Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content

Han Yang[1,2] Ruimao Zhang[2] Xiaobao Guo[2] Wei Liu[3] Wangmeng Zuo[1] Ping Luo[4]
[1]Harbin Institute of Technology [2]SenseTime Research
[3]Tencent AI Lab [4]The University of Hong Kong
{yanghancv, wmzuo}@hit.edu.cn, wl2223@columbia.edu, pluo@cs.hku.hk
{zhangruimao, guoxiaobao}@sensetime.com

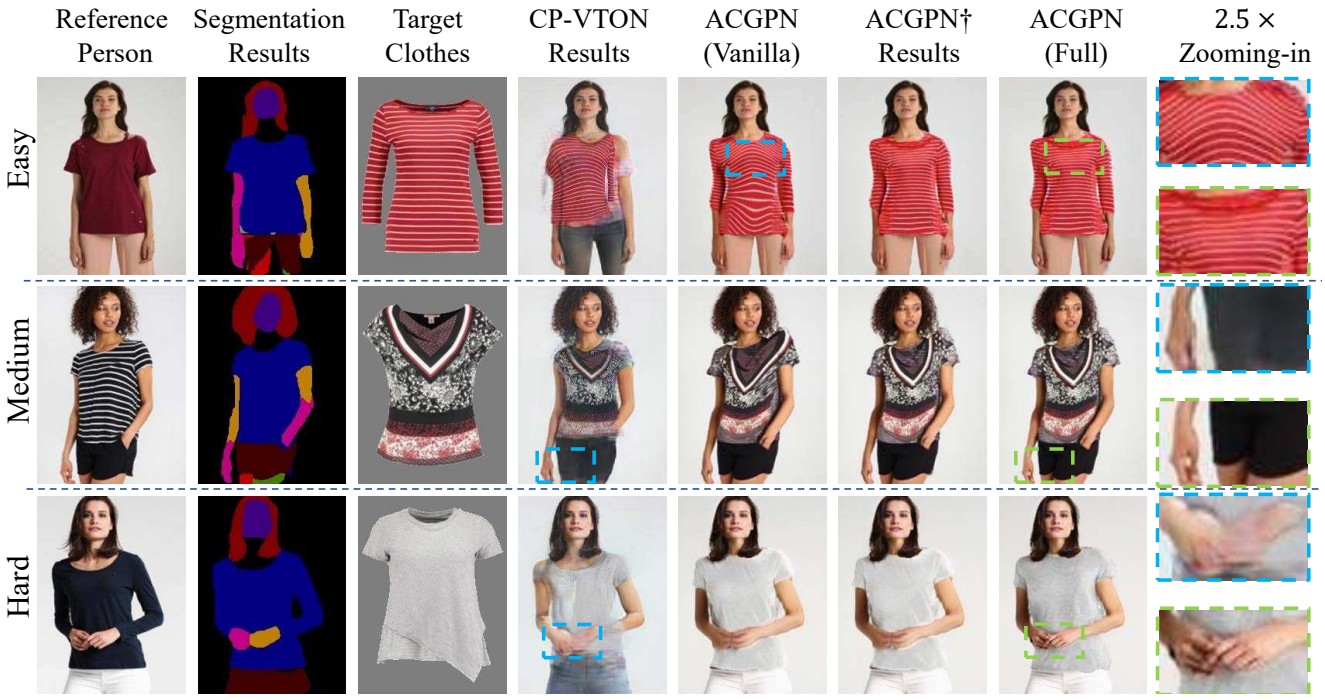

Figure 1. We define the difficulty level of the try-on task to easy, medium, and hard based on current works. Given a target clothing image and a reference image, our method synthesizes a person in target clothes while preserving photo-realistic details such as characteristics of clothes (texture, logo), posture of person (non-target body parts, bottom clothes), and identity of person. ACGPN (Vanilla) indicates ACGPN without the warping constraint or non-target body composition, ACGPN† adds the warping constraint on ACGPN (Vanilla). Also, zooming-in of the greatly improved regions are given on the right.

Abstract

*Image visual try-on aims at transferring a target clothing image onto a reference person, and has become a hot topic in recent years. Prior arts usually focus on preserving the characteristics of a clothing image (*e.g.*, texture, logo, and embroidery) when warping it to an arbitrary human pose. However, it remains a big challenge to generate photo-realistic try-on images when large occlusions and human poses are presented in the reference person (Fig. 1). To address this issue, we propose a novel visual try-on network, namely Adaptive Content Generating and Preserving Network (ACGPN). In particular, ACGPN first predicts the semantic layout of the reference image that will be changed after try-on* (e.g., *long sleeve shirt→arm, arm→jacket), and then determines whether its image content needs to be generated or preserved according to the predicted semantic layout, leading to photo-realistic try-on and rich clothing details. ACGPN generally involves three major modules. First, a semantic layout generation module utilizes semantic segmentation of the reference image to progressively predict the desired semantic layout after try-on. Second, a clothes warping module warps clothing images according to the generated semantic layout, where a second-order*

*difference constraint is introduced to stabilize the warping process during training. Third, an inpainting module for content fusion integrates all information (*e.g.*, reference image, semantic layout, and warped clothes) to adaptively produce each semantic part of human body. In comparison to the state-of-the-art methods, ACGPN can generate photo-realistic images with a much better perceptual quality and richer fine-details.*

1. Introduction

Motivated by the rapid development of image synthesis [16, 30, 21, 22] , image-based visual try-on [19, 12] aiming to transfer the target clothing item onto a reference person has achieved much attention in recent years. Although a considerable progress has been made [40, 4, 47, 2], it remains a challenging task to build up the photo-realistic virtual try-on system for the real-world scenario, partially ascribing to the semantic and geometric differences between the target clothes and reference images, as well as the interaction occlusions between the torso and limbs.

