The Visual Computer
https://doi.org/10.1007/s00371-024-03603-z

ORIGINAL ARTICLE

Slot-VTON: subject-driven diffusion-based virtual try-on with slot attention

Jianglei Ye[1] · Yigang Wang[1] · Fengmao Xie[2] · Qin Wang[1] · Xiaoling Gu[3] · Zizhao Wu[1]

Accepted: 5 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Virtual try-on aims to transfer clothes from one image to another while preserving intricate wearer and clothing details. Tremendous efforts have been made to facilitate the task based on deep generative models such as GAN and diffusion models; however, the current methods have not taken into account the influence of the natural environment (background and unrelated impurities) on clothing image, leading to issues such as loss of detail, intricate textures, shadows, and folds. In this paper, we introduce Slot-VTON, a slot attention-based inpainting approach for seamless image generation in a subject-driven way. Specifically, we adopt an attention mechanism, termed slot attention, that can unsupervisedly separate the various subjects within images. With slot attention, we distill the clothing image into a series of slot representations, where each slot represents a subject. Guided by the extracted clothing slot, our method is capable of eliminating the interference of other unnecessary factors, thereby better preserving the complex details of the clothing. To further enhance the seamless generation of the diffusion model, we design a fusion adapter that integrates multiple conditions, including the slot and other added clothing conditions. In addition, a non-garment inpainting module is used to further fix visible seams and preserve non-clothing area details (hands, neck, etc.). Multiple experiments on VITON-HD datasets validate the efficacy of our methods, showcasing state-of-the-art generation performances. Our implementation is available at: https://github.com/SilverLakee/Slot-VTON.

Keywords Virtual try-on · Diffusion models · Generative models · Slot attention · High-resolution image synthesis

1 Introduction

Virtual try-on stands as one of the state-of-the-art fashion technological applications, enabling users to virtually try on diverse clothes in both online and offline stores. This innovative approach eliminates the necessity for individuals to physically don garments, significantly enhancing the shopping experience and optimizing operations within the fashion and clothing industry. The main goal of VTON is to seamlessly transfer clothing from one image to another, with a specific focus on minimizing the loss of intricate clothing details. The task requires overcoming two major challenges [43]: (1) achieving a seamless fit of clothing in relevant areas and ensuring accurate reproduction of realistic clothing details, including shadows and wrinkles. (2) ensuring consistency in generating non-clothing elements such as human identity and pose. Many virtual try-on endeavors [8, 12, 13, 45, 48, 50] have relied on Generative Adversarial Networks (GANs) [9]. However, despite advancements in generating relatively lifelike human images, GAN-based generators suffer from inherent weaknesses, particularly in

✉ Yigang Wang
yigang.wang@hdu.edu.cn

Jianglei Ye
jiangleiye@hdu.edu.cn

Fengmao Xie
21041429@hdu.edu.cn

Qin Wang
wangqin@hdu.edu.cn

Xiaoling Gu
guxl@hdu.edu.cn

Zizhao Wu
wuzizhao@hdu.edu.cn

[1] School of Digital Media Technology, Hangzhou Dianzi University, Hangzhou, China

[2] School of Electronics and Information Engineering, Hangzhou Dianzi University, Hangzhou, China

[3] School of Computer Science, Hangzhou Dianzi University, Hangzhou, China

Published online: 23 August 2024

generated quality and model collapse issues. Therefore, when generating details, there may be errors in the final generated results.

Recently, diffusion models [16, 42] have achieved significant progress in image generation. Within the realm of virtual try-on, methodologies [5, 10, 54] grounded in diffusion models are increasingly outperforming those rooted in GANs. For instance, Kim et al. [21] introduces a novel method for virtual try-on, leveraging semantic correspondence learning within the latent space of pre-trained diffusion models to achieve stable and accurate results. Additionally, Morelli et al. [31] achieve the preservation of clothing texture and details through a text-inversion [7] approach, effectively maintaining fidelity.

While these approaches have significantly enhanced clothing detail preservation, they have overlooked challenges in real-world scenarios. In actual scenes, a clothing image may encompass various subjects beyond the clothing itself, such as environmental backgrounds and unrelated substances. These interference factors may lead to the blurring of the clothing subject, hindering the garment's ability to address complex scenarios like wrinkles and shadows due to human pose. Furthermore, existing works have neglected the preservation of elements beyond clothing intricacies (e.g., hands, human identity), and addressing seam issues between the wearer and backgrounds is crucial.

Motivated by these challenges, we present a subject-driven virtual try-on framework with diffusion models. This framework comprises two phases: data processing and generation. In the data processing phase, we employ a warping module to warp the clothing, which is subsequently fused with the clothes-agnostic person during the generation phase. Then, we introduced slot attention, which extracts a series of slot representations to achieve unsupervised separation of subjects in an image. Through this mechanism, multiple subjects within clothing images are segregated into distinct slots (e.g., clothing slots, background slots), with only clothing slots retained based on task specifications, thereby enhancing the clothing subject. To facilitate seamless integration, we further devise a fusion adapter capable of fusing multiple conditions, ensuring meticulous preservation of details throughout the framework. In the generation phase, we leverage the fusion conditions acquired as a condition to guide the diffusion model's generation process, thereby generating person images replete with intricate clothing details. Additionally, we also propose the non-garment inpainting module, which involves specific fine-tuning to the decoder for each image. This refinement ensures the preservation of intricate details in non-clothing areas of the generated results, effectively mitigating seams between the wearer and the background.

We conduct comprehensive validation on the VITON-HD dataset. The results demonstrated both superior quantitative and qualitative performance compared to state-of-the-art methods.

