Toward Realistic Virtual Try-on Through Landmark-Guided Shape Matching

Guoqiang Liu [1] *, **Dan Song** [2] *, **Ruofeng Tong** [1] †, **Min Tang** [1]
[1]Zhejiang University
[2]Tianjin University
lgqfhwy@zju.edu.cn, dan.song@tju.edu.cn, trf@zju.edu.cn, tang_m@zju.edu.cn

Abstract

Image-based virtual try-on aims to synthesize the customer image with an in-shop clothes image to acquire seamless and natural try-on results, which have attracted increasing attentions. The main procedures of image-based virtual try-on usually consist of clothes image generation and try-on image synthesis, whereas prior arts cannot guarantee satisfying clothes results when facing large geometric changes and complex clothes patterns, which further deteriorates the afterwards try-on results. To address this issue, we propose a novel virtual try-on network based on landmark-guided shape matching (LM-VTON). Specifically, the clothes image generation progressively learns the warped clothes and refined clothes in an end-to-end manner, where we introduce a landmark-based constraint in Thin-Plate Spline (TPS) warping to inject finer deformation constraints around the clothes. The try-on process synthesizes the warped clothes with personal characteristics via a semantic indicator. Qualitative and quantitative experiments on two public datasets validate the superiority of the proposed method, especially for challenging cases such as large geometric changes and complex clothes patterns. Code will be available at https://github.com/lgqfhwy/LM-VTON.

Introduction

Virtual try-on can greatly improve the efficiency and experience of shopping apparel products online, which will also reduce the return rate. Due to the huge commercial value, it has been a hot topic in both academic and industrial communities in recent years. Traditional virtual try-on methods rely on 3D human body reconstruction and physically-based cloth simulation to animate the state of clothes on body, which still faces plenty of difficulties such as massive computations, manual labor and scanning devices that are not widely available (Song et al. 2019). Another way adopted by some companies for virtual try-on captures and stores the apparel appearance on various body shapes with the help of a shape controllable robot, which is later used for displaying try-on effects on virtual user bodies. Such kind of virtual try-on method costs tremendous manual labor that most retailers cannot afford. With the rapid development of deep learning, the most recent and appealing kind of virtual try-on method is based on image synthesis, which is efficient and saves lots of manual efforts.

Given a customer photo and an in-shop clothes image, the object of image-based virtual try-on is to synthesize a try-on image with the following requirements: (1) the person in the result image owns the same identity (such as face and hair) as the customer photo; (2) the clothes in the result image preserves the characteristics of the in-shop clothes image, such as color, texture, logo and text; and (3) the person in the result image puts on the new clothes seamlessly and naturally while preserving the other personal characteristics such as exposed skin or pants (when trying on tops). Although rapid progress (Han et al. 2018; Belongie, Malik, and Puzicha 2002; Yang et al. 2020) has been made on this topic, it remains challenging to synthesize a realistic try-on photo satisfying the above requirements.

Existing methods usually first warp the in-shop clothes image according to person representation and then synthesize the try-on image using the customer image and warped clothes. Due to the large geometric variations between the in-shop clothes and target clothes on person, the clothes warping stage is the most challenging part in this task, which also plays a vital role to synthesize realistic try-on image. VITON (Han et al. 2018) warps the in-shop clothes image by shape context matching (Belongie, Malik, and Puzicha 2002) with the generated clothing mask, which narrows the geometry gaps. Instead of hand-crafted shape context matching, CP-VTON (Wang et al. 2018a) proposes a new learnable thin-plate spline (TPS) transformation network to describe the image warping. To make the TPS (Belongie, Malik, and Puzicha 2002) warping more stable and get rid of obvious distortions, ACGPN (Yang et al. 2020) uses a second-order difference constraint. However, as shown in Fig. 1, the results of previous methods still contain artifacts.

To achieve more realistic results, we propose LM-VTON, a virtual try-on network based on landmark-guided shape matching. As shown in Fig. 2, the proposed method consists of two main steps, where the first step generates the clothes image consistent to customer's pose and shape and the second step synthesizes the try-on image with original person seamlessly and naturally putting on the clothes. Specifically, for Step I, we introduce a novel landmark-based constraint to TPS warping(Belongie, Malik, and Puzicha 2002),

*Equal contribution.
†Corresponding author.