To illustrate the limitations of existing visual try-on methods, we divide the VITON dataset [12] into three subsets of difficulty levels according to the human pose in 2D reference images. As shown in Fig. 1, the first row gives an easy sample from the VITON dataset [12], where the person in the image is represented with a standard posture, *i.e.*, face forward and hands down. In such a case, the methods only need to align the semantic regions between the reference and target images. Some pioneering synthesized-based methods [19, 2, 32, 3, 37] belong to this category. From the second row, the image with the medium-level difficulty is generally with torso posture changes. And several models [12, 40, 4, 47] have been suggested to preserve the characteristics of the clothes, such as texture, logo, embroidery, and so on. Such a goal is usually attained by developing advanced warping algorithms to match the reference image with clothes deformation. The last row of Fig. 1 presents a hard example, where postural changes occur on both the torso and the limbs, leading to spatial interactions between the clothing regions and human body parts, *e.g.*, occlusions, disturbances, and deformation. Therefore, an appropriate algorithm is required to understand the spatial layout of the foreground and background objects in the reference image, and adaptively preserve such an occlusion relationship in the try-on process. However, content generation and preservation remain an uninvestigated problem in virtual try-on.

To address the above limitations, this paper presents a novel Adaptive Content Generation and Preservation Network (ACGPN), which first predicts the semantic layout of the reference image and then adaptively determines the content generation or preservation according to the predicted semantic layout. Specially, the ACGPN consists of three major modules as shown in Fig. 2. The first one is the Semantic Generation Module (SGM), which uses the semantic segmentation of body parts and clothes to progressively generate the mask of the exposed body parts (*i.e.*, the synthesized body part mask) and the mask of warped clothing regions. As opposed to prior arts, the proposed SGM generates semantic masks in a two-stage fashion to generate the body parts first and synthesize the clothing mask progressively, which makes the original clothes shape in the reference image completely agnostic to the network. The second part is the Clothes Warping Module (CWM), which is designed to warp clothes according to the generated semantic layout. Going beyond the Thin-Plate Spline based methods [12, 40, 4], a second-order difference constraint is also introduced to the Warping loss to make the warping process more stable, especially for the clothes with the complex texture. Finally, the Content Fusion Module (CFM) integrates the information from the synthesized body part mask, the warped clothing image, and the original body part image to adaptively determine the generation or preservation of the distinct human parts in the synthesized image.

With the above modules, ACGPN adopts a split-transform-merge strategy to generate a spatial configuration aware try-on image. Experiments on the VITON dataset [40] show that our ACGPN not only promotes the visual quality of generated images for the easy and medium difficulty levels (see Fig. 1), but also is effective in handling the hard try-on case with the semantic region intersections in an elegant way and producing photo-realistic results.

The main contributions of this paper can be summarized as follows. (1) We propose a new image-based virtual try-on network, *i.e.*, ACGPN, which greatly improves the try-on quality in semantic alignment, character retention, and layout adaptation. (2) We for the first time take the semantic layout into consideration to generate the photo-realistic try-on results. A novel adaptive content generation and preservation scheme is proposed. (3) A novel second-order difference constraint makes the training process of the warping module more stable, and improves the ability of our method to handle complex textures on clothes. (4) Experiments demonstrate that the proposed method can generate photo-realistic images that outperform the state-of-the-art methods both qualitatively and quantitatively.

2. Related Work

Generative Adversarial Networks. Generative Adversarial Networks (GANs) have greatly facilitated the improvements and advancements in image synthesis [16, 30, 21, 22] and manipulation [20, 23, 5]. A GAN generally consists of a generator and a discriminator. The generator learns to generate realistic images to deceive the discriminator, while the discriminator learns to distinguish the synthesized images from the real ones. Benefited from the

powerful abilities of GANs, it enjoys pervasive applications on tasks such as style transfer [50, 3], image inpainting [43, 15, 45, 46, 26], and image editing [20, 5, 23, 30]. The extensive applications of GANs further demonstrate the superiority in image synthesis.

Fashion Analysis and Synthesis. Fashion related tasks recently have received considerable attention due to their great potential in real-world applications. Most of the existing works focus on clothing compatibility and matching learning [25, 17, 39], clothing landmark detection [29, 44, 8, 24], and fashion image analysis [14, 11, 27]. Virtual try-on is among the most challenging tasks in fashion analysis.

Virtual Try-on. Virtual try-on has been an attractive topic even before the renaissance of deep learning [49, 7, 38, 13]. In the recent years, along with the progress in deep neural networks, virtual try-on has raised more and more interest due to its great potential in many real applications. Existing deep learning based methods on virtual try-on can be classified as 3D model based approaches [36, 1, 10, 31, 33] and 2D image based ones [12, 40, 4, 19], where the latter can be further categorized based on whether to keep the posture or not. Dong et al. [4] presented a multi-pose guided image based virtual try-on network. Analogous to our ACGPN, most existing try-on methods focus on the task of keeping the posture and identity. Methods such as VITON [12] and CP-VTON [40] use the coarse human shape and pose map as input to generate a clothed person. While methods such as SwapGAN [28], SwapNet [32] and VTNFP [47] adopt semantic segmentation [48] as input to synthesize a clothed person. Table 1 presents an overview of several representative methods. VITON [12] exploits a Thin-Plate Spline (TPS) [6] based warping method to first deform the inshop clothes and map the texture to the refined result with a composition mask. CP-VTON [40] adopts a similar structure to VITON but uses a neural network to learn the transformation parameters of TPS warping rather than using image descriptors, and achieves more accurate alignment results. CP-VTON and VITON only focus on the clothes, leading to coarse and blurry bottom clothes and posture details. VTNFP [47] alleviates this issue by simply concatenating the high-level features extracted from body parts and bottom clothes, thereby generating better results than CP-VTON and VITON. However, blurry body parts and artifacts still remain abundant in the results because VTNFP ignores the semantic layout of the reference image.

In Table 1, CAGAN uses analogy learning to transfer the garment onto a reference person, but can only preserve the color and coarse shape. VITON presents a coarse-to-fine structure which utilizes the coarse shape and pose map to ensure generalization to arbitrary clothes. CP-VTON adopts the same pipeline as VITON, while changing the warping module into a learnable network. These two methods perform quite well with retention of the character of clothes, but overlook the non-target body parts and bottom clothes. VTNFP ameliorates this ignorance by adding weak supervision of original body parts as well as bottom clothes to help preserve more details, which generates more realistic images than CAGAN, VITON and CP-VTON; however, VTNFP results still have a large gap between photo-realistic due to their artifacts.