Our main contributions are summarized as follows:

- We propose Slot-VTON, a virtual try-on pipeline that extracts the clothing subject through slot attention as a guiding condition and utilizes a novel diffusion model to achieve a high-fidelity virtual try-on.
- We design a fusion adapter integrating multiple conditions, which ensures the conditions can collaboratively guide the diffusion model.
- We introduce the non-garment inpainting module to solve the seam problem, which fine-tunes the decoder for each image within the established framework.
- We conducted comprehensive experiments to validate the efficacy of each component in our architecture. The results show that the person images generated by our method achieve state-of-the-art performance during the virtual try-on task.

2 Related work

In this section, we present a brief overview of image-based virtual try-on, latent diffusion models, as well as slot attention.

2.1 Image-based virtual try-on

Image-based virtual try-on is a technology that fits new in-shop clothes into a human image while preserving details. Pioneering works like VITON [13] have adopted a coarse-to-fine framework. Initially, it utilizes an encoder–decoder structure to generate a preliminary synthetic image followed by a refinement network that enhances the coarse representation using the warped clothing item obtained through a thin-plate spline (TPS) transformation [6]. Subsequently, some works focused on improving the warping module. Wang et al. [45] introduced CP-VTON, which incorporates a convolutional geometric matcher that learns geometric deformations (i.e., thin-plate spline transform). However, methods that use TPS for deformation processing, can only provide simple deformation handling. They transfer the clothing to the target area roughly and cannot handle more significant geometric deformations. In response to these issues, some works [8, 12, 14] utilize a flow-based method to simulate the flow field of clothing on specific body areas, ensuring a better fit.

Many efforts [3, 17, 28] have been paid to the image synthesis [11] stage. Previous works [13, 45] relied on human segmentation, and training try-on models required high-quality human parsing. Any minor errors in the segmentation would lead to highly unrealistic try-on images. To overcome

this, Issenhuth et al. [18] propose a novel student-teacher paradigm in which human segmentation is not employed. The Parser-Free model's image generation quality is further enhanced by PF-AFN [8], which employs a *teacher-tutor-student* knowledge distillation approach alongside real image supervision.

Most of the current virtual try-on methods heavily rely on GANs, but a significant issue is that the images generated by GANs struggle to preserve clothing details. Even with an improved warping module for clothing, when combined with the human image, it still lacks a high level of realism. With the emergence of diffusion models showcasing powerful generative capabilities, recent works are increasingly turning to diffusion models to achieve superior generation results.

2.2 Latent diffusion models

Inspired by principles from non-equilibrium statistical physics, diffusion models were introduced as a novel generative model for data distribution [42]. These models employ an iterative Markov process, initiating a forward diffusion process to disrupt the data structure, followed by a learned reverse diffusion process to restore the original data. Currently, the latent diffusion model (LDM) [35] introduces cross-attention as a general approach to conditional guided model generation. Building on this, recent advancements in text-to-image [24, 33, 37, 39] and image-to-image [32, 38, 44] synthesis showcase the prowess of latent diffusion models in generating high-quality visual images. Yang et al. [51] introduced the concept of inpaint for sample-guided image editing, which uses a self-supervised manner to offer fine-grained image control.

As latent diffusion models demonstrate their impressive generative capabilities, researchers are exploring their potential application in the virtual try-on task. Gou et al. [10] have extended the concept of inpaint to virtual try-on, and use additional warping modules to restore high-frequency details. While Morelli et al. [31] have effectively preserved garment texture and details through a text-inversion approach, which trains the added skip connection module to achieve higher fidelity generation.

In this paper, we leverage the strengths of latent diffusion models to propose a novel diffusion-based approach. Departing from traditional methods, we introduce a subject-driven approach, utilizing the slot attention mechanism to extract the clothing subject and direct the model's generation process effectively.

2.3 Slot attention

The slot attention mechanism was initially introduced by Locatello et al. [29] to decompose an image into a series of slot representations, with each slot corresponding to a specific subject within the image. SAVi [23], based on slot attention, runs on videos with a recurrent encoder–decoder structure. However, its mixture-based decoder lacks interaction between slots, limiting modeling capacity. To address this limitation, STEVE [41] proposes reconstructing intermediate features from another network [40]. Although it works on videos with textured objects and backgrounds with the help of cross-attention and feature-level [49] reconstruction, transformer-based [26] decoder performance remains low for complex data. Building upon the success of the diffusion model, Wu et al. [47] proposed SlotDiffusion, which further improved the quality of image generation by combining slot and diffusion techniques.

Taking advantage of these advancements, we abstract the conditions of clothing images into a set of slots, including clothing slots and background slots. Specifically, we focus solely on extracting the clothing slots as a condition, to mitigate the negative impact of irrelevant factors on the representation of the clothing. This subject-driven approach enhances the robustness of our model, enabling it to better handle challenging scenarios (Fig. 1).

3 Method

Given a clothing image and a person image, the objective is to realistically transfer the cloth onto the target person. In this work, we undertake the virtual try-on task in the form of inpainting grounded on stable diffusion models [35]. Despite the recent remarkable success of text-based image editing [20, 53], enabling users to submit images of clothes is a more practical and feasible approach to enhance the granularity of virtual try-on functions.

In the following, we first introduce the diffusion priors for virtual try-on in Sect. 3.1. Then, we present an overview of our method in 3.2, followed by detailed explanations in 3.3. Finally, we elaborate the loss functions in Sect. 3.4.

