Towards Multi-pose Guided Virtual Try-on Network

Haoye Dong[1,2], Xiaodan Liang[3],Xiaohui Shen[7], Bochao Wang[1],
Hanjiang Lai[1,2], Jia Zhu[4,5], Zhiting Hu[6], Jian Yin[1,2,*]
[1]School of Data and Computer Science, Sun Yat-sen University
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
[3]School of Intelligent Systems Engineering, Sun Yat-sen University
[4]School of Computer Science, South China Normal University
[5]Guangzhou Key Laboratory of Big Data and Intelligent Education
[6]Carnegie Mellon University, [7]ByteDance AI Lab.
{donghy7@mail2,laihanj3@mail,issjyin@mail}.sysu.edu.cn
xdliang328@gmail.com,jzhu@m.scun.edu.cn,shenxiaohui@bytedance.com

Abstract

Virtual try-on systems under arbitrary human poses have significant application potential, yet also raise extensive challenges, such as self-occlusions, heavy misalignment among different poses, and complex clothes textures. Existing virtual try-on methods can only transfer clothes given a fixed human pose, and still show unsatisfactory performances, often failing to preserve person identity or texture details, and with limited pose diversity. This paper makes the first attempt towards a multi-pose guided virtual try-on system, which enables clothes to transfer onto a person with diverse poses. Given an input person image, a desired clothes image, and a desired pose, the proposed Multi-pose Guided Virtual Try-On Network (MG-VTON) generates a new person image after fitting the desired clothes into the person and manipulating the pose. MG-VTON is constructed with three stages: 1) a conditional human parsing network is proposed that matches both the desired pose and the desired clothes shape; 2) a deep Warping Generative Adversarial Network (Warp-GAN) that warps the desired clothes appearance into the synthesized human parsing map and alleviates the misalignment problem between the input human pose and the desired one; 3) a refinement render network recovers the texture details of clothes and removes artifacts, based on multi-pose composition masks. Extensive experiments on commonly-used datasets and our newly-collected largest virtual try-on benchmark demonstrate that our MG-VTON significantly outperforms all state-of-the-art methods both qualitatively and quantitatively, showing promising virtual try-on performances.

*Corresponding author is Jian Yin

Figure 1. Some results of our model. The clothes and poses images are shown in the first row, while the person images shown in the first column. The results manipulated by both clothes and pose are shown in the other columns.

1. Introduction

Virtual try-on, which enables users to try on clothes to check the size or style in a virtual way, has a huge amount of commercial value and attracts extensive attention in computer vision. Many virtual try-on systems [13, 37] have been presented and achieve promising results when the pose is fixed. However, these approaches usually learn to synthesize the image conditioned on clothes only. When given a different pose, they tend to synthesize blurry images, losing most of the details and style, as shown in Figure 4.

Meanwhile, other existing works [22, 29, 44] leverage 3D models and measurements to preserve the body shape and generate visually realistic results. However, it needs expert knowledge and huge labor cost to collect the 3D annotated data and build the 3D models. When the 3D model of the person could not be obtained or is not accurate, these methods would become inapplicable as well. To address these limitations, we propose a practical try-on task that allows users to control both the clothes and poses without any 3D annotations. Given a person image, a desired clothes, and a desired pose, we generate the person image that wears the new clothes with preserved textural appearance, and reconstruct the pose simultaneously, as illustrated in Figure 1.

The challenge of advancing from fixed-pose virtual try-on to the multi-pose try-on task comes from the fact that the warping of target clothes and the manipulation of human pose have to be learned simultaneously. Without explicitly decomposing the two and modeling the intricate interplay among the appearance, clothes and pose, an image-based end-to-end solution as in those previous methods [13, 37, 46] would not be able to disentangle the pose and appearance space, usually resulting blurry artifacts.

Targeting at the problems mentioned above, we propose a novel Multi-pose Guided Virtual Try-On Network (MG-VTON) that can generate a new person image after fitting both desired clothes into the input image and manipulating the pose. Our MG-VTON is a multi-stage framework with generative adversarial learning. Concretely, we design a pose-clothes-guided human parsing network to estimate a plausible human parsing of the target image conditioned on the information from the source image (including the approximate body shape, the face mask and the hair mask), as well as the desired clothes and the target pose. The precise region of the body parts in the source image could guide the synthesis of human parsing in an effective way. Based on the synthesized human parsing map, a geometric matching model is then used to warp the target clothes and seamlessly fit it onto the person. In addition, we design a deep Warping Generative Adversarial Network (Warp-GAN) to synthesize the coarse result, alleviating the large misalignment caused by the different poses and the diversity of clothes appearance. Finally, we present a refinement network, utilizing multi-pose composition masks to recover the texture details and alleviate the artifacts caused by the large misalignment between the reference pose and the target pose.