		CA [19]	VI [12]	CP [40]	VT [47]	Ours
Representation	Use Coarse Shape	×	√	√	√	×
	Use Pose	×	√	√	√	√
	Use Segmentation	×	×	×	√	√
Preservation	Texture	×	√	√	√	√
	Non-target clothes	×	×	×	√	√
	Body Parts	×	×	×	×	√
Problem	Semantic Alignment	√	√	√	√	√
	Character Retention	×	√	√	√	√
	Layout Adaptation	×	×	×	×	√

Table 1. Comparison of representative virtual try-on methods. CA refers to CAGAN [19]; VI refers to VITON [12]; CP refers to CP-VTON [40], and VT refers to VTNFP [47]. We compare ACGPN with four popular image-based virtual try-on methods, *i.e.*, CAGAN, VITON, CP-VTON and VTNFP, and we compare them from three aspects: representations as input, preservation of source information, and problems to solve.

3. Adaptive Content Generating and Preserving Network

The proposed ACGPN is composed of three modules, as shown in Fig. 2. First, the Semantic Generation Module (SGM) progressively generates the mask of the body parts and the mask of the warped clothing regions via semantic segmentation, yielding semantic alignment of the spatial layout. Second, the Clothes Warping Module (CWM) is designed to warp the target clothing image according to the warped clothing mask, where we introduce a second-order difference constraint on Thin-Plate Spline (TPS) [6] to produce geometric matching yet character retentive clothing images. Finally, Steps 3 and 4 are united in the Content Fusion Module (CFM), which integrates the information from previous modules to adaptively determine the generation or preservation of the distinct human parts in the output synthesized image.The non-target body part composition is able to handle different scenarios flexibly in try-on task while mask inpainting fully exploits the layout adaptation ability of the ACGPN when dealing with the images from easy, medium, and hard levels of difficulties.

3.1. Semantic Generation Module (SGM)

The semantic generation module (SGM) is proposed to separate the target clothing region as well as to preserve the body parts (*i.e.*, arms) of the person, without changing the pose and the rest human body details. Many previous works focus on the target clothes but overlook human body generation by only feeding the coarse body shape directly into the

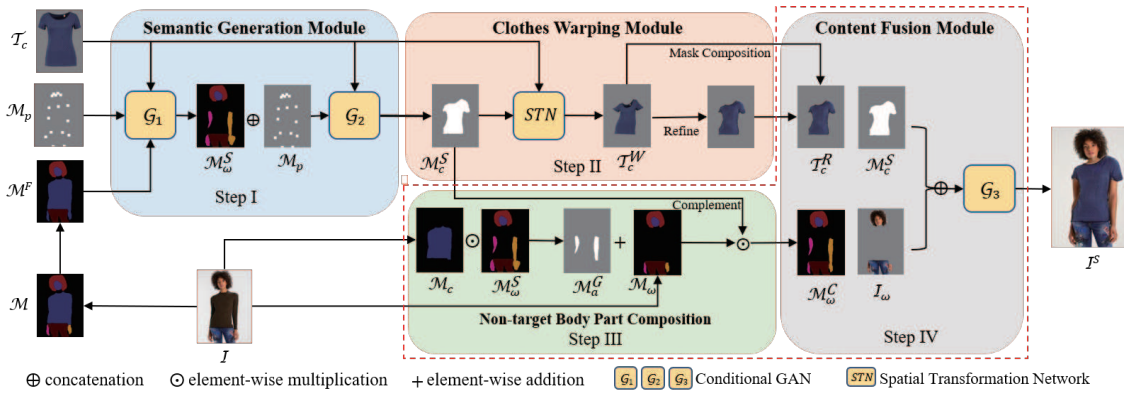

Figure 2. The overall architecture of our ACGPN. (1) In Step I, the Semantic Generation Module (SGM) takes the target clothing image $\mathcal{T}_c$, the pose map $\mathcal{M}_p$, and the fused body part mask $\mathcal{M}^F$ as the input to predict the semantic layout and to output the synthesized body part mask $\mathcal{M}^S_\omega$ and the target clothing mask $\mathcal{M}^S_c$; (2) In Step II, the Clothes Warping Module (CWM) warps the target clothing image to $\mathcal{T}^R_c$ according to the predicted semantic layout, where a second-order difference constraint is introduced to stabilize the warping process; (3) In Steps III and IV, the Content Fusion Module (CFM) first produces the composited body part mask $\mathcal{M}^C_\omega$ using the original clothing mask $\mathcal{M}_c$, the synthesized clothing mask $\mathcal{M}^S_c$, the body part mask $\mathcal{M}_\omega$, and the synthesized body part mask $\mathcal{M}^S_\omega$, and then exploits a fusion network to generate the try-on images $\mathcal{I}^S$ by utilizing the information $\mathcal{T}^R_c$, $\mathcal{M}^S_c$, and the body part image $\mathcal{I}_\omega$ from previous steps.

network, leading to the loss of the body part details. To address this issue, a mask generation mechanism is adopted in this module to generate semantic segmentation of the body parts and target clothing region precisely.

Specifically, given a reference image $\mathcal{I}$, and its corresponding mask $\mathcal{M}$, arms $\mathcal{M}_a$ and torso $\mathcal{M}_t$ are first fused into an indistinguishable area, resulting in the fused map $\mathcal{M}^F$ shown in Fig. 2 as one of the inputs to SGM. Following a two-stage strategy, the try-on mask generation module first synthesizes the masks of the body parts $\mathcal{M}^S_\omega$ ($\omega = \{h, a, b\}$ (h:head, a:arms, b:bottom clothes)), which helps adaptively preserve the body parts instead of the coarse feature in the subsequent steps. As shown in Fig. 2, we train a body parsing GAN $\mathcal{G}_1$ to generate $\mathcal{M}^S_\omega$ by leveraging the information from the fused map $\mathcal{M}^F$, the pose map $\mathcal{M}_p$, and the target clothing image $\mathcal{T}_c$. Using the generated information of the body parts, and its corresponding pose map and target clothing image, it is tractable to get the estimated clothing region. In the second stage, $\mathcal{M}^S_\omega$, $\mathcal{M}_p$ and $\mathcal{T}_c$ are combined to generate the synthesized mask of the clothes $\mathcal{M}^S_c$ by $\mathcal{G}_2$.