3.1 Diffusion prior

Diffusion-based models employ a two-step process to manipulate data distributions. Initially, they add noise via a forward process to destroy the structure of the data distribution (usually converted to Gaussian noise in practice). Then, they learn step by step to restore the original data distribution through a reverse process. In Slot-VTON, the state-of-the-art stable diffusion framework (SD) [25] is utilized as priors. This framework integrates a variational autoencoder [22] (consists of an encoder E and a decoder D) and a denoiser U-Net [36]. It follows a process where the image first being compressed into latent space using encoder E. Subsequently, guided by a text prompt C_y encoded using a CLIP text encoder, the

Fig. 1 Images generated by the proposed Slot-VTON model. Given an input person image and a try-on cloth, our method can generate high-fidelity virtual try-on results

U-Net denoises the Gaussian noise ϵ. This training process is achieved by minimizing a loss function $\mathcal{L}_{\text{LDM}}$, which involves the difference between the noise and its reconstruction at each time step of the forward diffusion process:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\epsilon(\mathbf{x}),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(Z_t, t, C_y)\|_2^2\right], \tag{1}$$

where $t \in 1, ..., T$ represents the time step, Z_t denotes the noise latent at time t and C_y denotes the text CLIP embedding. The text condition is incorporated into the cross-attention mechanism of U-Net to guide generation.

3.2 Framework overview

Our proposed Slot-VTON employs a pre-trained large-scale diffusion model to generate images $\hat{X}$, which is a person X_p wearing a target cloth X_c. We utilize a binary mask $m \in \{0, 1\}^{H\times W}$ to identify the area of the person image requiring inpainting, typically the upper body in virtual try-on scenarios. The upper body region $\hat{X} \odot m$ aligns with the clothing image X_c, while the complementary area aligns with the person image X_p, i.e., $X_p \odot (1 - m) \approx \hat{X} \odot (1 - m)$, where $\odot$ denotes element-wise multiplication. Additionally, we introduce a special weight fine-tuning mechanism in the non-garment inpainting module, to ensure a seamless transition at the interface between the two regions.

Figure 2 illustrates our method, where the yellow and purple segments denote the conditional data processing and generation phases, respectively. In the conditional data processing phase, we obtain clothes-agnostic person X_a, the clothes-agnostic segmentation map S_p and densepose P from the person image X_p. Similar to previous works [8, 10], the clothing image X_c is combined with the clothes-agnostic segmentation map S_p and densepose P through the warping module to predict the appearance flow field and distort the clothing to obtain warped clothing X_w. Additionally, we obtain the clothing slot C_s and other clothing feature embeddings through different processing. In the generation phase, the coarse result X_{aw}, synthesized from the clothes-agnostic person X_a and warped clothing X_w, serves as the input to the diffusion model. Multiple clothing conditions obtained are collectively input into the diffusion model as the condition guide, yielding the final result $\hat{X}$.

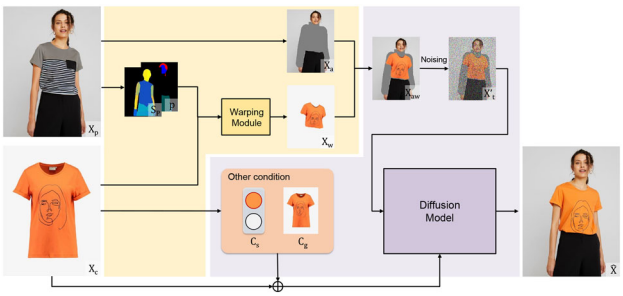

Fig. 2 The overview of our method. First, we obtain clothes-agnostic person X_a,the segmentation result S_p, densepose P from the person image X_p. the clothing image X_c was combined with the S_p and P through the warping module to obtain the warping cloth X_w. The coarse result X_{aw} synthesized by X_a and X_w is used as the input of the diffusion model. Then, the clothing item X_c is processed to obtain the clothing slot C_s. Integrating other added clothing signals, these conditions are fused as the control conditions of the diffusion model to collaboratively guide the generation of the diffusion model