To demonstrate our model, we collected a new dataset, named MPV, by collecting various clothes images and person images with different poses from the same person. Furthermore, we also conduct experiments on the DeepFashion [47] datasets for evaluation. Following the object evaluation protocol [38], we conduct a human subjective study on the Amazon Mechanical Turk (AMT) platform. Both quantitative and qualitative results indicate that our method achieves effective performance and high-quality images with appealing details. The main contributions of our work are summarized as follows:

- We introduce a novel task of virtual try-on conditioned on multiple poses, and collect a new dataset that covers different poses and various clothes.
- We propose a novel Multi-pose Guided Virtual Try-On Network (MG-VTON) that handles large pose variations by disentangling the warping of clothes appearance and the pose manipulation in multiple stages. Specifically, we propose a pose-clothes guided human parsing network to first synthesize the human parsing with the desired clothes and pose, which effectively guides the virtual try-on to achieve reasonable results via the correct region parts.
- We design a Warp-GAN that integrates human parsing with geometric matching to alleviate blurry issues caused by the misalignment among different poses.
- A pose-guided refinement network is further proposed to adaptively controls the composition mask according to different poses, which learns to recover details and remove artifacts.

2. Related Work

Generative Adversarial Networks (GANs). GANs [10] consists of two networks where the discriminator learns to classify between the synthesized images and the real images while the generator tries to fool the discriminator. Existing works have studied its connections with other generative models [15, 28], and applied the approach in various domains, such as style transfer [17, 45, 20], image inpainting [41, 12], video synthesis [6], and text generation [14, 43, 42]. Inspired by those impressive results of GANs, we also apply the adversarial loss to exploit a virtual try-on method with GANs.

Person image synthesis. The skeleton-guided approach [40] generates person image conditioning on target skeletons. PG2 [25] applied a coarse-to-fine framework that consists of a coarse stage and a refined stage. The work [26] further improved the results with a new decomposition strategy. The deformableGANs [35] and [1, 11, 5] attempted to alleviate the misalignment problem between different poses using transformation on the coarse rectangle region and warped the parts, respectively. [16, 7] added structured human body constraints in learning the generation model. V-UNET [8] introduced a variational U-Net [32] to synthesize person image by restructuring the shape with stickman labels. The work [30] applied CycleGAN [45] directly to manipulate pose. However, all those works fail to preserve

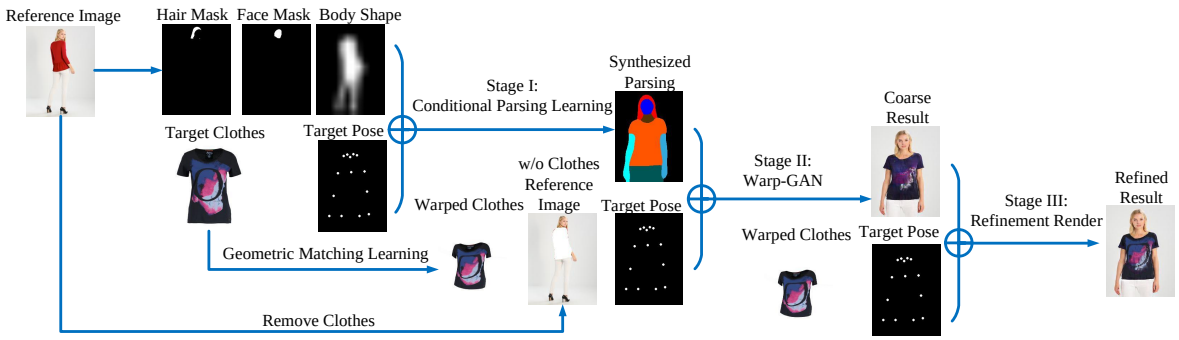

Figure 2. The overview of the proposed MG-VTON. Stage I: We first decompose the reference image into three binary masks. Then, we concatenate them with the target clothes and target pose as an input of the conditional parsing network to predict human parsing map. Stage II: Next, we warp clothes, remove the clothing from the reference image, and concatenate them with the target pose and synthesized parsing to synthesize the coarse result by using Warp-GAN. Stage III: We finally refine the coarse result with a refinement render, conditioning on the warped clothes, target pose, and the coarse result.

the textures consistency. The reason behind that is they ignore to consider the interplay among the human parsing, the clothing, and the pose. The human parsing can guide the generator to synthesize image in the precise region level that ensures the coherence of body structure.

