2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.02083

Virtual Try-On with Pose-Garment Keypoints Guided Inpainting

Zhi Li[1,2] Pengfei Wei[1] Xiang Yin[1] Zejun Ma[1] Alex C. Kot[2]
[1]Bytedance Ltd. [2]Nanyang Technological University
{zhi.li.2023, pengfei.wei, yinxiang.stephen, mazejun}@bytedance.com eackot@ntu.edu.sg

Abstract

Virtual try-on is an important technology supporting online apparel shopping, which provides consumers with a virtual experience to fit garments without physically wearing them. Recently, the image-based virtual try-on has received growing research attention. However, the synthetic results of existing virtual try-on methods usually present distortions in garment shape and lose pattern details. In this paper, we propose a pose-garment keypoints guided inpainting method for the image-based virtual try-on task, which produces high-fidelity try-on images and well preserves the shapes and patterns of the garments. In our method, human pose and garment keypoints are extracted from source images and constructed as graphs to predict the garment keypoints at the target pose. After which, the predicted keypoints are used as guide information to predict the target segmentation map and warp the garment image. The try-on image is finally generated with a semantic-conditioned inpainting scheme using the segmentation map and recomposed person image as conditions. To verify the effectiveness of our proposed method, we conduct extensive experiments on the VITON-HD dataset under both paired and unpaired experimental settings. The qualitative and quantitative results show that our method significantly outperforms prior methods at different image resolutions. The codes repository link is https://github.com/lizhi-ntu/KGI.

1. Introduction

Online shopping has exploded rapidly in the past decade because of its convenience and high cost-effectiveness. Buying clothes and other daily necessities without leaving home is gradually becoming a mainstream lifestyle among the young generation. With the popularity of online apparel shopping, virtual try-on technology has received growing interests from fashion brands and online retail platforms in recent years [14].

Virtual try-on technology aims to improve the online shopping experience by visualizing the fitting results without requiring consumers to physically wear the garments.

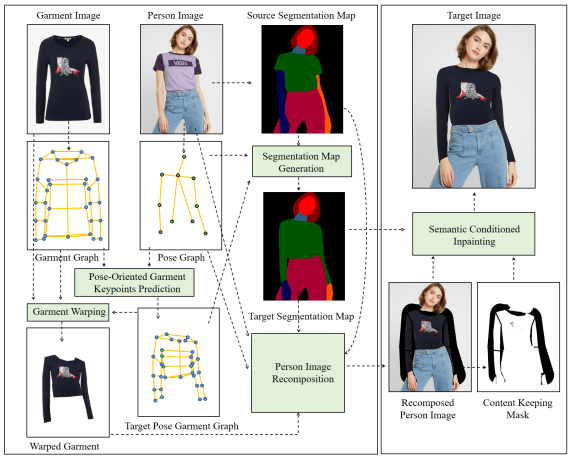

Figure 1. Pose-garment keypoints guided inpainting framework.

According to whether 3D modeling is used, existing virtual try-on methods are categorized into image-based approach and 3D model-based approach. Since capturing 3D information requires additional sensory devices and the 3D modeling of person and garments costs more expense, image-based technology has attracted more attention recently.

Image-based virtual try-on aims to produce high-fidelity fitting results given person and garment images. The generated try-on image is expected to preserve the appearance and pose of the given person image but replace the cloth region with the given garment image. Most existing approaches share the same idea of firstly warping the garment image for target pose then blending the person image with warped garment image and target segmentation map. However, a common issue has always existed, that is, inappropriate warping of the garment image or inaccurate estimation of the target segmentation map usually results in the distortion of the garment shape. For instance, when part of the cloth is occluded due to the person pose, as shown in the *Ground-Truth* in Figure. 2, the warped garment generated by existing methods, e.g., thin-plate spline transformation (TPS) [3, 9], suffers from severer distortion at the overlapping part, as observed in the upper-side of Figure. 2. Other distortions may happen at cuff or neckline. Moreover,

