Latent Diffusion-Enhanced Virtual Try-On via Optimized Pseudo-Label Generation

Chenghu Du[*1,4], **Junyin Wang**[*1], **Feng Yu**[4†], **Shengwu Xiong**[2,3†]

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology
[2] Shanghai Artificial Intelligence Laboratory
[3] Interdisciplinary Artificial Intelligence Research Institute, Wuhan College
[4] School of Computer Science and Artificial Intelligence, Wuhan Textile University
{duch, wjy199708, xiongsw}@whut.edu.cn, yufeng@wtu.edu.cn

Abstract

Efficiently applying fully supervised learning to virtual try-on tasks is challenging due to the lack of paired ground truth in available training samples. Recent works have achieved virtual try-ons by employing self-supervised learning-based inpainting paradigms. However, this approach is heavily dependent on the constraints of inpainting masks. An incorrect mask can mislead the generated results, while overly large mask areas can lose essential original information, thereby hindering the synthesis of high-quality results. To address these problems, we propose a latent diffusion model-based virtual try-on network that achieves fully supervised learning using the concept of cycle consistency and knowledge distillation. Specifically, we divide our approach into pretext and downstream tasks. In the pretext task, we generate a pseudo-label (pseudo-person image) to form paired training samples, which enables the downstream task to achieve fully supervised learning. To prevent the unreliable pseudo-person image from introducing irresponsible prior knowledge, we propose a noise-covering strategy, which aims at fully optimizing the pseudo-label to eliminate the impact of the incorrect inpainting mask as much as possible. Additionally, we propose a skin refinement loss to further enhance the generation of details in the skin region. Extended experiments demonstrate that our proposed method is superior to state-of-the-art methods.

Introduction

The Virtual Try-On (VTON) task aims to realistically put clothing items from in-store images onto the person in user photos. When shopping for clothes online, consumers cannot inspect and try on clothing items as they would in a physical store. This significant demand, with its potential for practical and commercial value, has attracted widespread attention from researchers, leading to the development of image-based virtual try-on technology.

In this task, all available training samples (Han et al. 2018; Choi et al. 2021; Morelli et al. 2022) consist only of paired images (clothing and the person wearing the clothing) and do not include corresponding ground truths. Therefore, a practical self-supervised inpainting approach is typically employed (Choi et al. 2021; Lee et al. 2022; Xie et al. 2023; Gou et al. 2023; Morelli et al. 2023; Du et al. 2024; Li et al. 2023; Kim et al. 2024), which involves three steps: i) Train a parser to estimate a clothing-agnostic mask (see Figure 1 (a)) that covers the upper body region and part of the surrounding background. ii) Use the clothing-agnostic mask to remove the upper body of the person image (clothing-agnostic person, see Figure 1 (b)). iii) Inpaint the removed region using generative models (generative adversarial networks (Goodfellow et al. 2020) or diffusion models (Ho, Jain, and Abbeel 2020)) to synthesize the desired try-on result.

Figure 1: Visualization of the shortcomings of the inpainting paradigm based on self-supervised learning. (a) Clothing-agnostic mask. (b) Clothing-agnostic person. (c) Result: residual original information. (d) Larger clothing-agnostic mask. (e) Result: lost structural information. (c) and (e) display generated pseudo-person images. If an incorrect clothing-agnostic person or an unreliable pseudo-person image is used as input for the downstream task, it can introduce irresponsible knowledge.

In this paradigm, the clothing-agnostic person serves as the input, and the clothing-agnostic mask serves as the conditional input. This setup allows the generative model to learn how to fill in missing skin and optimize the clothing appearance on the clothing-agnostic person. However, if the parser predicts incorrect clothing-agnostic masks, it can directly affect the try-on outcomes. For example, if the original clothing is not fully removed, or areas such as hair or face are mistakenly removed, this will result in either the addition or loss of information (see Figure 1 (d)). Furthermore, due to the excessively large size of clothing-agnostic masks, the removed regions lose all their structural details. Relying solely on the generative model for empirical inpainting leads to a significant loss of structural coherence (see Figure 1 (e)).