For training SGM, both stages adopt the conditional generative adversarial network (cGAN), in which a U-Net structure is used as the generator while a discriminator given in pix2pixHD [41] is deployed to distinguish generated masks from their ground-truth masks. For each of the stages, the CGAN loss can be formulated as

$$\begin{aligned}\mathcal{L}_1 =& \mathbb{E}_{x,y}\left[\log\left(\mathcal{D}\left(x,y\right)\right)\right] \\ &+ \mathbb{E}_{x,z}\left[\log\left(1-\mathcal{D}\left(x,\mathcal{G}\left(x,z\right)\right)\right)\right],\end{aligned} \tag{1}$$

where x indicates the input and y is the ground-truth mask. z is the noise which is an additional channel of the input sampled from a standard normal distribution.

The overall objective function for each stage of the proposed try-on mask generation module is formulated as $\mathcal{L}_m$,

$$\mathcal{L}_m = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2, \tag{2}$$

where $\mathcal{L}_2$ is the pixel-wise cross entropy loss [9], which improves the quality of synthesized masks from the generator with more accurate semantic segmentation results. λ_1 and λ_2 are the trade-off parameters for two loss terms in Eq. (2), which are set to 1 and 10, respectively, in our experiments.

The two-stage SGM can serve as a core component for accurate understanding of body-parts and clothes layouts in visual try-on and guiding the adaptive preserving of image content by composition. We believe that SGM is effective for other tasks that need to partition the semantic layout.

3.2. Clothes Warping Module (CWM)

Clothes warping aims to fit the clothes into the shape of the target clothing region with visually natural deformation according to human pose as well as to retain the characteristics of the clothes. However, simply training a Spatial Transformation Network (STN) [18] and applying Thin-Plate Spline (TPS) [6] cannot ensure the precise transformation especially when dealing with hard cases (*i.e.*, the clothes with complex texture and rich colors), leading to misalignment and blurry results. To address these problems, we introduce a second-order difference constraint on the clothes warping network to realize geometric matching and character retention. As shown in Fig. 3, compared to the result with our proposed constraint, target clothes transformation without the constraint shows obvious distortion on shape and unreasonable mess on texture.

Formally, given $\mathcal{T}_c$ and $\mathcal{M}^S_c$ as the input, we train the STN to learn the mapping between them. The warped clothing image $\mathcal{T}^W_c$ is transformed by the learned parameters from STN, where we introduce the following constraint $\mathcal{L}_3$

as a loss term,

$$\mathcal{L}_3 = \sum_{p \in \mathbf{P}} \lambda_r \left| \|pp_0\|_2 - \|pp_1\|_2 \right| + \left| \|pp_2\|_2 - \|pp_3\|_2 \right| + \lambda_s \left(\left| S\left(p, p_0\right) - S\left(p, p_1\right) \right| + \left| S\left(p, p_2\right) - S\left(p, p_3\right) \right| \right), \tag{3}$$

where λ_r and λ_s are the trade-off hyper-parameters. Practically we can minimize $max(\mathcal{L}_3 - \Delta, 0)$ for restriction, and Δ is a hyper-parameter. As illustrated in Fig. 3, $p(x, y)$ represents a certain sampled control point and $p_0(x_0, y_0), p_1(x_1, y_1), p_2(x_2, y_2), p_3(x_3, y_3)$ are the top, bottom, left, and right sampled control points of $p(x, y)$, respectively, in the whole control points set $\mathbf{P}$; $S\left(p, p_i\right) = \frac{y_i - y}{x_i - x}$ (i = 0, 1, 2, 3) is the slope between two points. $\mathcal{L}_3$ is proposed to serve as a constraint on TPS transformation by minimizing the metric distance of two neighboring intervals for each axis and the distance between slopes, which maintains the collinearity, parallelism, and immutability properties of affine transformation. To avoid the divided-by-zero error, the actual implementation of the second term is

$$\begin{aligned} &\left| S\left(p, p_i\right) - S\left(p, p_j\right) \right| \\ &= \left| (y_i - y)(x_j - x) - (y_j - y)(x_i - x) \right|, \end{aligned} \tag{4}$$

where $(i, j) \in \{(0, 1), (2, 3)\}$. The warping loss can be represented as $\mathcal{L}_w$, which measures the loss between the warped clothing image $\mathcal{T}_c^W$ and its ground-truth $\mathcal{I}_c$,

$$\mathcal{L}_w = \mathcal{L}_3 + \mathcal{L}_4, \tag{5}$$

where $\mathcal{L}_4 = \|\mathcal{T}_c^W - \mathcal{I}_c\|_1$. The warped clothes are then fed into the refinement network to further generate more details, where a learned matrix $\alpha (0 \leq \alpha_{ij} \leq 1)$ is then utilized to finally combine the two clothing images as the refined clothing image $\mathcal{T}_c^R$ by

$$\mathcal{T}_c^R = (1 - \alpha) \odot \mathcal{T}_c^W + \alpha \odot \mathcal{T}_c^R, \tag{6}$$

where $\odot$ denotes element-wise multiplication. α is also restricted by a regularization term (refer to CP-VTON [40]) and the VGG loss is also introduced on $\mathcal{T}_c^R$ and $\mathcal{T}_c^W$. For better quality, the GAN loss can be also used here. Consequently, the refined clothing image can fully retain the characteristics of the target clothes. We believe that our formulation of CWM is effective in enforcing the collinearity of local affine transforms while maintaining the flexibility of TPS warping globally, which is beneficial to producing geometrically matched and realistic warped results.

3.3. Content Fusion Module (CFM)

Going beyond semantic alignment and character retention, it remains a great challenge to realize layout adaptation on the visual try-on task. To this end, the target clothing region is required to clearly rendered, and fine-scale details of the body parts (*i.e.*, finger gaps) are needed to be adaptively preserved. Existing methods usually adopt the coarse body shape as a cue to generate the final try-on images, and fail to reconstruct fine details. In contrast, the proposed content fusion module (CFM) is composed of two main steps, *i.e.*, Steps 3 and 4 in Fig. 2. In particular, Step 3 is designed to fully maintain the untargeted body parts as well as adaptively preserve the changeable body part (*i.e.*, arms). Step 4 fills in the changeable body part by utilizing the masks and images generated from previous steps accordingly by an inpainting based fusion GAN, $\mathcal{G}_3$ in Fig. 2.