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.02083

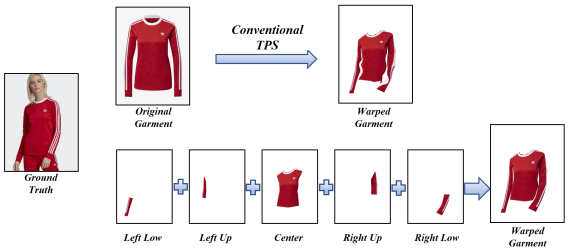

Figure 2. Garment warping results: conventional TPS *v.s.* ours.

the blending procedure may also blur the garment image or lead to the loss of pattern details. To alleviate these issues, we propose a pose-garment keypoints guided inpainting method (KGI) for image-based virtual try-on.

The motivation of KGI is that our interested information in the given garment and person images, i.e., the garment shape and the person pose, can be well represented by two sets of keypoints, namely original garment keypoints and pose keypoints. Using the two sets of keypoints, we can generate pose-oriented garment keypoints, precisely representing the warped garment shape following the person pose. Subsequently, we use pose-oriented garment keypoints to guide fine-grained garment warping and target segmentation map estimation. In the final try-on image generation, we further apply a semantic-conditioned inpainting scheme to avoid the issue of blurring and loss of pattern details. The framework of KGI is shown in Figure 1.

More specifically, we first extract keypoints from the given garment and person images, respectively. The two sets of extracted keypoints are constructed as graphs, and then fed into a two stream graph convolutional network to predict pose-oriented garment keypoints. Next, we use the predicted keypoints to perform garment warping and generate target segmentation map. For garment warping, we separate the garment into five sub-segments, namely left low, left up, center, right up and right low as shown in lower-side of Figure 2, and then use the paired original/pose-oriented kerpoints to warp each sub-segment individually. The final warped garment integrates the five warped sub-ones. By doing so, we can handle the overlapping deformation. For target segmentation map, we generate it using pose keypoints, pose-oriented garment keypoints, and the source segmentation map extracted from the given person image.

Finally, the warped garment image and the target segmentation map, together with the given person image, source segmentation map, and pose keypoints, are input for the try-on image generation. To avoid issues of blurring and loss of pattern details, we recompose a person image with incomplete fitting areas and inpaint the missing regions according to the semantic segmentation maps. Precisely, the given person image combined with pose keypoints and the source segmentation map are used to generate a person image with the garment and arms region cropped off. The warped garment are then populated, conditioned on target segmentation maps, into the cropped area of the person image, resulting in the recomposed person image. To fill up the missing area of the recomposed person image, we adopt inpainting conditioned on the target segmentation map. Note that the target segmentation map plays a role in providing the body semantic information, and this is why we call the final step semantic-conditioned inpainting. The main contribution of the paper is summarized as:

- We propose a pose-garment keypoints guided inpainting method for the image-based virtual try-on task.
- We propose a graph-based model to extract the pose-oriented garment keypoints for garment warping and target segmentation map estimation.
- We propose a semantic-conditioned inpainting scheme to generate the final try-on image.
- We conduct extensive experiments to verify the effectiveness of KGI and show quantitative and qualitative improvements compared with prior methods.

2. Related Works

In literature, virtual try-on methods are mainly divided into 3D model based [2, 17, 19, 28, 21] and 2D image based. With the exquisite modeling of person and garment, 3D model based approaches are capable of visualizing the fitting results at different views. However, the requirement of additional devices for information retrieval and the high computational cost of 3D modeling highly constrains the application of these methods, especially in scenarios with limited resources.