[*]These authors contributed equally.
[†]First corresponding author: Shengwu Xiong.

Other approaches to solve this task include knowledge distillation (Issenhuth, Mary, and Calauzenes 2020) and cycle consistency paradigms (Zhu et al. 2017). Their processes are gradually converging, both utilizing an auxiliary model (teacher model) to train another main model (student model) to learn the reversible mapping between two different domains. Due to the absence of ground truth, these paradigms necessitate generating corresponding pseudo-labels to supervise the main network involved (Issenhuth, Mary, and Calauzenes 2020; Ge et al. 2021b,a; Du et al. 2023). However, most of these pseudo-labels are unreliable and inaccurate. Especially when errors occur in the clothing regions of the pseudo-labels, it can introduce incorrect knowledge into the main network, thereby hindering the production of high-quality try-on results.

To address these challenges, following the principle of cycle consistency and knowledge distillation, we design a new virtual try-on architecture based on a latent diffusion model. We modify the roles of the two generative networks, assigning each of them specific tasks. That is, the first network performs the pretext task to assist the second network in performing the downstream (try-on) task. Specifically, in the pretext task, we assign the first generative network to generate pseudo-labels (pseudo-person images) to be used as input for the downstream task. The key is that we propose two strategies to significantly enhance the reliability of the pseudo-person image and eliminate the impact of the incorrect clothing-agnostic mask as much as possible.

Specifically, to address the unreliability of the pseudo-person image generated by the model in the pretext task during training and inference, which can mislead the downstream task, we propose a noise-covering strategy to integrate target clothing and pseudo-person images, which can refine the pseudo-person image and address the loss of structural coherence caused by the clothing-agnostic mask. It prevents unreliable clothing regions from misleading the downstream task. Then, in order to further eliminate the residual extra information such as fabric from the generated skin area, we propose a skin refinement loss to further enhance the generation of details in the skin region. The contributions of this work are as follows:

- We propose a new latent diffusion model-based virtual try-on architecture that completes the virtual try-on task with the concept of cycle consistency and knowledge distillation.
- We propose the noise covering strategy to prevent unreliable clothing regions of pseudo-person images from misleading the downstream task.
- We propose the skin refinement loss to further enhance the generation of details in the skin region.
- Extensive experiments demonstrate the superiority of our method over state-of-the-art methods.

Related Work

Virtual Try-Ons. Recently, image-based virtual try-ons have emerged as a popular topic within the field of artificial intelligence. This task entails transferring a given garment onto a specific individual. With the rise of diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Rombach et al. 2022), current virtual try-on methods can generate highly realistic try-on results within the pixel space. DCI-VTON (Gou et al. 2023) and LaDI-VTON (Morelli et al. 2023) followed the inpainting paradigm (PbE (Yang et al. 2023)) using the diffusion model as the generator to synthesize warped clothing and reference person images, providing significant inspiration. KGI (Li et al. 2023) first optimizes clothing deformation through keypoint transformation and then employs the diffusion model to synthesize try-on results. StableVITON (Kim et al. 2024) introduces a zero cross-attention block that learns semantic correspondence between clothing and the human body, enabling try-ons in the latent feature space. Anydoor (Chen et al. 2024) employed a diffusion-based image generator that seamlessly teleports target clothing to user-specified locations (mask) on person images. However, these inpainting-based methods are entirely restricted to the areas specified by the clothing-agnostic mask. If there are errors in the mask, the results can be affected by the original content. To address these challenges, we design a new virtual try-on architecture based on the latent diffusion model. Specifically, we propose several strategies to significantly enhance the reliability of the pseudo-person image and eliminate the impact of the incorrect clothing-agnostic mask as much as possible.