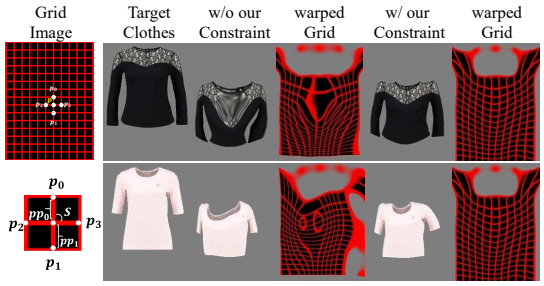

Figure 3. Comparison of STN warping results with and without the second-order difference constraint.

Non-target Body Part Composition. The composited body mask $\mathcal{M}_\omega^C$ is composed by the original body part mask $\mathcal{M}_\omega$, the generated body mask $\mathcal{M}_a^G$ which is the region for generation, and the synthesized clothing mask $\mathcal{M}_c^S$ according to

$$\mathcal{M}_a^G = \mathcal{M}_\omega^S \odot \mathcal{M}_c, \tag{7}$$

$$\mathcal{M}_\omega^C = (\mathcal{M}_a^G + \mathcal{M}_\omega) \odot (1 - \mathcal{M}_c^S), \tag{8}$$

$$\mathcal{I}_\omega = \mathcal{I}_{\omega'} \odot (1 - \mathcal{M}_c^S), \tag{9}$$

where $\odot$ denotes element-wise multiplication, and Eq. (9) is not shown in Fig. 2 for simplicity; $\mathcal{I}_{\omega'}$ is the original image I subtracting clothing region $\mathcal{M}_c$. Note that the composited body mask $\mathcal{M}_\omega^C$ always keeps a similar layout to the synthesized body part mask $\mathcal{M}_\omega^S$ by composition to eliminate the misaligned pixels in $\mathcal{M}_\omega^S$. It precisely preserves the non-target body part by combining the two masks (*i.e.*, $\mathcal{M}_\omega^S$ and $\mathcal{M}_\omega$), which are used to fully recover the non-targeted details in the following step to fully preserve $\mathcal{I}_\omega$ and generate coherent body parts with the guidance of $\mathcal{M}_a^G$. It is also worth noting that it can adaptively deal with different cases. For example, when transferring a T-shirt (short-sleeve) to a person in long-sleeve only the within region of $\mathcal{M}_a^G$ will perform generation and preserve all the others, while in the opposite case, $\mathcal{M}_a^G = \mathbf{0}$ and bulgy body parts will be shaded by clothes as in Eq. (8) and Eq. (9).

Mask Inpainting. In order to fully exploit the layout adaptation ability of the network during training, CFM uses masks $\mathcal{M}_k$ from the *Irregular Mask Dataset* [26] to randomly remove part of the arms in the body images $\mathcal{I}_\omega$ as $\mathcal{I}_\omega = (1 - \mathcal{M}_k \odot \mathcal{M}_a) \odot \mathcal{I}_{\omega'}$ for mimicking image inpainting, where $\mathcal{M}_a$ is the mask of arms and is similar to Eq. (9) in the form, making it possible to separate the regions of

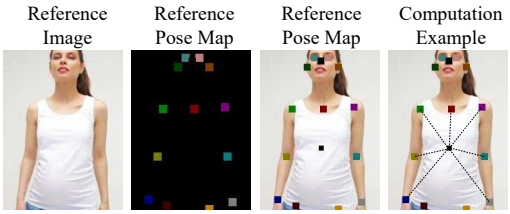

Figure 4. An example of computing the complexity score $\mathcal{C}$. Given a reference image and its pose map, the connected points shown in the last image are selected to calculate $\mathcal{C}$ for the reference image.

preservation and generation. To combine the semantic information, composited body mask $\mathcal{M}_{\omega}^{C}$ and synthesized clothing mask $\mathcal{M}_{c}^{S}$ are concatenated with the body part image $\mathcal{I}_{\omega}$ and refined clothing image $\mathcal{T}_{c}^{R}$ as the input. Thus, the texture information can be recovered by the proposed inpainting based fusion GAN, yielding the photo-realistic results. Therefore, in the inference stage, the network can adaptively generate the photo-realistic try-on image with rich details via the proposed CFM. Extensive experiments in Section 4 demonstrate that the proposed method can not only solve cases of easy and medium levels but also hard cases with significant improvement.

4. Experiments

4.1. Dataset

Experiments are conducted on the dataset (*i.e.*, VITON [12] dataset) that was used in VITON [12] and CP-VITON [40]. It contains about 19,000 image pairs, each of which includes a front-view woman image and a top clothing image. After removing the invalid image pairs, it yields 16,253 pairs, further splitting into a training set of 14,221 pairs and a testing set of 2,032 pairs. ACGPN is compared with VITON, CP-VTON and VTNFP. Without the official code of VTNFP, we compare the visual results reported in VTNFP's paper and reproduce it for quantitative comparison. Extensive ACGPN try-on results are given in appendix.

Dataset Partition. Images of the try-on task exhibit different difficulty levels as shown in Fig. 1. Easy case usually shows a standard posture with face forward and hands down; Medium level images present the twisting of the body torso or one of the hands overlapping with the body; hard cases show both torso twisting and two hands blocking in front of the body. Limbs intersections and torso occlusions raise a great challenge for the semantic layout prediction. To describe this, we propose to use reference points to represent body parts by leveraging pose maps as illustrated in Fig. 4. To quantitatively score the complexity of each image, we define the complexity of a certain image as

$$\mathcal{C} = \frac{\sum_{t \in \mathcal{M}_{p'}}^{N} \left\| t - \frac{\sum_{t \in \mathcal{M}_{p'}}^{N} t}{N} \right\|_1}{N}, \tag{10}$$

where $\mathcal{M}_{p'}$ represents points of left (right) arm, left (right) shoulder, left (right) hand, and torso. $t = (x_t, y_t)$ is a certain pose point and $N = 7$ indicates the number of reference points. We define the thresholds of easy to medium as 80, and medium to hard as 68, in the sense that when $\mathcal{C} < 68$ the layout intersections become complicated, and when $\mathcal{C} > 80$ the images tend to be standard posture, face forward and hands down. 423, 514, and 1095 images are split into hard, medium, and easy levels, respectively.