Recently, image-based virtual try-on techniques have been widely concerned and developed rapidly. Image-based virtual try-on aims to generate a fitting image based on the given person and garment images. Most existing methods follow the pipeline of warping the given garment image and blending the cloth region of the given person image with the warped garment. For instance, Han *et al*. [9] propose a method to blend the cloth region with coarse-to-fine strategy. They firstly generate a coarse try-on image and a cloth mask using a multi-task regression network. Afterward, they leverage the cloth mask for cloth warping with the TPS and blend the coarse image with a refinement network. Wang *et al*. [22] propose a method to learn a TPS transformation for cloth warping and a composition mask to ensure the smoothness of generated images while composing the warped clothes and the rendered image. Han *et al*. [8] introduce an appearance-flow-based generative model which estimates a dense flow between source and target clothing regions for cloth warping. Yu *et al*. [26] present a three-stage design strategy including cloth warp, segmentation

map prediction, and fusion for fine-scale image synthesis. Yang *et al*. [25] propose to progressively predict the desired layout of try-on and warp the cloth image according to the generated layout. An inpainting module is then applied to adaptively produce fitting results. Choi *et al*. [6] propose a method which firstly uses the segmentation map to guide the try-on synthesis, then roughly fits the cloth to person image, and lastly handles the misaligned areas with Alignment Aware Segment normalization. The garment warping and the segmentation generation are usually performed individually in existing works. Considering the misalignment between the warped clothes and the segmentation map results in the artifacts in the generated results, Lee *et al*. [15] propose a method with a multi-task condition generator for garment warping and segmentation map generation. He *et al*. [10] propose a method based on StyleGAN for appearance flow estimation with the global style features. To simplify the existing multi-stage based method, Bai *et al*. [1] propose a single-stage try-on framework which performs multi-flow estimation using a deformable attention scheme.

Instead of explicitly performing cloth warping, some virtual try-on methods are developed with delicate feature learning strategies. Raj *et al*. [20] propose a method transfers the cloth of person images with the garment image via feature disentanglement. Ge *et al*. [7] propose a method to produce highly-realistic try-on images by disentangling clothes warping, skin generation, and other essential components. Considering that the estimation of person segmentation is sub-optimal and time-consuming, some virtual try-on methods are developed without using segmentation map. Issenhuth *et al*. [13] propose a try-on method in a student-teacher paradigm. Chen *et al*. [5] propose a co-attention feature-remapping framework to generate the try-on results according to the driven-pose sequence in two stages.

3. Methods

As shown in Figure 1, the proposed Keypoints Guided Inpainting (KGI) method consists of three stages. At Stage 1, we predict pose-oriented garment keypoints. At Stage 2, the pose-oriented garment keypoints are used to guide cloth warping and target segmentation map generation. Afterwards, a person image is recomposed using the warped garment image and the target segmentation map. At Stage 3, the recomposed person image is progressively inpainted using a diffusion model conditioned on the target segmentation map and known pixels.

3.1. Pose-Oriented Garment Keypoints Prediction

We formulate the pose-oriented garment keypoints prediction as a garment graph regression problem conditioned on the pose graph, and devise a two stream graph neural network to solve the problem.

(a) (b) (c) (d)

Figure 3. Illustration of pose and garment graphs.

The keypoints are extracted with the off-the-shelf models for human pose estimation [4] and fashion landmark detection. Then we represent the keypoints with the graph structure composed of nodes and edges so as to better model the relationships of different keypoints. As presented in Figure 3 (a), the pose graph consists of 10 nodes and 18 edges. The nodes are defined as the joints of upper human body and the edges are corresponding human skeletons. As the edges are directional, the number of edges doubles the number of yellow lines. For the garment graph, the nodes are defined as 32 keypoints extracted from the given garment image, presented in Fig 3 (b). We define two types of edges to describe different semantic relationships of these nodes. There are 64 edges representing the contour and 28 edges representing the symmetry structure for the garment as shown in Figure 3 (c) and (d). For all the nodes, we use horizontal and vertical coordinate values as features.