Figure 1: Try-on results and corresponding zoomed results of the proposed method and previous state-of-the-art methods, i.e., VITON (Han et al. 2018), CP-VTON (Wang et al. 2018a) and ACGPN (Yang et al. 2020). Our method generates photo-realistic try-on results, which preserve clothing characteristics (especially in first row), perform natural warping (especially in second row), have clear texture (especially in third row) and deal well with large pose variations (especially in last row).

which poses a uniform constraint around the clothes and produces reasonable deformations with less distortions. Based on the better warped clothes, clothes refinement generates target clothes which preserve clothing characteristics with less blurry texture. For Step II, we learn a semantic indicator to facilitate the try-on module be perceptible to the contents of target clothes and the preserved characteristics from customer image, which makes the try-on seamless and natural. Extensive experiments show that the proposed method can handle well with large geometric changes and complex clothes patterns and achieve the state-of-the-art performance on the public dataset Zalando (Han et al. 2018) and MPV (Dong et al. 2019).

The main contributions of our work are summarized as follows:

- We propose a new virtual try-on network, i.e., LM-VTON, which addresses high-quality clothes generation and semantic synthesis of realistic try-on images.
- We for the first time introduce landmark-based constraint to TPS warping, which injects finer deformation constraints around the clothes, and the better warped clothes will further contribute to the afterwards clothes refinement and virtual try-on.
- Extensive experiments demonstrate the superiority of the proposed method against the state-of-the-art methods on two public datasets both qualitatively and quantitatively.

Related Work

Image Synthesis

Deep learning has aided many research areas (Xu et al. 2020, 2019; Nie et al. 2020), among which image synthesis plays an important role. For example, Gatys et al. (Gatys, Ecker, and Bethge 2015) introduce the first CNN-based method for texture synthesis. Pix2Pix (Isola et al. 2017) uses a conditional generative adversarial network (Goodfellow et al. 2014) to learn a mapping from input to output images. Later Pix2PixHD (Wang et al. 2018b) is proposed to synthesize high-resolution photo-realistic images. In recent years, person image synthesis has attracted increasing attentions. Lassner et al. (Lassner, Pons-Moll, and Gehler 2017) present the first image-based generative model of people in clothing for the full body. FashionGAN (Zhu et al. 2017) changes the fashion items on a person in image with text descriptions. Pose-guided person image synthesis is also an interesting subject that generates a clothed person images with conditioned pose (Zheng et al. 2019). PG^2 (Ma et al. 2017) utilizes a two-stage GANs architecture to generate the person image based on pose keypoints. BodyROI7 (Ma et al. 2018) synthesizes person images based on a novel, two-stage reconstruction pipeline that learns a disentangled representation of image factors including foreground, background and human pose. DSCF (Siarohin et al. 2018) synthesizes person images conditioned on the appearance and the pose by introducing deformable skip connections and nearest-neighbour loss to U-Net (Ronneberger, Fischer, and Brox 2015). Dong et al. (Dong et al. 2018) propose a Warping-GAN to resolve the challenges induced by geometric variability and spatial displacements.

Image-Based Virtual Try-On

As an important part of person image synthesis, image-based virtual try-on aims to generate photo-realistic try-on images with a standard in-shop clothes and a customer image in different clothes and various poses. Due to large geometric variations, such a task is a challenging image synthesis problem. VITON (Han et al. 2018) for the first time ad-