Diffusion Models. Diffusion models are a class of likelihood-based models that have gained significant attention in generative modeling. These models introduce a novel approach by iteratively adding noise to data and then learning to reverse this process to generate realistic samples (Huang et al. 2024). This concept is inspired by the physical phenomenon of molecular diffusion, where particles spread from areas of high concentration to low concentration over time (Sohl-Dickstein et al. 2015). One pioneering work in this area is Diffusion Probabilistic Models (DPMs) (Sohl-Dickstein et al. 2015), which use Langevin dynamics to iteratively diffuse noise throughout the data space, gradually transforming noise into data samples while preserving the data distribution. This method allows for high-quality sample generation without relying on explicit likelihood functions. Building on this foundation, Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020) were introduced as a more scalable and efficient variant of DPMs. DMs utilize an autoregressive model to parameterize the diffusion process, simplifying computation and training. By incorporating deep learning techniques, DMs achieve state-of-the-art performance in image generation tasks, surpassing traditional generative models like Variational Autoencoders (VAEs) (Kingma and Welling 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al. 2020). Furthermore, Latent Diffusion Models (LDMs) (Rombach et al. 2022) were developed to process high-resolution synthesis while reducing inference costs. LDMs map high-resolution data to a latent sample space for diffusing and denoising, making them beneficial for image editing and inpainting. In this task, we take LDMs as the synthesis backbone to generate high-realistic try-on results.

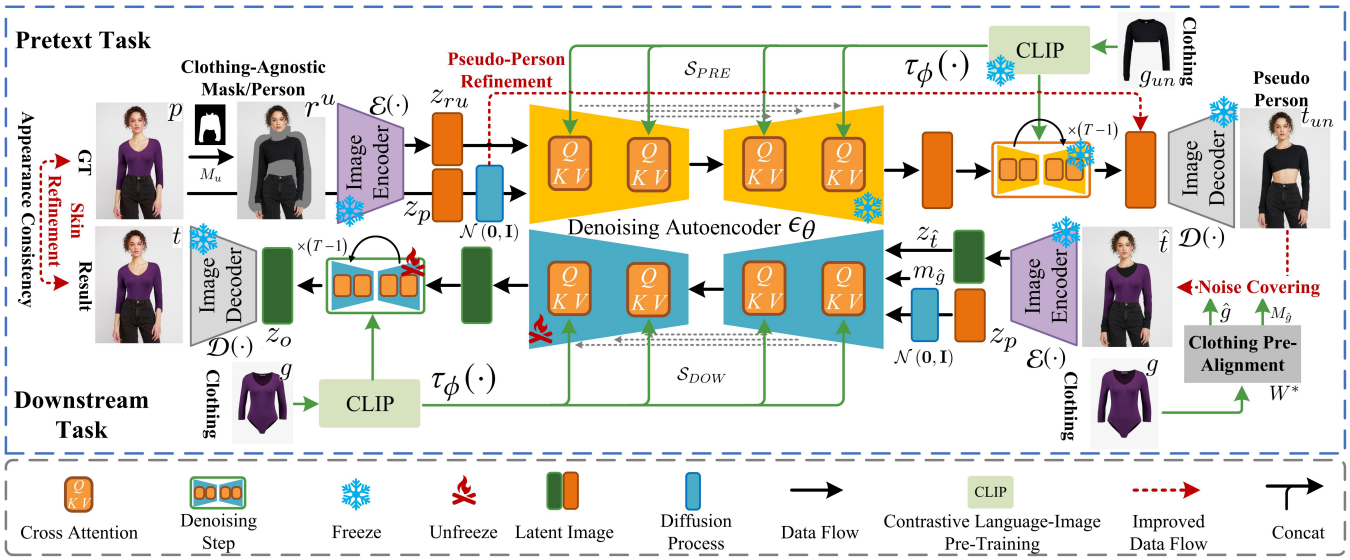

Figure 2: Overview of our proposed virtual try-on framework.