4.2. Implementation Details

Architecture. ACGPN contains SGM, CWM and CFM. All the generators in SGM and CFM have the same structure as U-Net [34] and all the discriminators are from pix2pixHD [41]. The structure of STN [18] in CWM begins with five convolutional layers followed by a max-pooling layer with stride 2. Resolution for all images in training and testing is 256×192. Followed by steps in Fig. 2, we first predict the semantic layout of the reference image, and then decide the generation and preservation of image content.

Training. We train the proposed modules separately and combine them to ultimately output the try-on image. Target clothes used in the training process are the same as in the reference image since it is intractable to grab the ground-truth images of try-on results. Each module in the proposed method is trained for 20 epochs by setting the weights of losses $\lambda_r = \lambda_s = 0.1$, $\lambda_1 = \lambda_2 = 1$, and batch-size 8. The learning rate is initialized as 0.0002 and the network is optimized by the Adam optimizer with the hyper-parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All the codes are implemented by deep learning toolkit PyTorch and eight NVIDIA 1080Ti GPUs are used in our experiments.

Testing. The testing process follows the same procedure as training but is only different in that the target clothes are different from the ones in the reference images. We test our model in easy, medium and hard cases, respectively, and evaluate the results qualitatively and quantitatively. More evaluation results are given in the following sections.

4.3. Qualitative Results

We perform visual comparison of our proposed method with VITON [12], CP-VTON [40], and VTNFP [47]. As shown in Fig. 5, from top to bottom, the difficulty levels of the try-on images are arranged from easy to hard. In all difficulty levels the images generated by VITON show many visual artifacts including color mixture, boundary blurring, cluttered texture, and so on. In comparison to VITON, CP-VITON achieves better visual results in the easy level but still results in unnecessary editing on bottom clothes and blurring on body parts in the medium and hard levels. Bad cases such as broken arms in the generated images can also be observed when there are intersections between arms and torso. To sum up, VITON and CP-VTON warp the image onto the clothing region and map the texture and embroi-

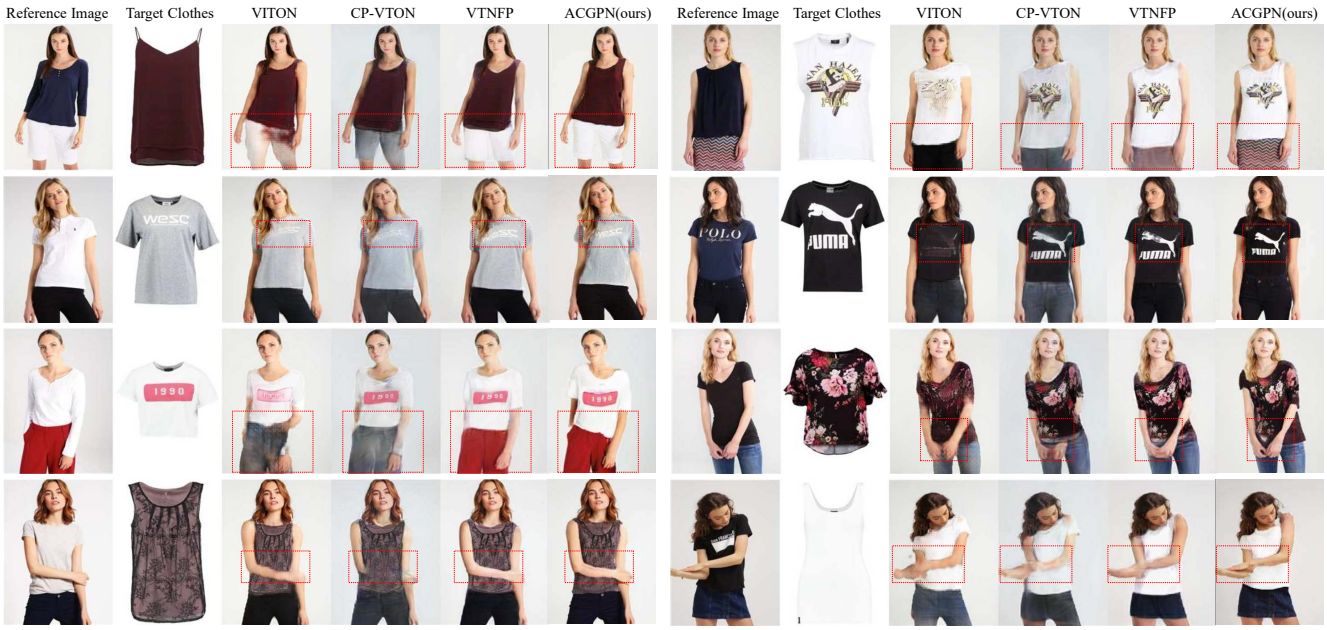

Figure 5. Visual comparison of four virtual try-on methods in easy to hard levels (from top to bottom). ACGPN generates photo-realistic try-on results, which preserves both the clothing texture and person body features. With the second-order difference constraint, the embroideries and texture are less likely to be distorted (*i.e.*, the 2nd row). With the preservation ability of the non-target body part composition, the body parts in our results are visually much more photo-realistic (*i.e.*, the 4th row). Especially different regions are marked in red-boxes.

ders, thereby possibly causing the incorrect editing on body parts and bottom clothes.

VTNFP uses segmentation representation to further preserve the non-target details of body parts and bottom clothes, but is still inadequate to fully preserve the details, resulting in blurry output. The drawbacks behind VTNFP lie in the unawareness of the semantic layout and relationship within the layout, therefore being unable to extract the specific region to preserve. In comparison to VITON and CP-VTON, VTNPF is better in preserving the characteristics of clothes and visual results, but still struggles to generate body parts details (*i.e.*, hands and finger gaps). It is worth noting that all the methods cannot avoid distortions and misalignments on the Logo or embroidery, remaining a large gap to photo-realistic try-on.