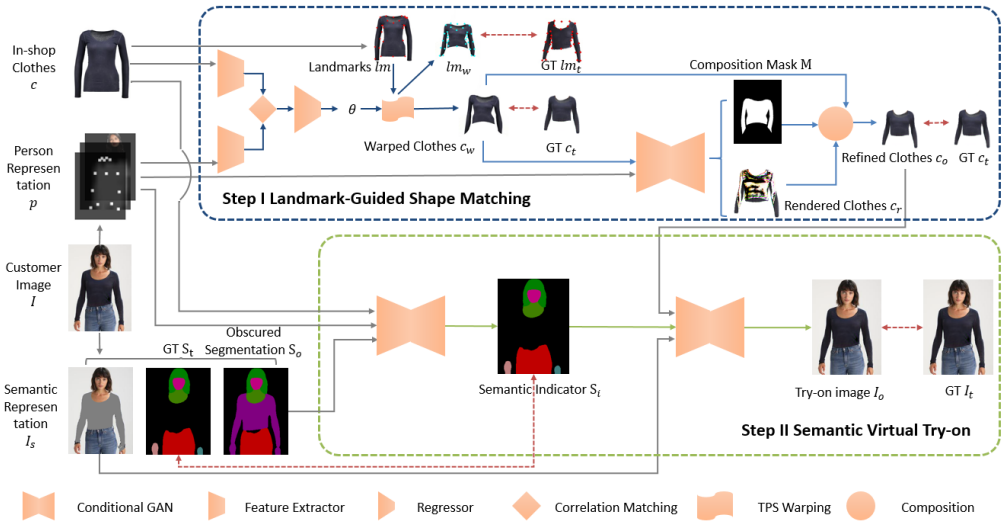

Figure 2: Overview of the proposed method. It should be noted that in the training stage the customer image I is the same as ground-truth I_t, because of the difficulty to collect triplets. Clothing-agnostic representation (i.e., p, I_s and S_o) is designed to solve this issue. However, in testing stage, the clothes in customer image is different from the in-shop clothes for fair evaluation.

dresses this task, and proposes the clothing-agnostic person representation and coarse-to-fine strategy. SP-VTON (Song et al. 2019) improves the person representation to keep body shapes, which can deal well with more complex clothes types. Instead of hand-crafted shape context matching adopted by VITON, CP-VTON (Wang et al. 2018a) proposes a new learnable thin-plate spline (TPS) transformation network to describe the clothes warping. For large pose variation and challenging clothes such as long-sleeved pattern, simply applying Thin-Plate Spline (TPS) transformation cannot guarantee precise transformation, which will also deteriorates the afterwards try-on synthesis. Yang et al. (Yang et al. 2020) introduce a second-order difference constraint to get rid of obvious distortions, where they have ability against drastic deformations but fail to deal with challenging cases of large geometric changes. Issenhuth et al. (Issenhuth, Mary, and Calauzènes 2020) propose a novel idea to get rid of human parser and pose estimator at inference time. We propose a landmark-guided shape matching to pose explicit correspondence between target clothes and warped clothes, which produces a better clothes warping and facilitates the afterwards clothes refinement and try-on.

Our Approach

Given an in-shop clothes c and a customer image I of a person wearing in clothes c', the goal of virtual try-on is to synthesize the try-on image I_o of the wearer in the new clothes c_o. As shown in Fig. 2, the proposed framework consists of two main modules. Firstly, the landmark-guided shape matching module (LGSM) uses in-shop clothes c and person representation p as input and learns target clothes c_o, whose shape are consistent with person's pose and shape. At this stage an end-to-end pipeline is trained, which introduces a landmark-based constraint in TPS warping and finally generates refined clothes c_o based on the warped clothes c_w and person representation p. Secondly, the semantic virtual try-on module (SVTO) utilizes person representation p, semantic representation I_s and refined clothes c_o to synthesize the try-on image I_o, which is trained to inpaint the refined clothes while preserving the other characteristics of original customer image to the final try-on image.

Preliminary

The customer image I is preprocessed to two kinds of representations, i.e., person representation p and semantic representation I_s. Person representation is designed for warping clothes, where we deform standard in-shop clothes c and generate the target clothes c_o consistent to person's pose and shape. Semantic representation I_s and obscured segmentation S_o are used to make the try-on module perceptible to the contents of target clothes and original personal characteristics to be preserved.

Person representation. We adopt the clothing-agnostic person representation used by VITON and CP-VTON (Han et al. 2018; Wang et al. 2018a), which contains three components: (1) pose heatmap acquired by (Cao et al. 2017), an 18-channel feature map with each channel corresponding to one human pose keypoint; (2) body shape map obtained by (Gong et al. 2017), a 1-channel feature map of a blurred binary mask indicating the clothing-agnostic person contour; and (3) head map, an RGB image including face and hair parsed by (Gong et al. 2017). In summary, the identity representation concatenates these components together to form a map with size $256 \times 192 \times 22$.