Preliminaries

Diffusion Models. Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) are probabilistic generative models, which consist of two processes: diffusion and reverse. The diffusion process is constrained to a Markov chain $q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I})$ of length T with noise level schedule $\{\beta_t\}_{t=1}^T$, which gradually adds Gaussian noise to source images (initial data distribution) $z_0 \sim q(z_0)$. Any noisy latent z_t at timestep t can be directly sampled using a closed-form sampling function:

$$z_t := \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (1)$$

where t uniformly sampled from $\{1, \ldots, T\}$. The coefficient $\alpha_t = 1-\beta_t$ defines the level of noise, and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. In the reverse process, a denoising autoencoder $\epsilon_\theta(\cdot)$ is trained to denoise ϵ from z_t to restore z_0 by minimizing the objective:

$$\min_\theta \mathbb{E}_{z_0,\epsilon,t}\left\|\epsilon_\theta(z_t, t) - \epsilon\right\|_2^2. \quad (2)$$

Conditional Diffusion Models. Given a target conditioning information embedding $\mathbf{c}$ (images, text prompts, or representations), the conditional diffusion model can sample an image z_0 from a conditional distribution $p_\theta(z|\mathbf{c})$ (Rombach et al. 2022). It can be formulated with a conditional denoising autoencoder: $\tilde{\epsilon} = \epsilon_\theta(z_t, t, \mathbf{c})$, where $\tilde{\epsilon}$ is a predicted noise, which is used to methodically remove the noise ϵ from z_t across T steps until the final result z_0 is obtained, formalized as:

$$\tilde{z}_0 := \left[z_t - \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}\right]/\sqrt{\bar{\alpha}_t}, \quad (3)$$

where $\tilde{z}_0$ is reconstructed z_0. In classifier-free guidance (Ho and Salimans 2021), conditional and unconditional predictions are linear amalgamated with a guidance scale w, to ensure substantial conditional influence and control over the generated output:

$$\tilde{\epsilon} = (1+w)\epsilon_\theta(z_t, t, \mathbf{c}) - w\epsilon_\theta(z_t, t). \quad (4)$$

Proposed Method

Problem Statement. Given a person image $p \in \mathbb{R}^{H\times W\times 3}$ and a target clothing image $g_{un} \in \mathbb{R}^{H\times W\times 3}$, the primary goal of virtual try-on tasks is to generate a try-on result $t_{un} \in \mathbb{R}^{H\times W\times 3}$, where the clothing worn in the person image is replaced with the target clothing from g_{un}. This is an unpaired image-to-image translation task (I2I) because available datasets consist only of pairs of images (clothing, person wearing the clothing), which cannot be used for fully supervised training due to the lack of ground truth. In other words, if using a triplet $\{$*(input 1, input 2), ground truth*$\}$ to represent the training sample unit used in this task, it can be represented as $\{(p, g), \varnothing\}$. Current feasible approaches (Gou et al. 2023; Kim et al. 2024) achieve self-supervised training by constructing new training sample units $\{(\{r^1, r^2, \ldots, r^M\}, g), p\}$ with representations $\{r^i\}_{i=1}^M$ of p. However, these generated representations, such as human parsing and keypoint maps, are not entirely reliable, thus negatively impacting the learning process.

Inspired by the concept of the cycle-consistency and following knowledge distillation, we use the first network to perform the **pretext task** to assist the second network in performing the **downstream (try-on) task** by using formed $\{(t_{un}, g), p\}$. During inference, we only utilize the second network. The entire architecture is depicted in Figure 2.