To handle the pose-conditioned garment graph regression task, we devise a two stream graph neural network with graph convolution blocks [29]. The architecture of the network is shown in Figure 4. The main stream takes the garment graph as the input for nodes feature regression. Moreover, the pose graph is embedded as the side stream to condition the regression task by hierarchically providing pose information to the main stream at different feature levels. The network is trained in a supervised manner to optimize network parameters Θ_{KP}. We use g_G, g_P, and g_T to denote the garment graph, the pose graph and the predicted graph, respectively. The objective function is represented as:

$$\arg\min_{\theta_{KP}} \mathcal{L}_{KP}(\Theta_{KP}(g_G, g_P), g_T), \tag{1}$$

where $\mathcal{L}_{KP}$ is the loss function. For a precise prediction, $\mathcal{L}_{KP}$ is defined as the combination of nodes loss $\mathcal{L}_N$ and edges loss $\mathcal{L}_E$ as presented in Eqs. (2-4).

$$\mathcal{L}_{KP}(g, g') = \lambda_N \cdot \mathcal{L}_N(g, g') + \lambda_E \cdot \mathcal{L}_E(g, g'), \tag{2}$$

$$\mathcal{L}_N(g, g') = \frac{1}{M_N} \sum_{i=1}^{M_N} ||x_i - x'_i||^2, \tag{3}$$

$$\mathcal{L}_E(g, g') = \frac{1}{M_E} \sum_{j=1}^{M_N} \sum_{i=1}^{M_N} a_{ij} \cdot (1 - \frac{\langle x_i - x_j, x'_i - x'_j \rangle}{|x_i - x_j||x'_i - x'_j|}) \tag{4}$$

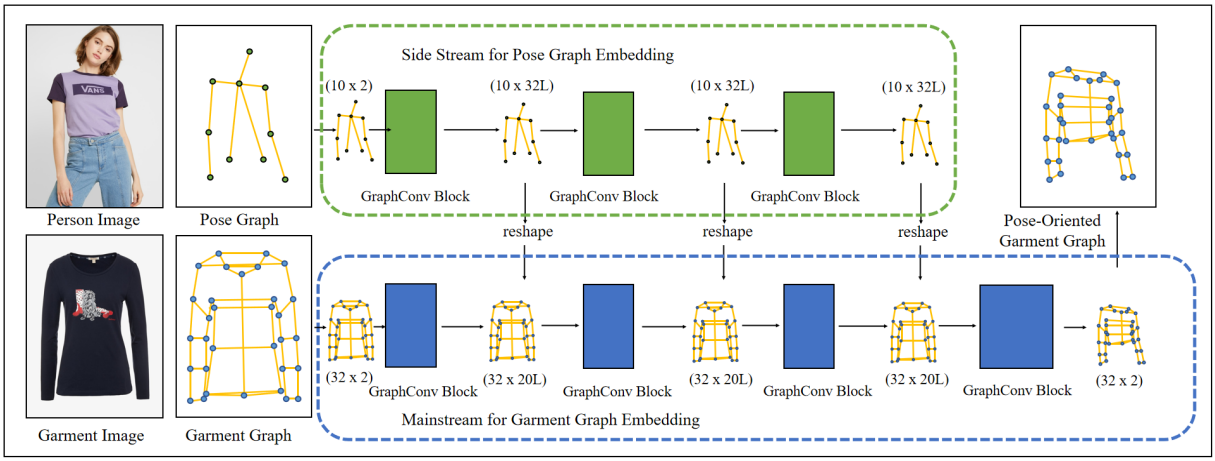

Figure 4. Illustration of the pose-oriented garment keypoints prediction module. The network consists of two graph convolution streams. The mainstream takes the garment graph of given garment image as the input and outputs the garment graph at the target pose. The side stream takes the pose graph as input and hierarchically embeds the pose information and provides conditions to the mainstream.

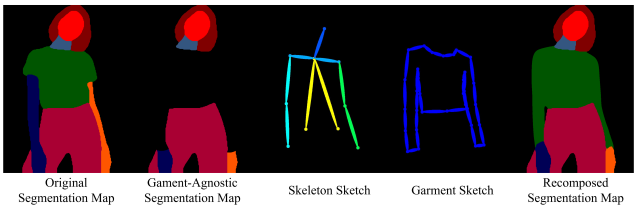

Figure 5. Illustration of the inputs for the target segmentation map generation.

where the M_N and M_E denote the numbers of nodes and edges in graphsx_i, x_j, x'_i and x'_j are the features of the ith and jth nodes in graph g and g', respectively, and a_{ij} is 1 if edge exists between nodes i and j and a_{ij} is 0 otherwise.