Semantic representation. Semantic segmentation has been proved by previous methods (Dong et al. 2019; Yang

Figure 3: Landmarks example for long-sleeved top. (a): detected landmarks on in-shop clothes; (b): detected landmarks on cropped ground-truth clothes. (c): transform the landmarks of (b) back for ground-truth clothes on person.

et al. 2020) for successfully indicating the contents of target clothes and original preserved characteristics. We adopt (Gong et al. 2017) for human parsing, which assigns 20 semantic labels for person image. Semantic representation I_s is obtained by removing the parsed clothes (i.e., the clothes desired to take off) from customer image, which indicates that the rest of I is the original characteristic that can be relied. Obscured segmentation S_o is acquired by combining related segments (e.g., the top clothes is related to arms and hands) into one segmentation. S_o is used together with in-shop clothes c (i.e., the clothes desired to try-on) to indicate the original contents in I conflicting to target clothes.

Clothing Landmarks. We introduce landmarks correspondence and add an explicit constraint for shape matching. We detect clothing landmarks via HRNet(Sun et al. 2019) for both in-shop clothes c and the ground-truth clothes c_t, respectively denoted as lm and lm_t. As different landmarks are set for different clothes types, we utilize Faster-RCNN (Ren et al. 2015) to detect different clothes categories and obtain the clothing bounding box. Fig. 3 gives an example for the landmarks of long-sleeved top. To achieve better landmarks detection for the clothes on person, we utilize the bounding box to crop and scale the image and then perform landmarks detection.

Landmark-Guided Shape Matching (LGSM)

This module aims to learn the geometric matching between in-shop clothes and warped clothes which is consistent to customer's pose and shape. For large geometric changes induced by pose and challenging clothes such as long-sleeved pattern, simply applying Thin-Plate Spline (TPS) transformation cannot guarantee precise transformation, which will also deteriorate the afterwards try-on synthesis. To address this issue, Yang et al. (Yang et al. 2020) introduce a second-order difference constraint to get rid of obvious distortions. We propose a landmark-guided shape matching to inject explicit correspondence between ground-truth clothes and warped clothes, which produces a better warping.

This module is trained in an end-to-end manner under the total loss as $L_{LGSM} = L_{warping} + L_{refinement}$, where $L_{warping}$ is designed for clothes warping and $L_{refinement}$ is designed for clothes refinement. The base network of clothes warping consists of two feature extractors to extract high-level features of in-shop clothes and person representation, a correlation layer to combine two features, a regressor to produce the spatial transformation, and a TPS transformation module to warp clothes. $L_{warping}$ is composed of the pixel-wise $L1$ loss between warped clothes c_w and ground-truth clothes c_t, and landmarks loss between the warped landmarks lm_w and ground-truth landmarks lm_t. The loss for clothes warping is defined as:

$$L_{warping} = \lambda_l \left\| T_\theta(c) - c_t \right\|_1 + \lambda_{lm} \left\| T_\theta(lm) - lm_t \right\|_1 \quad (1)$$

where T_θ denotes the TPS transformation and $|| \cdot ||_1$ represents L_1 loss, λ_l and λ_{lm} are used to balance each item.

Clothes refinement learns to supplement or eliminate contents of warped clothes according to person representation p and warped clothes c_w. The generative clothes refinement is inspired by the try-on module of CP-VTON (Wang et al. 2018a). As shown in Fig. 2, given the person representation p and warped clothes c_w, a rendered clothes c_r and a composition mask M are simultaneously acquired via U-Net (Ronneberger, Fischer, and Brox 2015). Afterwards, the generated rendered clothes c_r and warped clothes c_w are fused using composition mask M to synthesize the final clothes c_o, where

$$c_o = M \odot \hat{c_w} + (1 - M) \odot c_r \quad (2)$$

and $\odot$ represents element-wise matrix multiplication.