Clothing Pre-alignment

Before synthesizing the try-on result via LDM, the target clothing must first be aligned with the reference person's body posture to simulate real-world fabric properties and interactions. A simple and effective approach is to non-rigidly warp the target clothing g/g_{un} and inject it as conditioning into LDM, allowing LDM to refine the coarsely warped clothing by multiple denoising processes. Specifically, we employ an off-the-shelf warping model W^* (He et al. 2022; Xie et al. 2023) to estimate an appearance flow

f or $f_{un} \in \mathbb{R}^{H\times W\times 2}$ (Zhou et al. 2016). We then apply bilinear interpolation $\mathcal{B}$ to g or g_{un} using f or f_{un} to obtain the desired warped clothing $\hat{g}$ or $\hat{g}_{un}$. This process is formulated as i) $f = W^*(p, g)$, $\hat{g} = \mathcal{B}(g, f)$ and ii) $f_{un} = W^*(p, g_{un})$, $\hat{g}_{un} = \mathcal{B}(g_{un}, f_{un})$.

Pretext Task

The purpose of the pretext task is to assist the downstream task in achieving fully supervised learning by constructing a new triplet $\{(t_{un}, g), p\}$ based on the incomplete training sample triplet $\{(p, g), \varnothing\}$. To achieve this, we follow the approach in (Yang et al. 2023) to train an inpainting-based try-on PbE $\mathcal{S}_{PRE}$.

This is a self-supervised learning process. Specifically, we first obtain the clothing-agnostic mask $M_u \in \{0,1\}^{H\times W\times 1}$ (Lee et al. 2022). Then, the clothing-agnostic person ($r^u = p - M_u \otimes p$) is derived by using M_u, which is p without the upper body. Afterwards, we take g as conditioning and introduce a domain-specific encoder $\tau_\phi(\cdot)$ that projects g to an intermediate representation $\tau_\phi(g)$ with a cross-attention mechanism (Vaswani et al. 2017) to augment the denoising autoencoder ϵ_θ^{pre}, where $\tau_\phi(\cdot)$ is implemented with the frozen pre-trained CLIP (Radford et al. 2021) image encoder. The goal of $\mathcal{S}_{PRE}$ is to fill in the entire masked region to reconstruct p. This process is formulated as $\hat{p} = \mathcal{S}_{PRE}(r^u, \hat{g}, M_u, \tau_\phi(g))$, where $\hat{p}$ is the reconstructed p. Let

$$z_t := \sqrt{\bar{\alpha}_t} z_p + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where $z_t \in \mathbb{R}^{64\times 64\times 4}$, $z_p \in \mathbb{R}^{64\times 64\times 4} = \mathcal{E}(p)$, the loss to optimize ϵ_θ^{pre} is expressed as Equation (6):

$$\min_{\epsilon_\theta^{pre}} \mathbb{E}_{\mathcal{E}(p), z_{ru}, m_u, \tau_\phi(g), \epsilon\sim\mathcal{N}(0,I), t} \tag{6}$$
$$\left[\|\epsilon - \epsilon_\theta^{pre}(z_t, z_{ru}, m_u, \tau_\phi(g), t)\|_2^2 \right],$$

where $z_{ru} \in \mathbb{R}^{64\times 64\times 4} = \mathcal{E}(r^u)$, $m_u \in \{0,1\}^{64\times 64\times 1} =$ *Resize*(M_u). Following LDM, $\mathcal{E}$ is a pre-trained encoder used to embed the image into the latent space. *Resize*$(\cdot)$ represents scaling the mask proportionally so that its height and width match z_t. Afterwards, we utilize the well-trained $\mathcal{S}_{PRE}$ to generate desired try-on results t_{un} for arbitrary clothing g_{un}, forming the new triplet $\{(t_{un}, g), p\}$ to assist the downstream task.

Pseudo-Person Refinement. To prevent the generation of overly coarse t_{un}, we use m_u to refine the try-on result in the latent space, thereby largely preserving the original identity information of p. According to Equation (5), the refined latent variable z_f after denoising can be obtained by reversing the equation, thus, the desired try-on result t_{un} can be generated by decoding z_f with the pre-trained decoder $\mathcal{D}$, which is represented as Equation (7):

$$t_{un} := \mathcal{D}\big(m_u \otimes z_f + (1 - m_u) \otimes z_p\big), \tag{7}$$

where $\otimes$ denotes entry-wise multiplication. $\mathcal{D}$ is used to code latent variables into the pixel space.