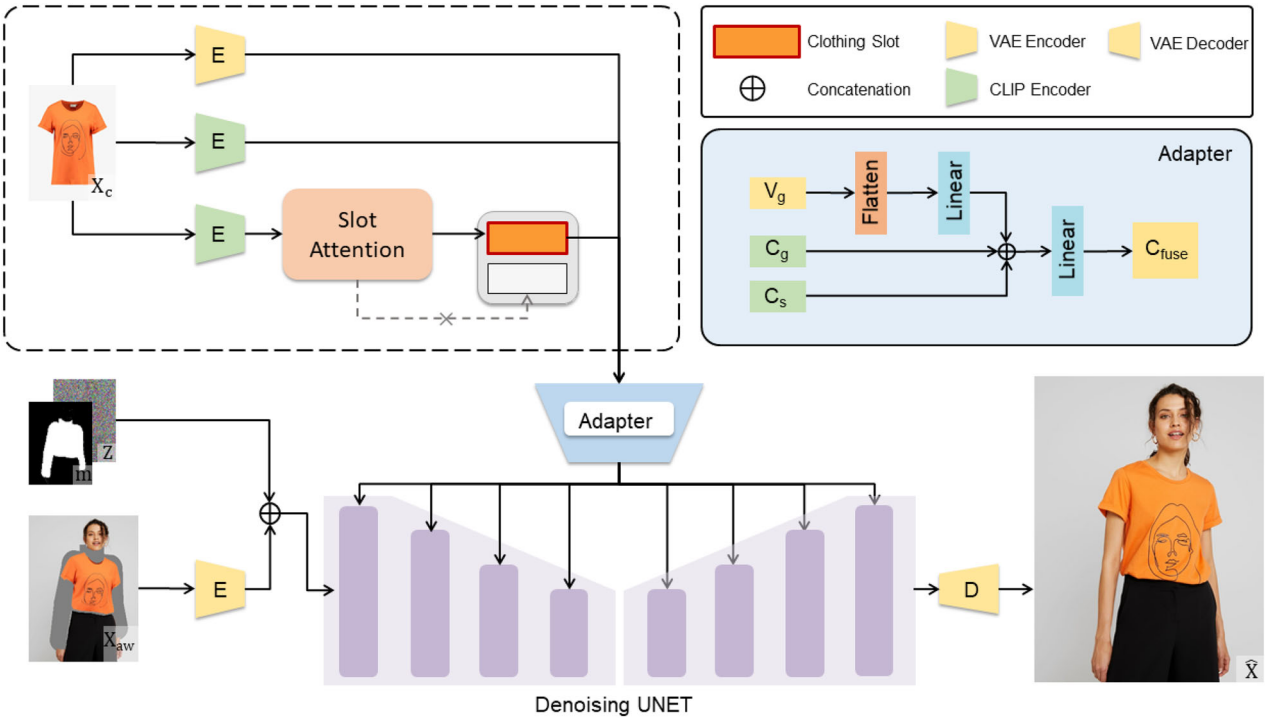

Fig. 3 The training pipeline of our proposed Slot-VTON. Clothing image X_c extracts clothing slot through slot attention, thereby eliminating the interference of other subjects to the clothing itself and enhancing the subject of the clothing item. The fusion condition C_{fuse} of the final input diffusion model is obtained through the fusion adapter

While completing the virtual try-on task through inpainting can yield higher-quality outcomes, it may lead to certain unsatisfactory scenarios. Particularly, the absence of processing in areas beyond the garment could result in noticeable seams between the background and the person. We tackle this issue through the non-garment inpainting module, which fine-tunes the decoder for each image.

3.3 Subject-driven diffusion model

Figure 3 illustrates the training pipeline of our diffusion model. The input of the diffusion model consists of the clothes-agnostic person X_a, the clothing mask m, and the noise image z. To preserve certain low-level clothing features such as outline and color, we integrate the warped clothing X_w with the agnostic person X_a resulting in the formation of the input coarse result X_{aw}. The clothing warping process follows the DCI-VTON [10] module, which utilizes an iterative refinement strategy to achieve the final appearance flow. This approach enables the capture of the intricate correspondence between clothes and person images, enhancing the efficiency in handling significant misalignments. Ultimately, the input undergoes reconstruction by the diffusion model guided by the adapter fusion module to produce the final result.

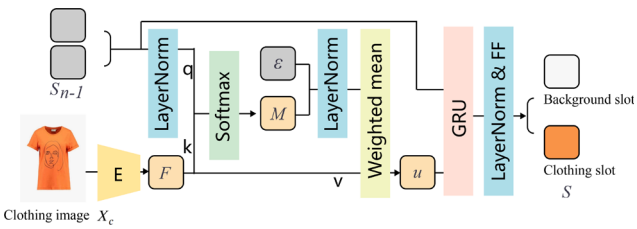

Fig. 4 Architecture of our slot attention mechanism.Through slot attention mechanism, we extract the clothing subject and discard the background subject, to maximize the preservation of clothing details

In a real scene, for a clothing image X_c, besides the clothing subject, there may be unwanted subjects such as the background or other objects, collectively referred to as interference subjects. To eliminate the adverse effects of these interference subjects on the clothing subject, we introduce the slot attention mechanism to assign each subject to a slot, as illustrated in Fig. 4. To facilitate training, we set only two subjects for one clothing image in our training set: the clothing itself and the background. We retain only the clothing slot and discard the others, enhancing the clothing subject.

Initially, stable diffusion (SD) [35] used contrastive language-image pre-training (CLIP) [34] embeddings with textual prompts as conditions to guide the generation of the diffusion model. Nevertheless, as previously mentioned, tex-

tual conditions alone cannot accurately describe the texture information of clothes. CLIP encodes both text and images into a shared embedding space, making it seem natural to directly replace text embeddings with CLIP image embeddings. However, in practice, it was discovered that the indirect visual information captured by CLIP was inadequate for capturing the intricate details in conditional images. Specifically, the generated clothing images often lacked realistic lighting and folds, appearing overly smooth. Therefore, we enhanced the original CLIP embeddings by incorporating clothing embedding encoded by SD's VAE module. This enhancement enabled the recovery of intricate visual details, such as lines and contour structures.

Slot Attention. In slot attention, the initial step involves encoding the input clothing image X_c into the input feature F using the CLIP image encoder. Subsequently, soft feature space clustering is conducted on F to derive N concept partitions. This process entails iterative cross-attention, with slots functioning as queries q and input features as keys k and values v. specifically, at the onset of the iterative refinement phase, slot S is initialized with random Gaussian noise. Then in the course of T iterations, we calculate the attention scores of both q and k and obtain the attention weight matrix M by performing softmax operation on them. Next, the M is weighted to get the updated slot value u. Finally, the updated value u and the original slot value s_{n-1} are input into the GRU model to realize the final slot value update. The slot in the last iteration is considered to be the final slot representation S.

During our training, the clothing image will be divided into two slots, namely the background slot and the clothing slot, and we will use the clothing slot as the final condition C_s.

Fusion Adapter. To fully leverage the guiding roles of all the mentioned embeddings in image generation, we designed a custom fusion adapter $\mathcal{A}$. This fusion adapter ensures that the conditions are combined to preserve as much detail of the clothing as possible during the generation process. By blending all these embeddings through $\mathcal{A}$, we generate a fused embedding for typical cross-attention operations in the network. As depicted in the top right corner of Fig. 3, this module amalgamates all conditions and reshapes the output to fit the expected format of a denoising U-Net cross-attention module. The final fusion embedding is defined as:

$$\mathcal{C}_{\text{fuse}} = \mathcal{A}(\text{Linear}(\text{Flatten}(V_g)), C_g, C_s), \tag{2}$$

where V_g represents the VAE embedding of the clothing image, C_g denotes the CLIP embedding of the clothing image, and C_s signifies the CLIP embedding of the clothing slot.