The loss for clothes refinement $L_{refinement}$ consists of L_1 loss, VGG perceptual loss (Johnson, Alahi, and Fei-Fei 2016) and mask regularization loss, which is defined as:

$$\begin{aligned} L_{refinement} &= \lambda_{l2} L_1(c_o, c_t) + \lambda_v L_{VGG}(c_o, c_t) + \lambda_m L_m(M) \\ &= \lambda_l \left\| c_o - c_t \right\|_1 + \lambda_v \sum_{i=1}^{5} \lambda_i \left\| \phi_i(c_o) - \phi_i(c_t) \right\|_1 + \lambda_m \left\| \mathbf{1} - M \right\|_1 \end{aligned} \quad (3)$$

where $\phi(\cdot)$ denotes the feature map of $i-th$ layer in a visual perceptual network VGG19 (Simonyan and Zisserman 2014) pre-trained on ImageNet. The regularization L_m encourages the generated clothes c_o to composite more from the warped clothes c_w. λ_{l2}, λ_v and λ_m are the weighting parameters for each loss term in Eq.3.

Semantic Virtual Try-on (SVTO)

This module aims to synthesize the try-on image from refined clothes c_o and semantic person representation I_s. Inspired by ACGPN (Yang et al. 2020), we first learn a semantic indicator S_i to indicate the contents of customer image to be preserved. Based on such an indicator S_i, the try-on image is synthesized by utilizing the refined clothes c_o and semantic person representation I_s.

As shown in Fig. 2, given the in-shop clothes c, person identity representation p and obscured semantic representation S_o (explained in Preliminary Section), we train a conditional GAN (Wang et al. 2018b) to generate the semantic indicator S_i. U-Net (Ronneberger, Fischer, and Brox 2015) is adopted as the generator while pix2pixHD (Wang et al. 2018b) is utilized as the discriminator. The pixel-wise cross entropy loss L_{ce} and adversarial loss L_{adv} are adopted to train the cGAN, which are defined as:

$$L_{ce} = -\sum_{i=1}^{n} S_i(i) \log S_t(i) \quad (4)$$

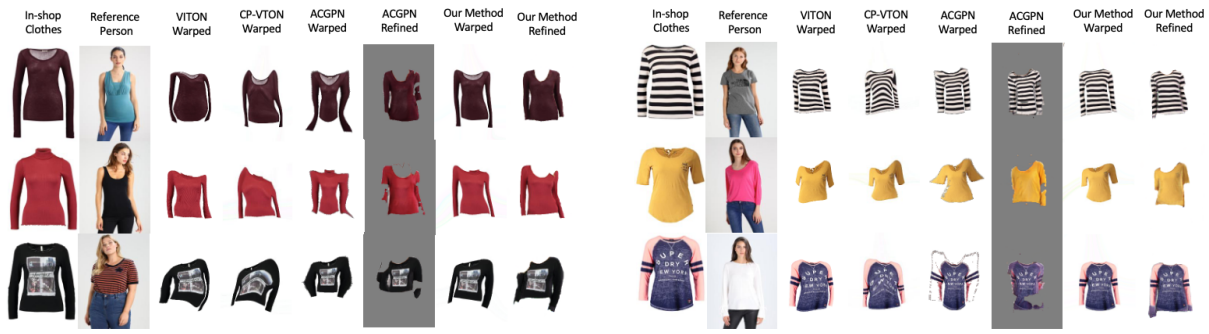

Figure 4: Evaluation on warped clothes and refined clothes with Zalando dataset.

$$L_{adv} = E_{x,y}\left[\log\left(\mathcal{D}\left(x,y\right)\right)\right] + E_{x}\left[\log\left(1-\mathcal{D}\left(x,\mathcal{G}\left(x\right)\right)\right)\right] \tag{5}$$

where S_i is the output semantic segmentation, S_t is the corresponding ground-truth, and n is the number of pixels. In Eq. 5, x denotes network output and y represents the ground-truth. Consequently, the loss for generating S_i is:

$$L_{semantic} = \lambda_c L_{ce} + \lambda_a L_{adv} \tag{6}$$

Given semantic person representation I_s, semantic indicator S_i and refined clothes c_o as input, the same architecture is designed for the afterwards cGAN. Besides the adversarial loss L_{adv} (Eq. 5), the pixel-wise L_1 loss and perceptual VGG loss L_{VGG} between the try-on image and ground-truth are used for training. The loss for generating try-on image I_o is:

$$L_{generation} = \lambda_{l3} L_1 + \lambda_{a2} L_{adv} + \lambda_{v2} L_{VGG} \tag{7}$$

where L_1 and L_{VGG} are defined the same as in Eq. 3 except the input. λ_{l3}, λ_{a2} and λ_{v2} are the corresponding weighting parameters.