Virtual try-on. VITON [13] and CP-VTON [37] all presented an image-based virtual try-on network, which can transfer a desired clothes on the person by using a warping strategy. VITON computed the transformation mapping by the shape context TPS [2] directly. CP-VTON introduced a learning method to estimate the transformation parameters. FashionGAN [46] learned to generate new clothes on the input image of the person conditioned on a sentence describing the different outfit. However, all of the above methods synthesized the image of person only on the fixed pose, which limits the applications in the realistic virtual try-on simulation. ClothNet [23] presented an image-based generative model to produce new clothes conditioned on color. CAGAN [18] proposed a conditional analogy network to synthesize person image conditioned on the paired of clothes, which limits the practical virtual try-on scenarios. ClothCap [29] utilized the 3D scanner to capture the clothes, the shape of the body automatically. [34] presented a virtual fitting system that requires the 3D body shape, which is laborious for collecting the annotation. In this paper, we introduce an effective method for learning to synthesize image with the new outfit on the person in different poses through adversarial learning.

3. MG-VTON

We propose a novel Multi-pose Guided Virtual Try-On Network (MG-VTON) that learns to synthesize the new person image for virtual try-on by manipulating both clothes and pose. Given an input person image, a desired clothes, and a desired pose, the proposed MG-VTON aims to produce a new image of the person by manipulating the desired clothes and poses. Inspired by the coarse-to-fine idea [13, 25], we adopt an outline-coarse-fine strategy that divides this task into three subtasks, including the conditional parsing learning, the Warp-GAN, and the refinement render. The Figure 2 illustrates the overview of MG-VTON.

We first apply the pose estimator [4] to estimate the pose. Then, we encode the pose as 18 heatmaps, which is filled with ones in a circle with radius 4 pixels and zeros elsewhere. A human parser [9] is used to predict the human parsing which is utilized to extract the binary mask of the face, the hair, and the shape of the body. Following VITON [13], we downsample the shape of the body to a lower resolution (16×12) and directly resize it to the original resolution (256×192), which helps to alleviate the artifacts caused by the variety of the body shape.

3.1. Conditional Parsing Learning

To preserve the structural coherence of the person image while manipulating both clothes and the pose, we design a pose-clothes-guided human parsing network, conditioned on the image of clothes, the pose heatmap, the approximated shape of the body, the mask of the face, and the mask of hair. As shown in Figure 4, the baseline methods failed to preserve some parts of the person (e.g., the color of the trousers and the style of hair.) because they fed the person image and clothes image into the model directly. In this work, we leverage the human parsing maps to address those problems, which can help the generator to synthesize the high-quality image on parts-level.

Formally, given an input image of person I, an input image of clothes C, and the target pose P, this stage learns to predict the human parsing map $S_t^{'}$ conditioned on clothes C and the pose P. As shown in Figure 3 (a), we first extract the hair mask M_h, the face mask M_f, the body shape M_b, and the target pose P by using a human parser [9] and a pose estimator [4], respectively. We then concatenate them with the image of clothes as the input of the conditional parsing network. The inference of $S_t^{'}$ can be formulated as