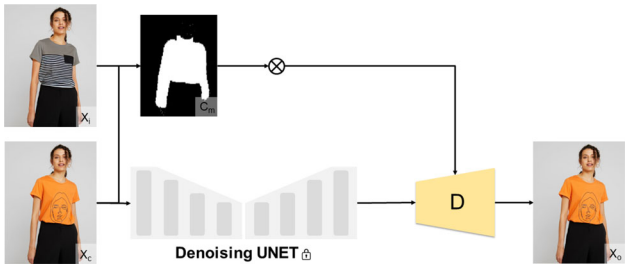

Fig. 5 Architecture of non-garment inpainting module. In this module, we solely fine-tune the decoder for each image, while keeping the other weight parameters of the model unchanged

Non-garment Inpainting Module. The *latent to image* method, which generates images within a latent space, significantly improves the efficiency of diffusion models in the generation process. However, this improvement comes at the cost of a certain degree of loss in the generated images. To address this issue and preserve intricate details within the person image, we incorporate the concept of inpainting. Specifically, our approach focuses on processing only the clothing region while leaving the areas outside this region untouched. Although this pixel-level stitching method retains most details, it introduces another problem known as the *seam problem* where the junction of two regions may not align well. Additionally, there may be a loss of details in interconnected regions such as hands and necks. To overcome these challenges, we develop a specialized non-garment inpainting module.

Figure 5 illustrates the structural details of this module. Thanks to the conclusion of blended latent diffusion [1] that refining the generator weights for each image leads to superior reconstruction results, we extend it to the virtual try-on task. By utilizing a precise clothing mask C_m, we perform fine-tuning of decoder weights for each coarsely generated image, enabling seamless cloning. Importantly, the model's other weight parameters remain unchanged throughout this procedure. The formula for this process is elucidated as follows:

$$\begin{aligned} w^* = \underset{w}{\text{argmin}} \, & \| D_w(z_0) \odot C_m - x_c \odot C_m \| \\ & + \mu \| D_w(z_0) \odot (1 - C_m) - x_i \odot (1 - C_m) \| \end{aligned}. \tag{3}$$

The coarse result after the clothing area treatment is termed as x_c, with the original image labeled as x_i. The hyperparameter μ, indicating the importance of background reconstruction, is fixed at a value of 10 in this research.

3.4 Loss functions

Similar to $\mathcal{L}_{\text{LDM}}$, the optimization function for the U-Net in Slot-VTON is shown below:

Fig. 6 Image reconstruction results from the Slot-VTON with and without the VGG loss. The VGG loss is essential for maintaining the structural details of clothing accurately

$$\mathcal{L}_{\text{Slot}} = \mathbb{E}_{\epsilon(\mathbf{x}),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(Z_t, Z_{aw}, m, t, C_{\text{fuse}})\|_2^2\right], \tag{4}$$

where $t \in 1, ..., T$ represents the time step, m is the mask identifying the area of the person image requiring inpainting, Z_t denotes the noise latent at time t, Z_{aw} is the coarse generated result of the person image, and C_{fuse} is the embedding of the fusion condition. In addition to the L_{Slot}, we incorporate a VGG [19] perceptual loss function. Specifically, our method is optimized by utilizing the total loss function outlined below:

$$\mathcal{L} = \mathcal{L}_{\text{Slot}} + \lambda\mathcal{L}_{\text{VGG}}. \tag{5}$$

In our experiment, the hyperparameter λ is empirically set to 1e-4. Our findings revealed the significant impact of the VGG perception loss on reconstructing intricate clothing contour details (Figs. 6, 7).

4 Experiments

4.1 Experiments setting

Datasets. Our experiments involved the VITON-HD dataset [4], conducted at a resolution of 512×384 and 256×192. This dataset comprises 13,679 pairs of front-view women and tops, with 11,647 pairs allocated to the training set and 2,032 pairs to the test set. The dataset is designed with both paired and unpaired settings. In the paired setting, the clothing matches the model, while in the unpaired setting, the clothing differs from the model.

Baselines and Evaluation Metrics. Our comparative study involved Slot-VTON and several state-of-the-art methods, namely CP-VTON [45], VITON-HD [4], HR-VITON [25], GP-VTON [48], and DCI-VTON [10]. These methods were trained from scratch on VITON-HD [4], utilizing the official code provided by the respective authors.

Quantitative evaluation of virtual try-on is a challenging task due to the lack of a real reference person with the target clothing. For both test setups, we used different metrics to evaluate the performance of our method. For paired Settings, we use the learned perceptual image patch similarity (LPIPS) [52] and structural similarity (SSIM) [46] to evaluate the consistency of the generated image with the ground-truth. We also used the Frechet Inception Distance (FID_p) [15] and the Kernel Inception Distance (KID_p) [2] to supplement the assessment completeness. For unpaired Settings, we use FID_u and KID_u to measure the authenticity of the generated image. In addition, we conducted two user studies on the authenticity of the generated image and the consistency with the input image.

Implementation Details. Our experiments focused on high-resolution image training and inference, using a single Nvidia A40 GPU. We employed the same appearance flow strategy as DCI-VTON [10] for the warping networks, maintaining consistent hyperparameters. This approach allows for more flexible transformations and effective handling of significant misalignments. Our baseline model was stable diffusion, and we utilized the CLIP pre-trained model as the primary condition image encoder. During training, we set the iteration times of slot attention to 10 and the number of slots to 2 (background slots and clothing slots, respectively). The model was trained using the AdamW optimizer [30] with a learning rate of 1e-5, a batch size of 2, and a total of 40 epochs. For the inference process, we employed the PLMS [27] sampling method with 100 sampling steps.