3.2. Segmentation Map Generation, Cloth Warping and Person Image Recomposition

With the predicted pose-oriented garment keypoints, we generate the target segmentation map and perform cloth warping. Afterwards, we recompose a person image used to be inpainted. Following [9, 15], the garment region and part of skin regions are removed from the source segmentation map (generated from the given person image) to produce a garment-agnostic segmentation map. As shown in Figure 5, we draw the sketches of human skeleton and garment contour using pose keypoints and pose-oriented garment keypoints. The sketches are stacked with the garment-agnostic segmentation map and fed into an autoencoder to generate the target segmentation map. The model is trained with a cross-entropy loss in a supervised manner.

In addition to the target segmentation map generation, pose-oriented garment keypoints are also used for fine-grained garment warping. Thin Plate Spline (TPS) [3] well preserves the pattern details during transformation hence has been widely used in virtual try-on methods [9, 22] for garment warping. Despite well handling non-rigid deformation, TPS is not effective when encountering folding and occlusions. To address the limitation of the conventional TPS, we divide the garment into five sub-segments, namely left low, left up, center, right up and right low, and then use the paired original and pose-oriented kerpoints to warp each sub-segment individually. The final warped garment is obtained by combining the five individual warped images. The comparison between the conventional TPS and our proposed scheme is clearly presented in Figure 2.

With the target segmentation map and warped garment images, we recompose a person image and treat it as the incomplete try-on image for final inpainting. Due to space limit, we put the details of the network architecture for target segmentation map generation and the implementation details of the person image recomposition in the supplementary materials.

3.3. Semantic-conditioned Inpainting

The final stage aims to inpaint the missing regions in the recomposed person image according to the target segmentation map and the existing pixels. We observe that the blending procedure of existing methods usually loses the details of the input images. To better preserve the patterns of the garment and the appearance of person, we produce a binary mask indicating whether the region is kept or not during inpainting. Inspired by [16], we develop a semantic-conditioned inpainting model based on the denoising diffusion. Different from [16], we enforce the model to inpaint the missing pixels of try-on image conditioned on the target segmentation map.

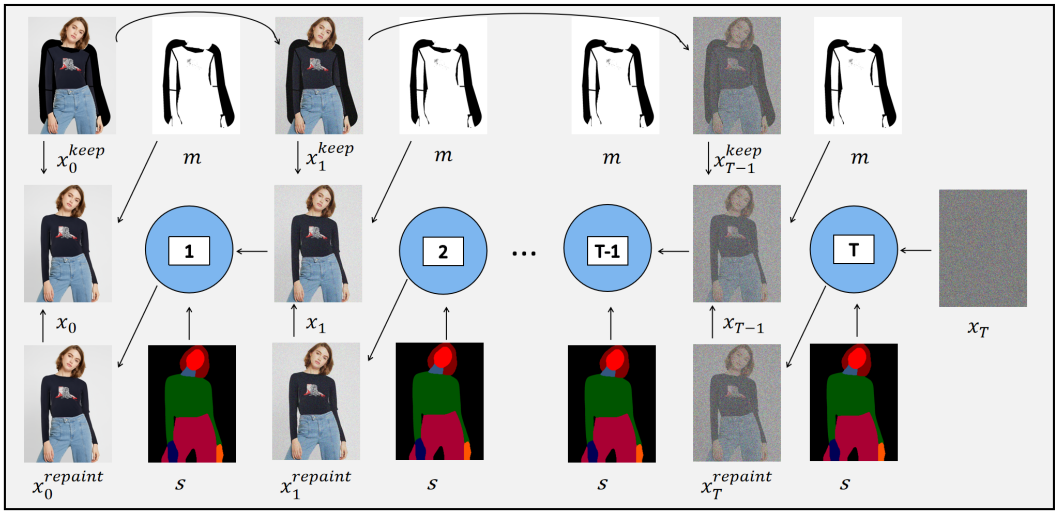

Figure 6. Illustration of the semantic conditioned inpainting with the reverse diffusion process. Given an image x_0^{keep}, the semantic conditioned inpainting start from the time step T with a noise x_T. s is the target segmentation map and m is the content keeping mask.