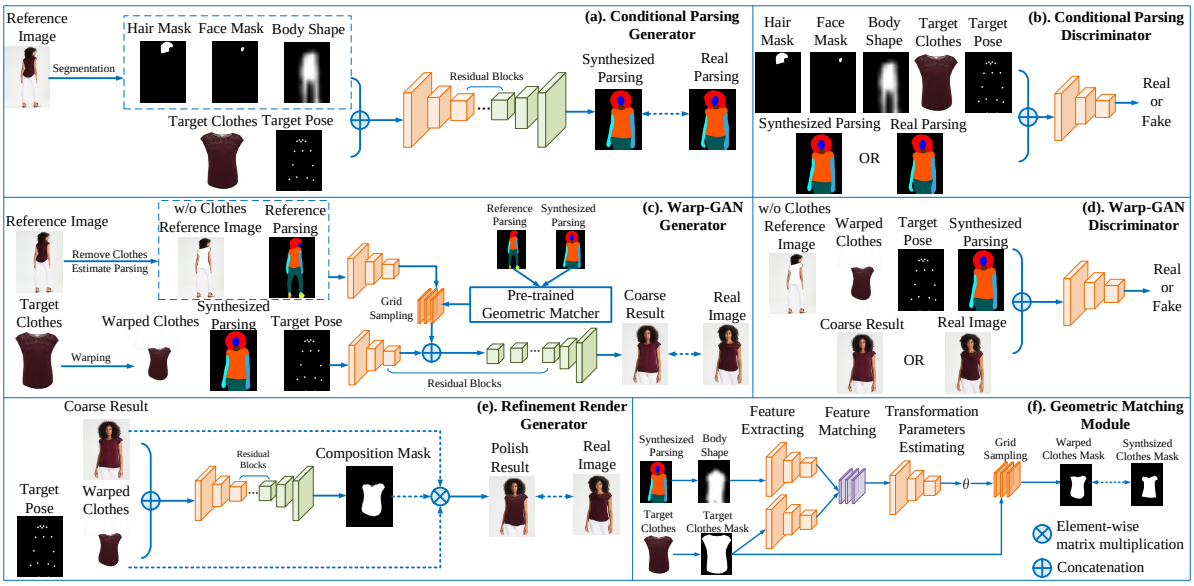

Figure 3. The network architecture of the proposed MG-VTON. (a)(b): The conditional parsing learning module consists of a pose-clothes-guided network that predicts the human parsing, which helps to generate high-quality person image. (c)(d): The Warp-GAN learns to generate a realistic image by using a warping features strategy due to the misalignment caused by the diversity of pose. (e): The refinement render network learns the pose-guided composition mask that enhances the visual quality of the synthesized image. (f): The geometric matching network learns to estimate the transformation mapping conditioned on the body shape and clothes mask.

maximizing the posterior probability:

$$p(S_t^{'}|(M_h, M_f, M_b, C, P)) = G(M_h, M_f, M_b, C, P). \tag{1}$$

We adopt a ResNet-like network as the generator G to build the conditional parsing model. We adopt the discriminator D directly from the pix2pixHD [38]. We apply the L1 loss for further improving the performance, which is advantageous for generating more smooth results [40]. Inspired by the LIP [9], we apply the pixel-wise softmax loss to encourage the generator to synthesize high-quality human parsing maps. Therefore, we formulated the problem of conditional parsing learning as:

$$\begin{aligned} &\min_G \max_D F(G, D) \\ &= \mathbb{E}_{M,C,P\sim p_{\text{data}}}[\log(1 - D(G(M, C, P), M, C, P))] \\ &+ \mathbb{E}_{S_t,M,C,P\sim p_{\text{data}}}[\log D(S_t, M, C, P)] \\ &+ \mathbb{E}_{S_t,M,C,P\sim p_{\text{data}}}[\|S_t - G(M, C, P)\|_1] \\ &+ \mathbb{E}_{S_t,M,C,P\sim p_{\text{data}}}[\mathcal{L}_{\text{parsing}}(S_t, G(M, C, P))], \end{aligned} \tag{2}$$

where M denotes the concatenation of M_h, M_f, and M_b. The loss $\mathcal{L}_{\text{parsing}}$ denotes the pixel-wise softmax loss [9]. The S_t denotes the ground truth human parsing. The p_{data} represents the distributions of the real data.

3.2. Warp-GAN

Since the misalignment of pixels would lead to generate the blurry results [35], we introduce a deep Warping Generative Adversarial Network (Warp-GAN) warps the desired clothes appearance into the synthesized human parsing map, which alleviates the misalignment problem between the input human pose and desired human pose. Different from deformableGANs [35] and [1], we warp the feature map from the bottleneck layer by using both the affine and TPS (Thin-Plate Spline) [3] transformation rather than process the pixel directly by using affine only. Thanks to the generalization capacity of [31], we directly use the pre-trained model of [31] to estimate the transformation mapping between the reference parsing and the synthesized parsing. We then warp the w/o clothes reference image by using this transformation mapping.

As illustrated in Figure 3 (c) and (d), the proposed deep warping network consists of the Warp-GAN generator G_{warp} and the Warp-GAN discriminator D_{warp}. We use the geometric matching module to warp clothes image, as described in the section 3.4. Formally, we take warped clothes image C_{w}, w/o clothes reference image $I_{\text{w/o\_clothes}}$, the target pose P, and the synthesized human parsing $S_t^{'}$ as input of the Warp-GAN generator and synthesize the result $\hat{I} = G_{\text{warp}}(C_{\text{w}}, I_{\text{w/o\_clothes}}, P, S_t^{'})$. Inspired by [19, 13, 24], we apply a perceptual loss to measure the distances between high-level features in the pre-trained model, which encourages generator to synthesize high-quality and realistic-looking images. We formulate the perceptual loss as:

$$\mathcal{L}_{\text{perceptual}}(\hat{I}, I) = \sum_{i=0}^{n} \alpha_i \|\phi_i(\hat{I}) - \phi_i(I)\|_1, \tag{3}$$

where $\phi_i(I)$ denotes the i-th $(i = 0, 1, 2, 3, 4)$ layer fea-

ture map in pre-trained network ϕ of ground truth image I. We use the pre-trained VGG19 [36] as ϕ and weightedly sum the L1 norms of last five layer feature maps in ϕ to represent perceptual losses between images. The α_i controls the weight of loss for each layer. Besides, following pixp2pixHD [38], the feature map at different scales from different layers of discriminator enhance the performance of image synthesis, we also introduce a feature loss and formulate it as:

$$\mathcal{L}_{\text{feature}}(\hat{I}, I) = \sum_{i=0}^{n} \gamma_i \|F_i(\hat{I}) - F_i(I)\|_1, \quad (4)$$

where $F_i(I)$ represent the i-th $(i = 0, 1, 2)$ layer feature map of the trained D_{warp}. The γ_i denotes the weight of L1 loss for corresponding layer.