4.2 Quantitative evaluation

We compared our method with previous virtual try-on methods: CP-VTON [45], VITON-HD [4], HR-VITON [25], GP-VTON [48], and DCI-VTON [10]. Table 1 shows a quantitative comparison with these methods. It can be observed that for unpaired Settings, our approach achieves performance results comparable to the current SOTA approach, DCI-VTON [10]. For paired settings, Slot-VTON also outperforms its competitors in terms of input consistency (i.e., LPIPS and SSIM).

User Study. To better evaluate the quality of our model generation, we conducted two user studies to measure the authenticity of the generated model and the consistency of its given inputs. Specifically, we collected 100 randomly selected pairs of composite images generated by different

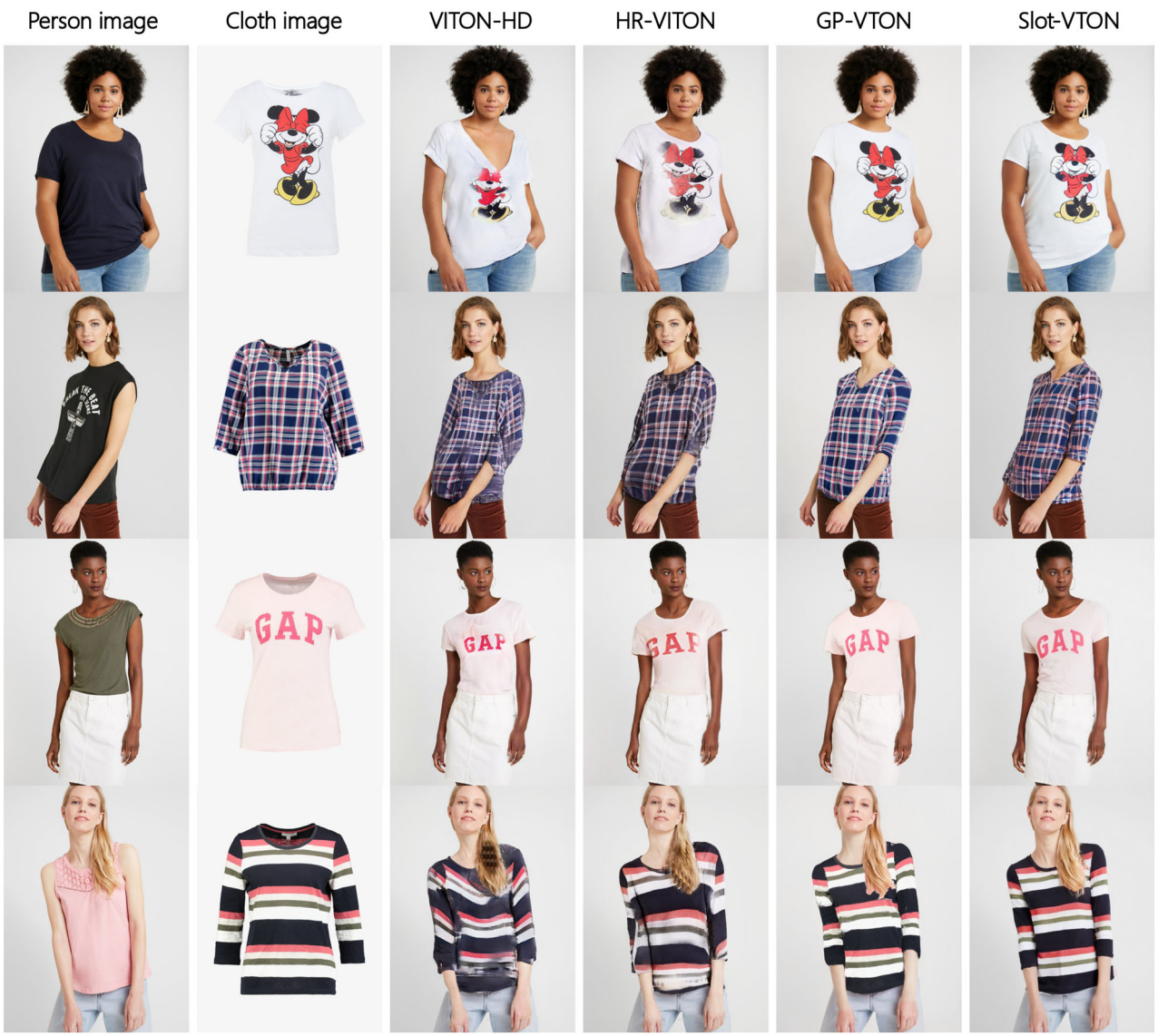

Fig. 7 Qualitative results generated by Slot-VTON and competitors on VITON-HD dataset. Our method outperforms others in preserving the fine details of the garment

Table 1 Quantitative comparison with baselines

Method	256 × 192				512 × 384					
	LPIPS ↓	SSIM ↑	FID_u ↓	KID_u ↓	LPIPS ↓	SSIM ↑	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓
CP-VTON* [45]	0.159	0.739	30.11	2.034	0.141	0.791	30.25	4.01	–	–
VITON-HD [4]	0.088	0.803	16.36	0.871	0.101	0.852	12.17	0.32	10.21	0.28
HR-VITON [25]	0.066	0.862	9.57	0.270	0.384	0.609	11.09	0.35	11.34	0.38
GP-VTON [48]	0.087	0.833	9.06	0.107	0.066	**0.895**	6.66	0.12	9.61	0.15
DCI-VTON [10]	**0.049**	0.906	8.52	0.077	0.057	0.887	6.24	**0.06**	9.02	0.12
Slot-VTON	0.052	**0.912**	**8.13**	**0.067**	**0.054**	**0.895**	**6.07**	**0.06**	**8.95**	**0.09**

The bold indicates the best performance
The * marker indicates results reported in previous works, which may differ in terms of metric implementation

Table 2 User study results on the unpaired test set of VITON-HD dataset

Model	Realism	Coherence
VITON-HD[4]	0.39%	0.52%
HR-VITON[25]	4.30%	1.13%
GP-VTON[48]	14.73%	17.21%
DCI-VTON[10]	17.57%	13.83%
Slot-VTON	**63.01%**	**67.31%**

The bold indicates the best performance

methods from the test set, with a resolution of 512×384. 20 participants were then invited to select what they thought was the best generation method for each pair of composite images, and we reported how often each method was selected as the best method for both aspects. As shown in Table 2, more than 63% of users believe that the images produced by our model are of higher quality.