As defined in DDPM [12], an image x_0 can be transformed into a white Gaussian noise by progressively adding noise in T time steps. In reverse, a noise sampled from the standard Gaussian distribution can be reconstructed to an image x_0 by predicting and removing the noise step by step. In this stage, we aim to train a denoising diffusion model conditioned on the target segmentation map while predicting noise. Afterwards, we use the denoising diffusion model to perform inpainting for the final try-on image synthesis, elaborated as follows.

As shown in Figure 6, given a range of time steps $[1, T]$, the inpainting process starts from time step T. We use x_0 to denote the image to be inpainted and x_T to denote an noise sampled from the Gaussian distribution. At each time step t>1, the image x_{t-1} is the composition of x_{t-1}^{keep} and $x_{t-1}^{inpaint}$ controlled by the content keeping mask m, as represented in Eq. (5):

$$x_{t-1} = m * x_{t-1}^{keep} + (1 - m) * x_{t-1}^{inpaint}, \tag{5}$$

where x_{t-1}^{keep} is produced by adding noise to x_0 and $x_{t-1}^{inpaint}$ is produced by the denoise from x_t with the diffusion model ϵ_θ conditioned on the segmentation map s. The calculation of x_{t-1}^{keep} and $x_{t-1}^{inpaint}$ are formalized as Eqs. (6) and (7),

$$x_{t-1}^{keep} = \sqrt{\bar{\alpha}_t} x_0 + (1 - \bar{\alpha}_t)\epsilon, \tag{6}$$

$$x_{t-1}^{inpaint} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t, s)) + \sigma_t z \tag{7}$$

Following the aforementioned reverse diffusion steps, the missing pixels in the input image x_0 will be inpainted progressively and the inpainting step is conditioned by the segmentation map and known pixels in the input image.

Figure 7. Examples of generation results under the paired setting.

To get the semantic conditioned inpainting model, we train a denoising diffusion model with the original person image and its segmentation map. Following [23], the network architecture consists of spatially-adaptive normalization [18] to embed the segmentation map into the diffusion model. The overall objective function $\mathcal{L}_{SI}$ consists of two loss terms which are formalized as in Eqs. (8-10):

$$\mathcal{L}_{SI} = \mathcal{L}_{simple} + \mathcal{L}_{vlb} \tag{8}$$

$$\mathcal{L}_{simple} = E_{t,x_0,\epsilon,s}[||\epsilon - \epsilon_\theta(x_t, t, s)||^2] \tag{9}$$

$$\mathcal{L}_{vlb} = KL(p_\theta(x_{t-1}|x_t, s)||q_\theta(x_{t-1}|x_0, x_t)) \tag{10}$$

where t is a given time step sampled from $[0, T]$, s is the semantic segmentation map, x_0, x_{t-1}, x_t are images at corresponding time steps, ϵ and ϵ_{theta} are the noise and the denoising diffusion model, respectively, q and p_θ are diffusion process posterior and the distribution of estimations.

Method	256 x 192				512 x 384				1024 x 768			
	SSIM↑	LPIPS↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
CP-VTON [22]	0.739	0.159	30.11	2.034	0.791	0.141	30.25	4.012	0.786	0.158	43.28	3.762
ACGPN [25]	0.833	0.074	11.33	0.344	0.858	0.076	14.43	0.587	0.850	0.112	43.29	3.730
VITON-HD [6]	0.811	0.084	16.36	0.871	0.843	0.076	11.64	0.300	0.873	0.077	11.59	0.247
HR-VITON [15]	0.864	0.062	9.38	0.153	0.878	0.061	9.90	0.188	0.892	0.065	10.91	0.179
KGI (Ours)	**0.878**	**0.062**	**6.38**	**0.084**	**0.892**	0.064	**6.50**	**0.072**	**0.900**	0.066	**6.93**	**0.077**

Table 1. Paired Setting Results on VITON-HD Dataset.