Experiments and Analysis

Dataset

As most image-based virtual try-on methods do (Han et al. 2018; Wang et al. 2018a; Yang et al. 2020), we use the dataset collected by Han et al.(Han et al. 2018) (denoted as Zalando dataset in this paper) to comprehensively compare the proposed method with the state-of-the-art methods. Zalando dataset contains around 19,000 pairs of front-view woman photo and top clothing image (i.e., I and c). There are 16,253 available pairs for training, 14,221 of which are organized as training set and the rest are used as testing set. Since in practical scenarios the customer should wear different clothes, the in-shop clothes images within the testing set are randomly shuffled.

As most clothes in the collected Zalando dataset are shorts, we use another dataset MPV (Dong et al. 2019) to show the ability of the proposed method to deal with more challenging conditions. MPV dataset contains various poses and views such as whole front, whole back, half front and half back, and also have more long-sleeved tops. We select the whole and half front view of MPV dataset to evaluate our approach. The selected data are split to the training set and the testing set with 16,585 and 3,344 pairs respectively.

Implementation Details

Clothing landmarks prediction. Both Zalando dataset(Han et al. 2018) and MPV dataset(Dong et al. 2019) do not provide landmarks for clothes. We first train landmarks detection model HRNet(Sun et al. 2019) on DeepFashion2 dataset (Ge et al. 2019) which has annotated landmarks for fashion images, then predict landmarks for clothes within Zalando and MPV datasets. It is worth noting that DeepFashion2 dataset contains 491K diverse images of 13 popular clothing categories, while different clothes types have different numbers of landmarks. We train Faster-RCNN (Ren et al. 2015) to recognize clothes categories and obtain clothes bounding box. Then we train HRNet(Sun et al. 2019) to detect landmarks for 13 types of popular clothes, where the output channel is modified to 192, each representing a landmark's position of one clothes category.

Training. As shown in Fig. 2, landmark-guided shape matching module and semantic virtual try-on module are separately trained. As it is difficult to collect a triplet (customer image I, in-shop clothes c, ground-truth try-on image I_t), existing methods make I the same as I_t and design clothing-agnostic person representation. For LGSM module, we set $\lambda_l = \lambda_l 2 = \lambda_v = \lambda_m = 1$ and $\lambda_{lm} = 0.01$. For SVTO module, we set $\lambda_c = \lambda_2 = 1$, $\lambda_{v2} = 10$. Each module is trained for 20 epochs with batch size 4. The learning rate is initialized to 0.0002 and the Adam optimizer is adopted with the hyper-parameter $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All the codes are implemented on PyTorch and run on one NVIDIA 2080Ti GPU.

Testing. The testing procedure almost follows the training phase as illustrated in Fig. 2. All the red lash lines in the figure should be removed, and we do not need to predict landmarks for clothes in the testing procedure. It should also be noted that for testing, the clothes in customer image are different from the in-shop clothes.

Figure 5: Evaluation on try-on images with Zalando dataset.

Qualitative Results

The qualitative results of proposed method are evaluated by comparing with VITON (Han et al. 2018), CP-VTON (Wang et al. 2018a), and ACGPN (Yang et al. 2020). Since all of the methods contain clothes warping and try-on procedures, we compare the results of these two key procedures.

Fig. 4 shows the visual results of clothes warping. Both ACGPN (Yang et al. 2020) and the proposed method refine the warped clothes, so the refined results are also shown in the figure. VITON (Han et al. 2018) first generates a clothing mask consistent to person representation, then does shape context matching for correspondence points to perform TPS warping. CP-VTON (Wang et al. 2018a) learns the parameters controlling TPS warping in an end-to-end pipeline under the supervision of ground-truth clothes c_t. It is difficult to tell whether CP-VTON competes VITON from visual warped clothes, but using learnable parameters in an end-to-end pipeline instead of hand-crafted shape context matching should be the advantage. Besides learning deformation parameters, ACGPN (Yang et al. 2020) introduces a second-order difference constraint to make the warping stable for handling complex textures. Consequently, the results of ACGPN warped clothes show performance against drastic deformations, which also have cannot deal well with large geometric changes.