In contrast, ACGPN performs much better in simultaneously preserving the characteristics of clothes and the body part information. Benefited from the proposed second-order spatial transformation constraint in CWM, it prevents Logo distortion and realizes character retention, making the warping process to be more stable to preserve texture and embroideries. As shown in the first example of the second row in Fig. 5, Logo 'WESC' is over-stretched in results of the competing methods; however, in ACGPN, it is clear and undistorted. The proposed inpainting-based CFM specifies and preserves the unchanged body parts directly. Benefited from the prediction of semantic layout and adaptive preservation of body parts, ACGPN is able to preserve the fine-scale details which are easily lost in the competing methods, clearly demonstrating its superiority over VITON, CP-VTON and VTNFP.

4.4. Quantitative Results

We adopt Structural SIMilarity (SSIM) [42] to measure the similaity between synthesized images and ground-truths, and Inception Score (IS) [35] to measure the visual quality of synthesized images. Higher scores on both metrics indicate higher quality of the results.

Table 2 lists the SSIM and IS scores by VITON [12], CP-VTON [40], VTNFP [47], and our ACGPN. Unsurprisingly, the SSIM score decreases along with the increase of difficult level, demonstrating the negative correlation between difficulty level and try-on image quality. Nonetheless, our ACGPN outperforms the competing methods by a large margin in both metrics for all difficulty levels. For the easy case, ACGPN surpasses VITON, CP-VTON and VTNFP by 0.067, 0.101 and 0.044 in terms of SSIM, respectively. For the medium case, the gains by ACGPN are 0.062, 0.099 and 0.040, respectively. As for the hard case, ACGPN also outperforms VITON, CP-VTON and VTNFP by 0.049, 0.099, and 0.040. In terms of IS, the overall gains against VITON, CP-VTON and VTNFP are respectively 0.179, 0.072 and 0.045, further showing the superiority of ACGPN by means of quantitative metrics.

4.5. Ablation Study

Ablation study is conducted to evaluate the effectiveness of the major modules in ACGPN in Table 2. Here,

Method	SSIM				IS
	All	Easy	Medium	Hard	
VITON [12]	0.783	0.787	0.779	0.779	2.650
CP-VTON [40]	0.745	0.753	0.742	0.729	2.757
VTNFP [47]	0.803	0.810	0.801	0.788	2.784
ACGPN†	0.825	0.834	0.823	0.805	2.805
ACGPN*	0.826	0.835	0.823	0.806	2.798
ACGPN	**0.845**	**0.854**	**0.841**	**0.828**	**2.829**

Table 2. The SSIM [42] and IS [35] results of five methods. ACGPN† and ACGPN* are ACGPN variants for ablation study.

ACGPN† refers to directly using $\mathcal{M}_{\omega}^{S}$ instead of $\mathcal{M}_{\omega}^{C}$ in CFM to generate a try-on image, and ACGPN* refers to using $\mathcal{M}_{\omega}^{C}$ as the input. Both models use $\mathcal{I}_{\omega}$ with the removal of arms. Comparing to ACGPN†, ACGPN* and ACGPN, it shows that the non-target body part composition indeed contributes to yield better visual results. We also notice that ACGPN† and ACGPN* also outperform VITON [12], CP-VTON [40] and VTNFP [47] by a margin, owing to the accurate estimation of the semantic layout. Visual comparison results in Fig. 6 further show the effectiveness of body part composition in adaptive preservation. With the composition, the human body layout can be clearly stratified. Otherwise, we can only get correct body part shape but may generate wrong details as in (f) of Fig. 6.

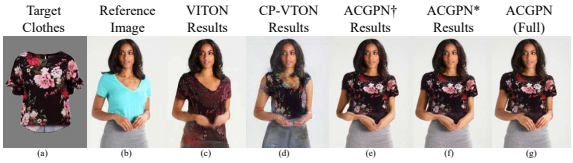

Figure 6. Visual comparison of our non-target body part composition. (c) generates incorrect target clothes and blurry body parts; (d) produces body parts with deformation; (e) and (f) show some distorted body parts; (g) generates the convincing result.

An experiment is also conducted to verify the effectiveness of our second-order difference constraint in CWM. As shown in Fig. 7, we choose target clothes with complicated embroiders for example. From Fig. 7(c), the warping model may generate distorted images without the constraint.

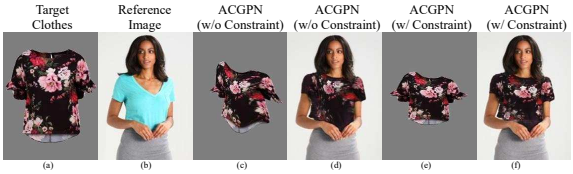

Figure 7. Ablation study on the effect of the second-order difference constraint. (c), (e) are the warped clothes, and (d), (f) are the synthesized results. Although ACGPN eliminates the artifacts in distorted warped clothing image (c), it still largely influences its verisimilitude of (d).

It is worth noting that, due to the effectiveness of semantic layout prediction, ACGPN without the constraint can still produce satisfying results, and the target clothes with pure color or simple embroideries are less vulnerable to the degeneration of warping. Regarding the target clothes with complex textures, the second-order difference constraint plays an important role in generating photo-realistic results with correct detailed textures (see in Fig. 7(d)(f)).

Method	Easy	Medium	Hard	Mean
CP-VTON [40]	15.4%	11.2%	4.0%	10.2%
ACGPN	**84.6%**	**88.8%**	**96.0%**	**89.8%**
VITON [12]	38.8%	18.2%	13.3%	23.4%
ACGPN	**61.2%**	**81.8%**	**86.7%**	**76.6%**
VTNFP [47]	45.6%	31.0%	23.4%	33.3%
ACGPN	**54.4%**	**69.0%**	**76.6%**	**66.7%**

Table 3. User study results on the VITON dataset. The percentage indicates the ratio of images which are voted to be better than the compared method.