4. Experiments

To verify the effectiveness of KGI, we conduct experiments on the recent VITON-HD dataset [6]. The dataset contains 13,679 pairs of garment and person image data. Following [15], we conducted experiments under both the paired and unpaired experimental settings for fair comparison with prior methods. For unpaired experimental setting, the quantitative results are evaluated with Frechet Inception Distance [11] and and Kernel Inception Distance (KID), which are commonly used to evaluate the performance of generative models by comparing the distributions of generated images and ground truths. For paired experimental setting, in addition to FID and KID, we also compute the Structural Similarity (SSIM) [24] and Learned Perceptual Image Patch Similarity (LPIPS) [27] for performance evaluation. Moreover, we did some ablation study to analyze the necessity of each components of KGI. In addition to numerical comparison, we also visualize the try-on results and perform qualitative comparison with [6, 15]. Due to the page limit, the implementation details about the network architecture, the hyper parameters for model training, the user study results, and some ablation study experiments are provided in the supplementary materials.

4.1. Paired Setting Results

In the paired virtual try-on experimental setting, the garment region of the person image are replaced with its paired garment. The generated try-on images are expected to be similar to the original person images. Following [15], we conduct experiments at three image resolutions: 1024x768, 512x384, and 256x192, respectively. We compared the quantitative results of our method with four image-based virtual try-on methods.

The quantitative results are presented in Tabel 1. From the numerical results, we observe that KGI consistently performs the best in terms of SSIM, FID and KID evaluation metrics at different image resolutions. Compared with prior methods, KGI has higher SSIM scores. In terms of the FID and KID performance, the superiority of KGI are more obvious. Following [15], we report KID value multiplied by a scale factor 100. For LPIPS, our method clearly performs better than CP-VTON [22], ACPGN [25] and VITON-HD [6] methods at all image resolution settings.

Method	FID↓	KID↓
VITON-HD [6]	11.65	0.256
HR-VITON [15]	11.03	0.228
KGI (Ours)	**10.33**	**0.174**

Table 2. Unpaired Setting Results on VITON-HD Dataset.

In addition to the quantitative comparison, we further visualize the generated fitting results of HR-VITON [15], VITON-HD [6] and our method for qualitatively comparison. The image examples are shown in Figure 7. From the images presented in the first row, we can observe that the synthetic results of HR-VITON and VITION-HD method present obvious color distortion and lose pattern details compared to the ground-truth. In contrast, the try-on image generated by KGI well preserves the color and detailed pattern information, which looks more realistic and more similar to the ground-truth image. The rationale is that KGI uses the paired original and pose-oriented garment keypoints to warp the garment image, which well preserves the color and patterns of the garment images. Moreover, after the image recomposition, a content keeping mask is used during the final inpainting. From the examples in the second row, we can see that the results of baselines have severe semantic errors. Especially for the image generated by HR-VITON, the length of the hem is not consistent with the actual garment and the boundary between the garment and the waist is blurring. Our method effectively reduces these errors with a more accurate target segmentation map estimation guided by the pose-oriented garment keypoints.

4.2. Unpaired Setting Results

Considering that the application scenario of the virtual try-on technique is to fit on arbitrary garments with the given person image. We further carry out experimental verification under the unpaired setting. Following [15], the experiments are conducted at 1024x768 image resolution. For quantitative comparison with prior methods, we generate the try-on images using the same person-garment pairs as prior works [6, 15] at the 1024x768 image resolution. Since there is no ground-truth under the unpaired setting, SSIM and LPIPS is not applicable. Following [15], we

Figure 8. Illustration of the generation results of different person and garment images. Images in the first row and the first column are given person images and garment images respectively. The rest of images are generation results with our proposed method.

evaluate the performance with FID and KID metrics only. As shown in Table. 2, KGI outperforms the two baselines in terms of both evaluation metrics. Compared to the recent HR-VITON method, KGI reduces the FID and KID by 0.7 and 0.54, respectively. In addition to quantitative results, we visualize the unpaired setting which is similar to the practical application scenario. Figure 8 presents the try-on results of KGI with arbitrary garment and person pairs. We can see that our method is able to produce high fidelity try-on images with arbitrary garments and people of different body shapes and poses. The try-on images generated by KGI well preserve the the shape, color, and textures of the garment image without obvious artifacts and semantic errors. We present the generation results of HR-VITON, VITON-HD and our method in Figure 9. From the figure we find that the images generated by HR-VITON and VITON-HD show more obvious color distortion and loss of details. In addition, images generated with VITON-HD appear body shape distortion. The images generated by KGI demonstrate higher fidelity.