Besides learning deformation parameters, we introduce a landmark-based constraint which injects a uniform constraint around the clothes. Image warping is limited to describe clothes deformation, especially when the limbs are overlapped with the torso. Generative clothes refinement aims to refine the warped clothes to be more close to ground-truth clothes, and it relies on the quality of warped clothes. Therefore, we achieve better refined results based on a better warped results.

Fig. 5 shows the visual results of virtual try-on. Although VITON (Han et al. 2018) and CP-VTON (Wang et al. 2018a) do not perform clothes refinement, they can obtain nice try-on results for plain clothes with little texture. However, they cannot clearly preserve texture or complex patterns. Attributing to the clothes refinement which generates clothes more close to the ground-truth clothes, the texture or complex patterns of the results are less blurred. However, ACGPN (Yang et al. 2020) cannot deal well with large pose variations, they achieve poor results at sleeves. From Fig. 5 we can find that the proposed method generates more realistic try-on results, where the clothes characteristics are preserved (e.g., the texture and stripes in the figure) and the person pose can be diverse.

It is also worth noting that the red turtle neck sweater in Fig. 5 is turned into a round-neck sweater, which is the limitation of current works (including the proposed method). The main reason is that the person representation adopted in existing works is not strictly clothing-agnostic. On one hand, the body shape is represented via down-sampling the clothed body segmentation to a low resolution, which does not totally get rid of the original clothes. On the other hand, the parsing tool adopted by existing works (i.e., LIP (Gong et al. 2017)) does not set a label for neck, so the neck and chest exposed out of sweater is parsed as background. Therefore, a better clothing-agnostic person representation is desired for preserving the clothing types, especially at the collars.

Method	Zalando	MPV
VITON (Han et al. 2018)	2.515	2.394
CP-VTON (Wang et al. 2018a)	2.600	2.519
ACGPN (Yang et al. 2020)	2.694	-
Our method	**2.765**	**3.043**

Table 1: Inception score (IS) on Zalando and MPV dataset.

Quantitative Results

We adopt Inception Score (IS) (Salimans et al. 2016) and Structural SIMilarity (SSIM) (Wang et al. 2004) for quantitative evaluation. Inception score (IS) is usually used to evaluate the synthesis quality and higher IS indicates diverse and semantically meaningful images. The results of IS on Zalando dataset and MPV dataset is illustrated in Tab. 1, and the proposed method performs best. Structural SIMilarity (SSIM) is used to measure the structural similarity between the synthesized image and the ground-truth, where higher values indicate better results. Since the testing set of Zalando dataset does not prepare ground-truth image, we

Figure 6: Ablation study on landmark-guided shape matching module. (a): in-shop clothes, (b): ground-truth, (c): warped clothes w/o landmark constraint, (d): warped clothes w/ landmark constraint, (e): refined clothes, (f): try-on warped clothes w/o landmark constraint, and (g): try-on refined clothes (the proposed).

Method	SSIM
VITON (Han et al. 2018)	0.639
CP-VTON (Wang et al. 2018a)	0.705
Our method	**0.888**

Table 2: SSIM results on MPV dataset.

cannot compute SSIM on Zalando dataset. The SSIM results on MPV dataset is shown in Tab. 2, and we achieve the best results compared with VITON (Han et al. 2018) and CP-VTON (Wang et al. 2018a). Since ACGPN (Yang et al. 2020) only releases the inference model, we cannot re-train on MPV dataset.

User Study

We follow the rules introduced in the most recent work WUTON(Issenhuth, Mary, and Calauzènes 2020) for user study. A/B tests are performed on 10 users, and each volunteer has to vote 100 times between ACGPN(Yang et al. 2020) and our synthesized images. The users choose the results generated by our method 74.3% of the time.

Ablation Study

To illustrate the effectiveness of specific designs in the landmark-guided shape matching, we show the warped clothes without using landmarks, warped clothes using landmarks, refined clothes and corresponding try-on results in Fig. 6. From the results we can observe that the landmark constraint plays an important role in preserving clothing characteristics such as texture, strips and sleeves. However, warping the in-shop clothes is limited to express the target clothes due to large variations. Therefore, it is necessary to propose clothes refinement to generates clothes which are more consistent with person representation. Also validated by the results in Fig. 6, synthesis with refined clothes generates more realistic try-on images.