Downstream (Try-on) Task

After obtaining the triplet $\{(t_{un}, g), p\}$, in this section, we take it to train a new try-on synthesizer $\mathcal{S}_{DOW}$ with fully supervised learning, where (t_{un}, g) serve as the input, and p as the ground truth.

Noise Covering. We assume that the pseudo-label t_{un} provided by the pretext task is always coarse and contains errors due to the influence of the incorrect clothing-agnostic mask M_u. Therefore, these errors will be concentrated in M_u, which mainly consists of the clothing part, arm part, and background. The influence of the background can be temporarily disregarded as its interference with the try-on effect is minimal. Beyond that, the focus lies on the presentation of the arm part and clothing part. To eliminate the potential impact of the incorrect clothing part of t_{un}, we directly cover the warped target clothing $\hat{g}$ deformed by W^* onto t_{un}. This process can be expressed as Equation (8):

$$\hat{t} = M_{\hat{g}} \otimes \hat{g} + (1 - M_{\hat{g}}) \otimes t_{un}, \tag{8}$$

where $\hat{t}$ is t_{un} covered by $\hat{g}$, and $M_{\hat{g}} \in \{0,1\}^{H\times W\times 1}$ is the mask of $\hat{g}$. The advantage of this approach is that we do not need to consider the quality of the clothing region in t_{un}, and the information outside the clothing region has been well-preserved by the pseudo-person refinement. In addition, this task has shifted from fusing two images to one of elimination or refinement, and it can even be retrained without adjusting the architecture of PbE.

In this setup, two scenarios can be roughly anticipated. When the clothing region of t_{un} is long-sleeved and g is short-sleeved, the primary task of $\mathcal{S}_{DOW}$ is to remove the excess fabric from t_{un}'s clothing to reconstruct the skin area. When the clothing region of t_{un} is short-sleeved and g is long-sleeved, the skin generation process is no longer necessary, and there is no need to consider skin removal, as the excess skin is already covered by $\hat{g}$. In this case, the main task of $\mathcal{S}_{DOW}$ is to refine the clothing region to produce realistic try-on results.

In addition, to prompt the denoising autoencoder $\epsilon_\theta^{dow}(\cdot)$ to eliminate the residual original information and the excess fabric, we take $M_{\hat{g}}$ as conditioning for LDMs to denoising. $M_{\hat{g}}$ is of the same size as the target clothing, so it does not cause issues like structural coherence loss mentioned previously. During training, $\hat{t}$ and p are embed into latent space $z_p, z_{\hat{t}} = \mathcal{E}(\hat{t})$, where $z_{\hat{t}} \in \mathbb{R}^{64\times 64\times 4}$. Then, Gaussian noise ϵ is introduced into the latent variable z_p (Equation (5)) and input into $\epsilon_\theta^{dow}(\cdot)$ for diffusion:

$$\mathcal{L}_1 = \mathbb{E}_{z_p, z_{\hat{t}}, m_{\hat{g}}, \tau_\phi(g), \epsilon\sim\mathcal{N}(0,I), t} \tag{9}$$
$$\left[\left\|\epsilon - \epsilon_\theta^{dow}(z_t, z_{\hat{t}}, m_{\hat{g}}, \tau_\phi(g), t)\right\|_2^2 \right],$$

where $m_{\hat{g}} \in \{0,1\}^{64\times 64\times 1} =$ *Resize*$(M_{\hat{g}})$. According to Equation (5), the refined latent variable z_o after denoising can be obtained by reversing the equation, thus, the desired try-on result t can be generated by decoding z_o with the pre-trained decoder $\mathcal{D}$, *i.e.* $t = \mathcal{D}(z_o)$. Finally, for the skin part, since ground truth is available in the downstream task, the loss function (9) can effectively optimize the skin areas that need to be generated. However, the power of global supervision is limited, and relying solely on the loss function (9) is insufficient to completely eliminate the residual original information of $\hat{t}$.