Furthermore, we also apply the adversarial loss $\mathcal{L}_{\text{adv}}$ [10, 27] and L1 loss $\mathcal{L}_1$ [40] to improve the performance. We design a weight sum losses as the loss of G_{warp}, which encourages the G_{warp} to synthesize realistic and natural images in different aspects. We formulate it as:

$$\mathcal{L}_{G_{\text{warp}}} = \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{perceptual}} + \lambda_3 \mathcal{L}_{\text{feature}} + \lambda_4 \mathcal{L}_1, \quad (5)$$

where λ_i $(i = 1, 2, 3, 4)$ denotes the weight of corresponding loss, respectively.

3.3. Refinement Render

In the coarse stage, the identification information and the shape of the person can be preserved, but the texture details are lost due to the complexity of the clothes image. Pasting the warped clothes onto the target person directly may lead to generate the artifacts. Learning the composition mask between the warped clothes image and the coarse results also generates the artifacts [13, 37] due to the diversity of pose. To solve the above issues, we present a refinement render utilizing multi-pose composition masks to recover the texture details and remove some artifacts.

Formally, we define C_{w} as an image of warped clothes obtained by geometric matching learning module, $\hat{I}$ as a coarse result generated by the Warp-GAN, P as the target pose heatmap, and G_{p} as the generator of the refinement render. As illustrated in Figure 3 (e), taking C_{w}, $\hat{I}$, and P as input, the G_{p} learns to predict a towards multi-pose composition mask and synthesize the rendered result. We formulate the result of the refinement render as:

$$\hat{I}_{\text{p}} = G_{\text{p}}(C_{\text{w}}, \hat{I}, P) \odot C_{\text{w}} + (1 - G_{\text{p}}(C_{\text{w}}, \hat{I}, P)) \odot \hat{I}, \quad (6)$$

where $\odot$ denotes the element-wise matrix multiplication. We also adopt the perceptual loss to enhance the performance that the objective function of G_{p} can be written as:

$$\mathcal{L}_{\text{p}} = \mu_1 \mathcal{L}_{\text{perceptual}}(\hat{I}_{\text{p}}, I) + \mu_2 \|1 - G_{\text{p}}(C_{\text{w}}, \hat{I}, P)\|_1, \quad (7)$$

where μ_1 denotes the weight of perceptual loss and μ_2 denotes the weight of the mask loss.

3.4. Geometric matching learning

Inspired by [31], we adopt the convolutional neural network to learn the transformation parameters, including feature extracting layers, feature matching layers, and the transformation parameters estimating layers. As shown in Figure 3 (f), we take the mask of the clothes image and the mask of body shape as input which is first passed through the feature extracting layers. Then, we predict the correlation map by using the matching layers. Finally, we apply a regression network to estimate the TPS (Thin-Plate Spline) [3] transformation parameters for the clothes image directly based on the correlation map.

Formally, given an input image of clothes C and its mask C_{mask}, following the stage of conditional parsing learning, we obtain the approximated body shape M_b and the synthesized clothes mask $\hat{C}_{mask}$ from the synthesized human parsing. This subtask aims to learn the transformation mapping function $\mathcal{T}$ with parameter θ for warping the input image of clothes C. Due to the unseen of synthesized clothes but have the synthesized clothes mask, we learn the mapping between the original clothes mask C_{mask} and the synthesized clothes mask $\hat{C}_{mask}$ obey body shape M_b. Thus, we formulate the objective function of the geometric matching learning as:

$$\mathcal{L}_{\text{geo\_matching}}(\theta) = \|\mathcal{T}_\theta(C_{mask}) - \hat{C}_{mask}\|_1, \quad (8)$$

Therefore, the warped clothes C_w can be formulated as $C_w = \mathcal{T}_\theta(C)$, which is helpful for addressing the problem of misalignment and learning the composition mask in the above subsection 3.2 and subsection 3.3.

4. Experiments

In this section, we first make visual comparisons with other methods and then discuss the results quantitatively. We also conduct the human perceptual study and the ablation study, and further train our model on our newly collected dataset MPV test it on the Deepfashion to verify the generation capacity.