4.3 Ablation study

By using 512×384 resolution as the basic setting on the VITON-HD dataset, we conducted ablation studies on each part of the network to verify its effectiveness. As shown in Table 3, the experiments in the first two lines mainly focused on the upper body repair area of the model, the slot conditions of the garment were removed in the first line, and the added clothing signal encoded by the variational autoencoder (VAE) was removed in the second line. Subsequently, the third line of the experiment mainly focuses on the parts except clothing and deletes the non-garment inpainting module.

Effectiveness of Slot and other condition. Figure 8 illustrates the visualization results of two proposed conditions for the clothing area of the model. The slot condition notably enhances the structural features of the generated clothing, improving the clarity and completeness of the clothing lines. This condition significantly contributes to the final image generation outcome. Additionally, the added VAE condition effectively complements the original results, mitigating the loss of certain details.

Effectiveness of non-garment module. Figure 9 illustrates the visualized results pre- and post-removal of the non-garment inpainting module. Upon removal, the model-produced image exhibits noticeable seam issues, alongside a partial loss of feature details at regional connections, such as the model's hand. The inclusion of this module effectively addresses these challenges.

Effectiveness of VGG loss. Figure 6 illustrates that solely employing the L_{Slot} loss function during training, with all other conditions held constant, remains insufficient for addressing specific extreme cases, despite the fusion condition's ability to preserve general clothing outline details. The inclusion of the VGG perceptual loss enhances our capability to handle such scenarios.

4.4 Qualitative evaluation

The composite images produced by our method and others on VITON-HD dataset are shown in Fig. 7. VITON-HD [4]

Fig. 8 The visual comparison of w/o Slot Condition, w/o VAE Condition and ours

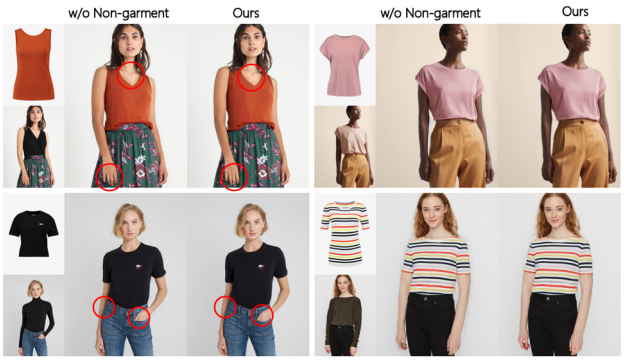

Fig. 9 Visualization results of with and without non-garment inpainting module. (Left) This module repairs certain intricate non-garment details, including the hands and neck. (Right) This module effectively solves the seam problem between background and the cloth

Table 3 Ablation studies of network components in our model

Method	LPIPS↓	SSIM↑	FID_u ↓	KID_u ↓
w/o Slot Condition	0.063	0.892	9.30	0.148
w/o VAE Condition	0.067	0.886	9.15	0.125
w/o Non-garment module	0.066	0.888	9.02	0.100
Ours	**0.054**	**0.895**	**8.95**	**0.093**

We multiply KID by 100 for better comparison
The bold indicates the best performance

correctly generated the image of the model wearing the target cloth. However, issues with the cloth's structure, like an excessive neckline in the first line, were observed. HR-VITON [25] struggles with complex textures, such as logos and patterns on clothing. GP-VTON [48] effectively preserves texture details but exhibits bluntness in the connection between the body and the garment, as seen in the fourth line. Slot-VTON maintains cloth characteristics, including shadow folds, and adeptly handles the interaction between the body and garment for high-quality image generation.

5 Conclusion

In this work, we propose a slot attention-based inpainting approach for the virtual try-on task, aimed at effectively addressing the challenge of preserving intricate clothing details. To combine the slot unsupervised subject extraction ability with the powerful generative ability of diffusion model, we propose a fusion adapter module to effectively integrate multiple conditions including slot, as a guide to the seamless generation of the diffusion model. Additionally, we develop a non-garment inpainting module to optimize the preservation of details in the non-garment area and effectively resolve seam problems. The experimental results from the VITON-HD dataset demonstrate the superior effectiveness of our approach. In future work, we aim to further investigate methods for achieving higher-quality seamless generation while ensuring generation efficiency.

Author Contributions Jianglei Ye is responsible for Conceptualization, Methodology, Software, Writing—Original Draft, Visualization; Yigang Wang is responsible for Supervision, Data Curation, Software, Visualization; Fengmao Xie is responsible for Data Curation, Software, Visualization; Qin Wang is responsible forMethodology, Resources, Software; Xiaoling Gu is responsible for Methodology, Resources, Software; Zizhao Wu isresponsible for Supervision, Conceptualization, Writing—Reviewand Editing.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Trans. Graph. (TOG) **42**, 1–11 (2023)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans (2018). arXiv preprint arXiv:1801.01401
3. Chang, Y., Peng, T., Yu, F., He, R., Hu, X., Liu, J., Zhang, Z., Jiang, M.: Vtnct: an image-based virtual try-on network by combining feature with pixel transformation. Vis. Comput. **39**, 2583–2596 (2023)
4. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14131–14140 (2021)
5. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Lazebnik, S.: Street tryon: learning in-the-wild virtual try-on from unpaired person images (2023). arXiv preprint arXiv:2311.16094
6. Duchon, J.: Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976, pp. 85–100. Springer (1977)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: personalizing text-to-image generation using textual inversion (2022). arXiv preprint arXiv:2208.01618
8. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8485–8493 (2021)
9. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014). arXiv:1406.2661
10. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 7599–7607 (2023)
11. Guo, H., Sheng, B., Li, P., Chen, C.P.: Multiview high dynamic range image synthesis using fuzzy broad learning system. IEEE Trans. Cybern. **51**, 2735–2747 (2019)
12. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: a flow-based model for clothed person generation. In: Proceedings of