4.6. User Study

To further assess the results of try-on images generated by VITON [12], CP-VTON [40], VTNFP [47] and ACGPN, we conduct a user study by recruiting 50 volunteers. We first test 200 images by different methods from easy, medium, and hard cases, respectively, and then group 1,800 pairs in total (each method contains 600 test images in three levels and each pair includes images from different methods). Each volunteer is assigned 100 image pairs in an A/B manner randomly. For each image pair, the target clothes and reference images are also attached in the user study. Each volunteer is asked to choose a better image meeting three criterion : (a) how well the target clothing characteristics and posture of the reference image are preserved; (b) how photo-realistic the whole image is; (c) how good the whole person seems. And we give the user unlimited time to choose the one with better quality. The results are shown in Table 3. It reveals the great superiority of ACGPN over the other methods, especially in hard cases. The results demonstrate the effectiveness of the proposed method in handling body part intersections and occlusions on visual try-on tasks.

5. Conclusion

In this work, we proposed a novel adaptive content generating and preserving network, dubbed ACGPN, which aims at generating photo-realistic try-on results while preserving both the characteristics of clothes and details of the human identity (posture, body parts, and bottom clothes). We presented three carefully designed modules, *i.e.*, Mask Generation Module (GMM), Clothes Warping Module (CWM), and Content Fusion Module (CFM). We evaluated our ACGPN on the VITON [12] dataset with three levels of try-on difficulties. The results clearly show the great superiority of ACGPN over the state-of-the-art methods in terms of quantitative metrics, visual quality, and user study.

Acknowledgement This work was partially supported by HKU Seed Fund for Basic Research and Start-up Fund, and the NSFC project under Grant No. U19A2073.

References

[1] Rémi Brouet, Alla Sheffer, Laurence Boissieux, and Marie-Paule Cani. Design preserving garment transfer. *ACM Trans. Graph.*, 31(4):36:1–36:11, 2012.

[2] Szu-Ying Chen, Kin-Wa Tsoi, and Yung-Yu Chuang. Deep virtual try-on with clothes transform. In *ICS*, volume 1013 of *Communications in Computer and Information Science*, pages 207–214. Springer, 2018.

[3] Yunjey Choi, Min-Je 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*, pages 8789–8797. IEEE Computer Society, 2018.

[4] Haoye Dong, Xiaodan Liang, Bochao Wang, Hanjiang Lai, Jia Zhu, and Jian Yin. Towards multi-pose guided virtual try-on network. *CoRR*, abs/1902.11026, 2019.

[5] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Zhenyu Xie, Bowen Wu, Ziqi Zhang, Xiaohui Shen, and Jian Yin. Fashion editing with multi-scale attention normalization. *CoRR*, abs/1906.00884, 2019.

[6] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.

[7] Jun Ehara and Hideo Saito. Texture overlay for virtual clothing based on PCA of silhouettes. In *ISMAR*, pages 139–142. IEEE Computer Society, 2006.

[8] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[10] Peng Guan, Loretta Reiss, David A. Hirshberg, Alexander Weiss, and Michael J. Black. DRAPE: dressing any person. *ACM Trans. Graph.*, 31(4):35:1–35:10, 2012.

[11] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4491, 2019.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: an image-based virtual try-on network. In *CVPR*, pages 7543–7552. IEEE Computer Society, 2018.

[13] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. Virtual try-on through image-based rendering. *IEEE Trans. Vis. Comput. Graph.*, 19(9):1552–1565, 2013.

[14] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. *arXiv preprint arXiv:1904.09261*, 2019.

[15] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE Computer Society, 2017.

[17] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, pages 2262–2267. IJCAI/AAAI, 2011.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[19] Nikolay Jetchev and Urs Bergmann. The conditional analogy GAN: swapping fashion articles on people images. In *ICCV Workshops*, pages 2287–2292. IEEE Computer Society, 2017.

[20] Youngjoo Jo and Jongyoul Park. SC-FEGAN: face editing generative adversarial network with user's sketch and color. *CoRR*, abs/1902.06838, 2019.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*. OpenReview.net, 2018.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019.

[24] Sumin Lee, Sungchan Oh, Chanho Jung, and Changick Kim. A global-local embedding module for fashion landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[25] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Trans. Multimedia*, 19(8):1946–1955, 2017.

[26] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[27] Jingyuan Liu and Hong Lu. Deep fashion analysis with feature map upsampling and landmark-driven attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[28] Yu Liu, Wei Chen, Li Liu, and Michael S. Lew. Swapgan: A multistage generative approach for person-to-person fashion style transfer. *IEEE Trans. Multimedia*, 21(9):2209–2222, 2019.

[29] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016.

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346. Computer Vision Foundation / IEEE, 2019.

[31] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap: seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36(4):73:1–73:15, 2017.

[32] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *ECCV (12)*, volume 11216 of *Lecture Notes in Computer Science*, pages 679–695. Springer, 2018.

[33] Damien Rohmer, Tiberiu Popa, Marie-Paule Cani, Stefanie Hahmann, and Alla Sheffer. Animation wrinkling: augmenting coarse cloth simulations with realistic-looking wrinkles. *ACM Trans. Graph.*, 29(6):157, 2010.

[34] 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.

[35] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2226–2234, 2016.

[36] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. *Comput. Graph. Forum*, 38(2):355–366, 2019.

[37] Wei Sun, Jawadul H. Bappy, Shanglin Yang, Yi Xu, Tianfu Wu, and Hui Zhou. Pose guided fashion image synthesis using deep generative model. *CoRR*, abs/1906.07251, 2019.

[38] 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.

[39] Andreas Veit, Balazs Kovacs, Sean Bell, Julian J. McAuley, Kavita Bala, and Serge J. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, pages 4642–4650. IEEE Computer Society, 2015.

[40] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 607–623. Springer, 2018.

[41] 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*, pages 8798–8807. IEEE Computer Society, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[43] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *CVPR*, pages 5840–5848. Computer Vision Foundation / IEEE, 2019.

[44] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 172–180. ACM, 2017.

[45] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *CoRR*, abs/1806.03589, 2018.

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514. IEEE Computer Society, 2018.

[47] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[48] Ruimao Zhang, Wei Yang, Zhanglin Peng, Pengxu Wei, Xiaogang Wang, and Liang Lin. Progressively diffused networks for semantic visual parsing. *Pattern Recognit.*, 90:78–86, 2019.

[49] 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*, page 33. ACM, 2012.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.