. We also use FID [17] to measure the visual quality of synthesized images on both paired-setting (denoted as 'Paired') and unpaired-setting (denoted as 'Unpaired'). Table 4 lists the quantitative comparisons. Our model clearly outperforms the other methods and sets a new state of the art for both datasets. We also see that each model gains some improvements by adopting the new structural information (D to E, F to G, and H to J). These improvements

Figure 9. Comparison between three virtual try-on methods on clothes with simple texture (left side) and complex texture (right side). Our method outperforms the others on three aspects: (1) clothes pose matching (*e.g.* top right) (2) texture preserving (*e.g.* bottom right) (3) body parts completing (*e.g.* middle left). All these advantages make the final result more realistic and natural.

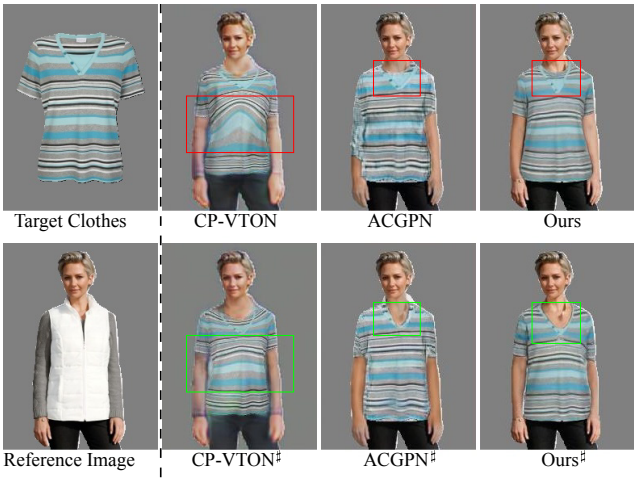

Figure 10. Comparison of three methods with and without clothes parsing and DensePose as inputs. Results show that all methods benefit from the new inputs that the performances on neckline and pose fitting are improved significantly.

are not as significant as what we get in previous stages (our model even drops 0.001 on SSIM). Like what we mentioned in Sec. 5.1, this phenomenon is due to the limitation of the paired testing setting. When testing on unpaired data, as shown in Fig. 10, using the new structural information improves the try-on results.

Moreover, We show qualitative comparisons in Fig. 9, including clothes with simple textures (on the left side) and complex textures (on the right side). Since CP-VTON does not predict the target human parsing, it often fail to fit the garment to the human body correctly, as shown in the top right results of Fig. 9. Besides, the other two methods use TPS transformation only to warp the clothes, where the preserves parts could be occluded in the final results, while our method uses appearance flow to abandon occluded parts. This technique makes our method perform better on neckline (*e.g.* bottom left result). Additionally, in CP-VTON, the single TPS transformation causes distortion. ACGPN tries to handle this problem by introducing image generation, which makes the result more blurry. In contrast, our method performs better on texture preserving by cascade cloth warping and minor necessary shading, which not only avoids distortion (*e.g.* top right result) but also provides clearer texture (*e.g.* bottom right result).

6. Conclusion

We presented RAIV, a new image-based virtual try-on dataset with 16,052 image pairs and rich annotations. It surpasses the widely used VITON dataset in terms of image resolution, annotation quality, and annotation kind. Taking advantage of RAIV, we proposed a novel model that consists of three stages. Extensive experiments show that our SICWmodel outperforms state of the art on both VITON and RAIVdatasets.

References

[1] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. volume 11, pages 567–585. IEEE, 1989. 2, 4

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

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2

[6] Jun Ehara and Hideo Saito. Texture overlay for virtual clothing based on pca of silhouettes. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 139–142. IEEE, 2006. 2

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[8] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 5

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 5

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 5

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[12] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 1, 2

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

[14] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1, 2, 3, 5, 6

[15] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. Virtual try-on through image-based rendering. *IEEE transactions on visualization and computer graphics*, 19(9):1552–1565, 2013. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[17] 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. 7

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

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015. 4

[20] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. Sievenet: A unified framework for robust image-based virtual try-on. In *WACV*, 2020. 2

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

[22] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017. 2

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[24] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *CVPR*, 2020. 7

[25] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020. 2

[26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[27] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 2

[28] Amir Hossein Raffiee and Michael Sollami. Garmentgan: Photo-realistic adversarial fashion transfer. *arXiv preprint arXiv:2003.01894*, 2020. 2

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, 2020. 5

[30] Iasonas Kokkinos Riza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1, 3, 5

[31] Damien Rohmer, Tiberiu Popa, Marie-Paule Cani, Stefanie Hahmann, and Alla Sheffer. Animation wrinkling: augmenting coarse cloth simulations with realistic-looking wrinkles. *ACM Transactions on Graphics (TOG)*, 29(6):1–8, 2010. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[33] Debapriya Roy, Sanchayan Santra, and Bhabatosh Chanda. Lgvton: A landmark guided approach to virtual try-on. *arXiv preprint arXiv:2004.00562*, 2020. 2

[34] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, pages 355–366. Wiley Online Library, 2019. 1, 2

[35] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 3

[36] Hiroshi Tanaka and Hideo Saito. Texture overlay onto flexible object with pca of silhouettes and k-means method for search into database. In *MVA*, pages 5–8, 2009. 2

[37] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 5, 7

[38] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[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*, 13(4):600–612, 2004. 7

[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 *CVPR*, 2020. 1, 2, 4, 5, 6, 7

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

[42] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2, 4

[43] 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*, pages 1–4. 2012. 2

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2