4.1. Datasets

Since each person image in the dataset used in VITON [13] and CP-VTON [37] only has one fixed pose, we collected the new dataset from the internet, named MPV, which contains 35,687 person images and 13,524 clothes images. Each person image in MPV has different poses. The image is in the resolution of 256×192. We extract the 62,780 three-tuples of the same person in the same clothes but with different poses. We further divide them into the train set and the test set with 52,236 and 10,544 three-tuples, respectively. Note that we shuffle the test set with different clothes and diverse pose for quality evaluation. DeepFashion [47] only has the pairs of the same person in different

Figure 4. Visual comparison with different methods on MPV dataset. Note that the previous methods cannot preserve the identity of the trousers and the head. DeformableGAN + CP-VTON is the model where we first use DeformableGAN [35] to change the pose and then use CP-VTON [37] to wear clothes. Please zoom in for best view.

poses but lacks of the image of clothes. To verify the generalization capacity of the proposed model, we extract 10,000 pairs from DeepFashion and randomly select clothes image from the test set of the MPV for testing.

4.2. Evaluation Metrics

We apply three measures to evaluate the proposed model, including subjective and objective metrics: 1) We perform pairwise A/B tests deployed on the Amazon Mechanical Turk (**AMT**) platform for human perceptual study. 2) we use Structural SIMilarity (**SSIM**) [39] to measure the similarity between the synthesized image and ground truth image. In this work, we take the target image (the same person wearing the same clothes) as the ground truth image used to compare with the synthesized image for computing SSIM. 3) We use Inception Score (**IS**) [33] to measure the quality of the generated images, which is a conventional method to verify the performances for image generation.

4.3. Implementation Details

Setting. We train the conditional parsing network, Warp-GAN, refinement render, and geometric matching network for 200, 15, 5, 35 epochs, respectively, using ADAM optimizer [21], with the batch size of 40, learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$. We use two NVIDIA Titan XP GPUs and Pytorch platform on Ubuntu 14.04.

Architecture. As shown in Figure 3, each generator of MG-VTON is a ResNet-like network, which consists of three downsample layers, three upsample layers, and nine residual blocks, each block has three convolutional layers with 3x3 filter kernels followed by the bath-norm layer and Relu activation function. For the discriminator, we apply the same architecture as pix2pixHD [38], which can handle the feature map in different scale with different layers. Each discriminator contains four downsample layers which include InstanceNorm and LeakyReLU activation function.

4.4. Baselines

VITON [13] and **CP-VTON** [37] are the state-of-the-art image-based virtual try-on methods which assume the pose of the person is fixed. They all used warped clothes image to improve the visual quality, but lack of the ability to generate image under arbitrary poses. In particular, VTION directly applied shape context matching [2] to compute the transformation mapping. CP-VTON borrowed the idea from [31] to estimate the transformation mapping using a convolutional network. Furthermore, we incorporate a state-of-the-art method DeformableGAN [35] with CP-VTON form other two baseline: DeformableGAN + CP-VTON and CP-VTON + DeformableGAN. **DeformableGAN + CP-VTON** first applys a pose-guided network DeformableGAN to convert the person in the reference image to the desired pose, then applys a virtual try-on network CP-VTON to try on the desired clothes. On the contrary, **CP-VTON + DeformableGAN** first uses CP-VTON to try on, then changes the pose by DeformableGAN. To obtain fairness, we first enriched the input of the VITON, CP-VTON, and DeformableGAN. Then, we retrained the VITON, CP-VTON, and DeformableGAN on MPV dataset with the same splits (train set and test set) as our model.

4.5. Quantitative Results

We conduct experiments on two benchmarks and compare against two recent related works using two widely used metrics **SSIM** and **IS** to verify the performance of the image synthesis, summarized in Table. 2. Higher scores are better. The results show that our proposed methods significantly achieve higher scores and consistently outperform all baselines on both datasets thanks to the cooperation of our conditional parsing generator, Warp-GAN, and the refinement

Table 1. Human study on MPV and DeepFashion. Each cell lists the percentage where our MG-VTON is preferred over the other method.

	VITON	CP-VTON	DeformableGAN + CP-VTON	CP-VTON + DeformableGAN	MG-VTON (w/o Parsing)	MG-VTON (w/o Render)	MG-VTON (w/o Mask)
MPV	83.1%	85.9%	89.2%	99.6%	98.5%	82.4%	84.6%
DeepFashion	88.9%	83.3%	93.2%	99.2%	99.0%	84.6%	75.5%

Table 2. Comparisons on MPV and DeepFashion.

	MPV		DeepFashion
Model	SSIM	IS	IS
VITON [13]	0.6395	2.394 ± 0.205	2.302 ± 0.116
CP-VTON [37]	0.7054	2.519 ± 0.107	1.977 ± 0.266
DeformableGAN + CP-VTON	0.6935	$\mathbf{3.354 \pm 0.047}$	3.130 ± 0.054
CP-VTON + DeformableGAN	0.7151	2.746 ± 0.068	2.649 ± 0.047
MG-VTON (w/o Parsing)	0.7539	2.578 ± 0.116	2.556 ± 0.056
MG-VTON (w/o Render)	**0.7544**	2.694 ± 0.119	2.813 ± 0.047
MG-VTON (w/o Mask)	0.7332	3.309 ± 0.137	$\mathbf{3.368 \pm 0.055}$
MG-VTON (Ours)	0.7442	3.154 ± 0.142	3.030 ± 0.057

Figure 5. Effect of the quality of human parsing. The quality of human parsing significantly affects the quality of the synthesized image in the virtual try-on task.