4.3. Ablation Study

In KGI, inpainting is conditioned on the target segmentation map and the recomposed person image with warped garments. To verify the necessity of the target segmentation map and warped garments, we carry out experiments for ablation study. From the experimental results shown in Table 3, we find that the performance degrades significantly without the two modules. Figure 10 shows examples of generation results. We can intuitively observe that without the target segmentation map, despite the inpainted pixels are similar to the garments texture, the generated image is semantic incorrect. Without the warped garment, the inpainted pixels in the generated image are with arbitrary textures.

Method	SSIM↑	LPIPS↓	FID↓	KID↓
full method	0.892	0.064	6.50	0.072
w/o segmentation	0.862	0.105	8.22	0.145
w/o warped cloth	0.861	0.187	14.34	0.464

Table 3. Contribution of Different Components as Conditions

No. Step	SSIM↑	LPIPS↓	FID↓	KID↓
5	0.879	0.083	8.81	0.236
10	0.889	0.066	6.98	0.094
20	0.891	0.065	6.62	0.078
50	0.892	0.064	6.50	0.072

Table 4. The Impact of Number of Diffusion Step During Inference

Figure 9. Examples of generation results of HR-VITON, VITON-HD and our KPI methods under the unpaired setting.

Figure 10. Results generated without each conditions.

Figure 11. Results with different lengths of diffusion process.

The semantic-conditioned inpainting is based on a diffusion process with T time steps. By increasing T, the inpainting process costs more time and computational expenses. Thus, we carry out experiments to verify the impact of T. As shown in Table 4, the performance consistently improves in terms of all evaluation metrics with the increase of T. As shown in Figure 11, the results of qualitative visualization are consistent with the quantitative analysis, the pattern of the inpainted regions are more realistic and harmonious with longer diffusion process.

5. Conclusions

In this paper, we propose a pose-garment keypoints guided inpainting (KGI) method for image-based virtual try-on task, which produces high fidelity try-on images and well preserves the patterns and shapes of the garments. In the proposed method, pose keypoints and garment keypoints are extracted from the source images and constructed as graphs to predict pose-oriented garment keypoints. After which, the predicted keypoints are used as guide information for garment warping and the target segmentation map generation. The given person image is recomposed with the warped garment image based on the semantic information of the target segmentation map. The missing region of the recomposed person image is finally filled with a semantic-conditioned inpainting scheme. To verify the effectiveness of KGI, we conduct extensive experiments on VITON-HD dataset under both paired and unpaired settings. The qualitative and quantitative results show that KGI significantly outperforms prior methods at different image resolutions.

In this work, the experiments on VITON-HD dataset contain samples of upper-body wears only. While the geometric structures of lower-body wears and foot wears can be described by landmarks as well. The KGI method could be extended for other garment items in future work.

6. Acknowledgement

This work was conducted at Bytedance AI Lab.

References

[1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 409–425. Springer, 2022.

[2] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 344–359. Springer, 2020.

[3] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[5] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13809–13818, 2021.

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021.

[7] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16928–16937, 2021.

[8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[10] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

[14] Jiyeon Kim and Sandra Forsythe. Adoption of virtual try-on technology for online apparel shopping. *Journal of interactive marketing*, 22(2):45–59, 2008.

[15] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 204–219. Springer, 2022.

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[17] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020.

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[19] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2242–2251, 2019.

[20] Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. Swapnet: Garment transfer in single view images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018.

[21] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021.

[22] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018.

[23] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.

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

[25] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020.

[26] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10511–10520, 2019.

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[28] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021.

[29] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019.