We also compute the numerical results (i.e., IS and SSIM) with/without landmarks. As shown in Table.3, the introduced landmarks also improve the numerical performance.

Method	IS		SSIM
	Zalando	MPV	MPV
w/o landmark	2.663	2.565	0.801
Our method	**2.765**	**3.043**	**0.888**

Table 3: IS and SSIM results for ablation study.

	CP-VTON	ACGPN	Our method
Total	298ms	834ms	540 ms

Table 4: Inference runtime.

Inference Runtime Analysis

In table 4, we compare the inference runtime of our method with those of CP-VTON(Wang et al. 2018a) and ACGPN(Yang et al. 2020) with a 2080ti GPU. Compared with existing related works, we additionally introduce landmarks detection. However, landmarks detection is only performed during the training stage and we do not need landmarks in testing stage. Our method consumes more time than CP-VTON(Wang et al. 2018a) because of the additional clothes refinement and semantic indication. ACGPN(Yang et al. 2020) need more time than our method to generate the clothing mask and compute the second-order difference.

Conclusion

In this paper, we propose a virtual try-on network based on landmark-guided shape matching, i.e., LM-VTON, to synthesize realistic try-on results. For clothes image generation, we introduce a landmark-based constraint to inject finer deformation constraints around the clothes, and based on the warped clothes we perform the generative clothes refinement. For try-on image synthesis, we learn a semantic indicator to facilitate the try-on module be perceptible to the contents of target clothes and original preserved characteristics. We conduct qualitative and quantitative experiments on two public datasets, and the results demonstrate the superiority of the proposed method, especially for challenging cases such as large pose variations and complex clothes patterns.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61902277, 61832016, 61972342, 61972341 and 61732015), the Open Project Program of the State Key Lab of CAD & CG, Zhejiang University (Grant No. A2012), and the China Postdoctoral Science Foundation (2020M680884).

References

Belongie, S.; Malik, J.; and Puzicha, J. 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (4): 509–522.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. In *Advances in Neural Information Processing Systems*, 472–482.

Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; and Yin, J. 2019. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, 9026–9035.

Gatys, L.; Ecker, A. S.; and Bethge, M. 2015. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, 262–270.

Ge, Y.; Zhang, R.; Wang, X.; Tang, X.; and Luo, P. 2019. 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*, 5337–5345.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 932–940.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

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

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Issenhuth, T.; Mary, J.; and Calauzènes, C. 2020. Do Not Mask What You Do Not Need to Mask: a Parser-Free Virtual Try-On. *arXiv preprint arXiv:2007.02721* .

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Lassner, C.; Pons-Moll, G.; and Gehler, P. V. 2017. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, 853–862.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, 406–416.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 99–108.

Nie, W.; Ren, M.; Liu, A.; Mao, Z.; and Nie, J. 2020. M-GCN: Multi-branch graph convolution network for 2D image-based on 3D model retrieval. *IEEE Transactions on Multimedia* .

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Siarohin, A.; Sangineto, E.; Lathuilière, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3408–3416.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Song, D.; Li, T.; Mao, Z.; and Liu, A.-A. 2019. SP-VITON: shape-preserving image-based virtual try-on network. *Multimedia Tools and Applications* 1–13.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*.

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

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018b. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4): 600–612.

Xu, N.; Liu, A.-A.; Wong, Y.; Nie, W.; Su, Y.; and Kankanhalli, M. 2020. Scene Graph Inference via Multi-Scale Context Modeling. *IEEE Transactions on Circuits and Systems for Video Technology* .

Xu, N.; Zhang, H.; Liu, A.-A.; Nie, W.; Su, Y.; Nie, J.; and Zhang, Y. 2019. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Transactions on Multimedia* 22(5): 1372–1383.

Yang, H.; Zhang, R.; Guo, X.; Liu, W.; Zuo, W.; and Luo, P. 2020. Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7850–7859.

Zheng, N.; Song, X.; Chen, Z.; Hu, L.; Cao, D.; and Nie, L. 2019. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia*, 266–274.

Zhu, S.; Urtasun, R.; Fidler, S.; Lin, D.; and Change Loy, C. 2017. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*, 1680–1688.