Figure 6. Some results from our model trained on MPV and tested on DeepFashion, which synthesizes the realistic image and captures the desired pose and clothes well.

render. Note that the MG-VTON (w/o Render) achieves the best SSIM score, and the DeformableGAN + CP-VTON achieves the best IS score, but they obtain worse visual quality results and achieve lower scores in AMT study compare with MG-VTON (ours), as illustrated in the Table 1 and Figure 7. As shown in Figure 4, MG-VTON (ours) synthesizes more realistic-looking results than MG-VTON (w/o Render), but the latter achieve higher SSIM score, which also can be observed in [19]. Hence, we believe that the proposed MG-VTON can generate high-quality person image for multi-pose virtual try-on with convincing results.

4.6. Qualitative Results

We perform visual comparisons of the proposed method with VITON [13], CP-VTON [37], DeformableGAN + CP-VTON, CP-VTON + DeformableGAN, , MG-VTON (w/o Parsing), MG-VTON (w/o Render), and MG-VTON (w/o Mask), illustrated in Figure 4, which shows that our model generates reasonable results with convincing details. Although the baseline methods have synthesized a few details of clothes, it is far from the practice towards multi-pose virtual try-on scenario. In particular, they fail to preserve the identity and the textures of the clothing. Besides, the clothing of the lower-body also cannot be preserved while the clothing of upper-body is replaced. Furthermore, the baseline methods cannot synthesize the hairstyle and face well that result in blurry images. The reasons behind are that they overlook the high-level semantics of the reference image and the relationship between the reference image and target pose in the virtual try-on task. Different from them, we adopt clothes and pose guided network to generate the target human parsing, which is helpful to alleviate the problem that lower-body clothing and hairstyle cannot be preserved. In addition, we also design a deep warping network with an adversarial loss carefully to solve the issue that the identity cannot be preserved. Furthermore, we capture the interplay of among the poses and present a multi-pose based refined network that learns to erase the noises and artifacts.

4.7. Human Perceptual Study

We perform a human study on MPV and Deepfashion [47] to evaluate the visual quality of the generated image. Similar to pix2pixHD [38], we deployed the A/B tests on the Amazon Mechanical Turk (AMT) platform. There are 1,600 images with size 256×192. We have shown three images for reference (reference image, clothes, pose) and two synthesized images with the option for picking. The workers are given two choices with unlimited time to pick the one image looks more realistic and natural, considering

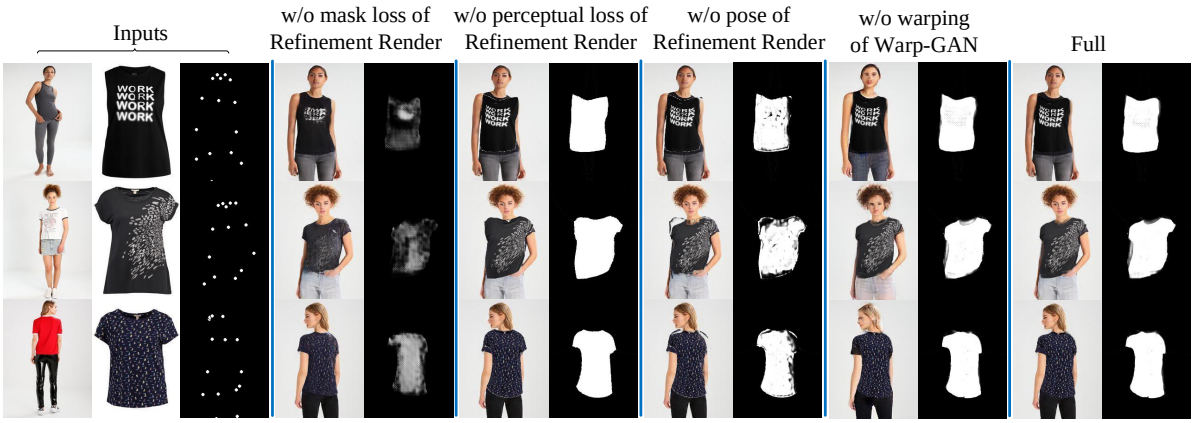

Figure 7. Ablation study on MPV dataset. Zoom in for details.