the IEEE/CVF International Conference on Computer Vision, pp. 10471–10480 (2019)
13. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: an image-based virtual try-on network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7543–7552 (2018)
14. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3470–3479 (2022)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. **33**, 6840–6851 (2020)
17. Hu, X., Zhang, J., Huang, J., Liang, J., Yu, F., Peng, T.: Virtual try-on based on attention u-net. Vis. Comput. **38**, 3365–3376 (2022)
18. Issenhuth, T., Mary, J., Calauzenes, C.: 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, pp. 619–635. Springer (2020)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 694–711. Springer (2016)
20. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: text-based real image editing with diffusion models (2023). arXiv:2210.09276
21. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: learning semantic correspondence with latent diffusion model for virtual try-on (2023). arXiv preprint arXiv:2312.01725
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). arXiv preprint arXiv:1312.6114
23. Kipf, T., Elsayed, G.F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., Greff, K.: Conditional object-centric learning from video (2021). arXiv preprint arXiv:2111.12594
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1931–1941 (2023)
25. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision, pp. 204–219. Springer (2022)
26. Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., Feng, D.D.: Eapt: efficient attention pyramid transformer for image processing. IEEE Trans. Multimed. **25**, 50–61 (2021)
27. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds (2022). arXiv preprint arXiv:2202.09778
28. Liu, Y., Jiang, T., Si, P., Zhu, S., Yan, C., Wang, S., Yin, H.: Unpaired semantic neural person image synthesis. Vis. Comput. 1–15 (2024)
29. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Adv. Neural Inf. Process. Syst. **33**, 11525–11538 (2020)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2017). arXiv preprint arXiv:1711.05101
31. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: latent diffusion textual-inversion enhanced virtual try-on (2023). arXiv preprint arXiv:2305.13501
32. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings, pp 1–11 (2023)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis (2023). arXiv preprint arXiv:2307.01952
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR. pp. 8748–8763 (2021)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pp. 234–241. Springer (2015)
37. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510 (2023)
38. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022a)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. **35**, 36479–36494 (2022)
40. Singh, G., Deng, F., Ahn, S.: Illiterate dall-e learns to compose (2021). arXiv preprint arXiv:2110.11405
41. Singh, G., Wu, Y.F., Ahn, S.: Simple unsupervised object-centric learning for complex and naturalistic videos. Adv. Neural Inf. Process. Syst. **35**, 18181–18196 (2022)
42. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, PMLR. pp. 2256–2265 (2015)
43. Song, D., Zhang, X., Zhou, J., Nie, W., Tong, R., Liu, A.A.: Image-based virtual try-on: a survey (2023). arXiv preprint arXiv:2311.04811
44. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1921–1930 (2023)
45. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 589–604 (2018)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004)
47. Wu, Z., Hu, J., Lu, W., Gilitschenski, I., Garg, A.: Slotdiffusion: object-centric generative modeling with diffusion models. Adv. Neural Inf. Process. Syst. **36** (2024)
48. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23550–23559 (2023)
49. Xie, Z., Zhang, W., Sheng, B., Li, P., Chen, C.P.: Bagfn: broad attentive graph fusion network for high-order feature interactions. IEEE Trans. Neural Netw. Learn. Syst. **34**, 4499–4513 (2021)

50. Yan, K., Gao, T., Zhang, H., Xie, C.: Linking garment with person via semantically associated landmarks for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17194–17204 (2023)
51. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18381–18391 (2023)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
53. Zhang, Z., Han, L., Ghosh, A., Metaxas, D., Ren, J.: Sine: single image editing with text-to-image diffusion models (2022). arXiv:2212.04489
54. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: a tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4606–4615 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Jianglei Ye is currently a postgraduate at the Faculty of Digital Media Technology, Hangzhou Dianzi University. Her research interests include fashion image processing, computer vision, and multimedia analysis.

Yigang Wang is currently a Professor at the School of Media and Arts, Hangzhou Dianzi University. He received the B.S. degree from the Department of Mathematics, Xi'an Jiaotong University, Xi'an, China, in 1992, and the M.S. and Ph.D. degrees from the Department of Applied Mathematics, Zhejiang University, Hangzhou, China, in 1995 and 1998, respectively. His research interests include computer graphics and virtual reality.

Fengmao Xie is currently an undergraduate student at the Faculty of Digital Media Technology, Hangzhou Dianzi University. His research interests include multimedia analysis and processing.

Qin Wang currently enrolled as a postgraduate student at Hangzhou Dianzi University's Faculty of Digital Media Technology, focuses his research on fashion image processing, multimedia analysis, and motion generation.

Xiaoling Gu received the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2017. She is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. Her research interests include computer vision, machine learning, and fashion data analysis. She was also an invited reviewer or a program committee member for top conferences and prestigious journals.

Zizhao Wu is currently an Associate Professor with the Faculty of Digital Media Technology, Hangzhou Dianzi University. He received the Ph.D. degree from the Department of Computer Science and Technology, Zhejiang University, in 2013. His main research interests include computer vision and computer graphics.