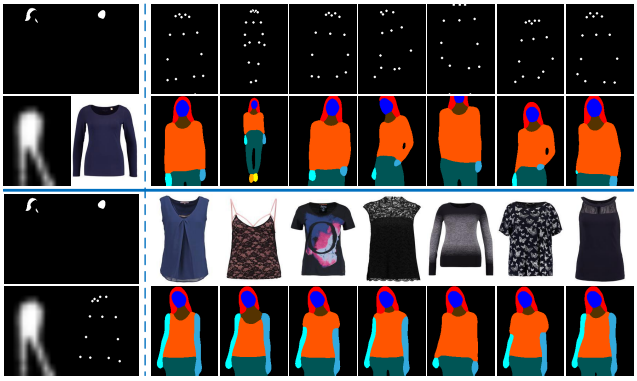

Figure 8. Effect of clothes and pose for the human parsing, which is manipulating by the pose and the clothes.

how well target clothes and pose are captured and whether the identity and the appearance of the person are preserved. Specifically, the workers are shown the reference image, target clothes, target pose, and the shuffled image pairs. We collected 8,000 comparisons from 100 unique workers. As illustrated in Table 1, the image synthesized by our model obtained higher human evaluation scores and indicate the high-quality results compare to the baseline methods.

4.8. Ablation Study

We conduct an ablation study to analyze the important parts of our method. Observed from Table. 2, MG-VTON (w/o Mask) achieves the best scores. However, as shown in Figure 4, it may inevitably generate artifacts. In Figure 7 and Figure 4, we further evaluate the effect of the components of our MG-VTON that human parsing, the multi-pose composition mask loss, the perceptual loss, and the pose in the refinement render stage, and the warping module in Warp-GAN are important to enhance the performance.

We also conduct an experiment to verify the effect of the human parsing in our MG-VTON. As shown in Figure 5, there is a positive correlation between the quality of the human parsing with that of the result. We further to verify the effect of the synthesized human parsing by manipulating the desired pose and clothes, as illustrated in Figure 8. We manipulate the human parsing instead of the person image directly, and we can synthesize the person image in an easier and more effective way. Furthermore, we introduce an experiment that trained on our collected dataset MPV and test on the DeepFashion dataset to verify the generalization of the proposed model. As the Figure 6 shown, our model captures the target pose and clothes well.

5. Conclusions

In this work, we make the first attempt to investigate the multi-pose guided virtual try-on system, which enables clothes transferred onto a person image under different poses. We propose an MG-VTON that generates a new person image after fitting the desired clothes into the input image and manipulating human poses. Our MG-VTON decomposes the virtual try-on task into three stages, incorporates a human parsing model is to guide the image synthesis, a Warp-GAN learns to synthesize the realistic image by alleviating misalignment caused by different pose, and a refinement renders recovers the texture details. We construct a new dataset for the multi-pose guided virtual try-on task covering person images with more poses and clothes diversity. Experiments demonstrate that our MG-VTON significantly outperforms existing methods both qualitatively and quantitatively with promising performances.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U1611264, 61472453, U1401256, U1501252, U1711261, U1711262, 61602530, 61836012, 61622214), the Guangzhou Key Laboratory of Big Data and Intelligent Education (201905010009), the National High Level Talents Special Support Plan (Ten Thousand Talents Program), the Natural Science Foundation of Guangdong Province under Grant No. 2017A030312006, and the Key R&D Program of Guangdong Province (2018B010107005).

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

[2] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002.

[3] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6):567–585, 1989.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-GAN for pose-guided person image synthesis. In *NeurIPS*, pages 474–484, 2018.

[6] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019.

[7] Haoye Dong, Xiaodan Liang, Chenxing Zhou, Hanjiang Lai, Jia Zhu, and Jian Yin. Part-preserving pose manipulation for person image synthesis. In *ICME*, pages 1234–1239, 2019.

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

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

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

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

[12] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry S. Davis. Compatible and diverse fashion image inpainting. In *ICCV*, 2019.

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

[14] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *ICML*, 2017.

[15] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. In *ICLR*, 2018.

[16] Zhiting Hu, Zichao Yang, Ruslan R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. In *NeurIPS*, 2018.

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

[18] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. *ICCVW*, 2(6):8, 2017.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Zorah Laehner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2018.

[23] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *CVPR*, 2017.

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[25] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[26] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

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

[28] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[29] 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):73, 2017.

[30] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, 2018.

[31] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

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

[34] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Bjorn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *International Conference on 3d Body Scanning Technologies*, pages 406–413, 2014.

[35] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. *arXiv preprint arXiv:1801.00055*, 2017.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[37] Bochao Wang, Huabin Zhang, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018.

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[40] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. Skeleton-aided articulated motion generation. In *ACM MM*, 2017.

[41] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, 2017.

[42] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*, 2018.

[43] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

[44] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, volume 2, page 3, 2017.

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

[46] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.

[47] Shi Qiu Xiaogang Wang Ziwei Liu, Ping Luo and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.