Skin Refinement. To address this issue, we design a skin refinement loss to expand the difference between the new

Figure 3: Qualitative comparison on the VITON-HD dataset. Red dashed boxes represent defects.

skin region $M_{skin} \in \{0,1\}^{H\times W\times 1}$ of t_{un} and p. This region represents the arm and neck regions where fabric needs to be removed, which can be obtained through the following operation:

$$M_{skin} = M_{\hat{g}_{un}} - (M_{\hat{g}} \otimes M_{\hat{g}_{un}}). \tag{10}$$

Thus, the loss can be expressed as Equation (11):

$$\mathcal{L}_2 = -\sum_{i=1}^{5} \|\phi_i(t_{un} \otimes M_{skin}) - \phi_i(p \otimes M_{skin})\|_1, \tag{11}$$

where $\|\|_1$ denotes ℓ_1 norm. $\phi_i(\cdot)$ denotes the feature map of the input of the i^{th} layer in the visual perception network ϕ, which is a VGG19 pre-trained on ImageNet. The layer $i \geq 1$ stands for 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2', respectively. Through the aforementioned operations, the difference region of skin between t_{un} and p will be further exaggerated.

Learning Objectives. The overall learning objective is:

$$\min_{\epsilon_\theta^{dow}} \mathcal{L}_1 + \lambda \cdot \mathcal{L}_2, \tag{12}$$

where λ is hyper-parameter.

Experiments

Datasets. Our experiments use VITON-HD (Choi et al. 2021), VITON (Han et al. 2018), which are two challenging datasets in virtual try-on. VITON consists of 16,253 image groups, each with a resolution of 256×192. Each group comprises a frontal-view woman image, a top clothing image, a semantic map, and a pose heatmap. The dataset is divided into a training set with 14,221 groups and a testing set with 2,032 groups. VITON-HD is a high-resolution version of the dataset, featuring a resolution of 512×384. It includes 13,679 image groups and is split into a training set with 11,647 groups and a testing set with 2,032 groups.

Figure 4: Qualitative comparison on the VITON dataset.

Training and Inference Details. We employ StableDiffusion v1.4 (Rombach et al. 2022) as the backbone for our architecture and initialize its denoising U-Net with the weights

Methods	SSIM ↑	FID ↓
ACGPN (Yang et al. 2020)	0.84	16.64
LM-VTON (Liu et al. 2021a)	0.85	17.18
DCTON (Ge et al. 2021a)	0.83	14.82
PF-AFN (Ge et al. 2021b)	0.89	<u>10.09</u>
ZFlow (Chopra et al. 2021)	0.88	15.17
RT-VTON (Yang, Yu, and Liu 2022)	-	11.66
SDAFN (Bai et al. 2022)	0.88	12.05
DressCode (Morelli et al. 2022)	0.89	13.71
PL-VTON (Zhang et al. 2023)	0.87	10.96
POVNet (Li, Zhang, and Forsyth 2023)	0.89	13.37
USC-PFN (Du et al. 2023)	<u>0.91</u>	10.47
Ours	**0.91**	**9.86**

Table 1: Quantitative comparison on VITON dataset. The best is in **bold**, and the second best is in <u>underline</u>.

Methods	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
VITON-HD	0.862	0.117	12.117	3.23
HR-VITON	0.878	0.105	11.265	2.73
GP-VTON	0.883	0.083	9.701	1.26
LaDI-VTON	0.864	0.096	9.480	1.99
PbE	0.802	0.143	11.939	3.85
DCI-VTON	0.880	0.080	**8.754**	**1.10**
KGI	0.892	0.064	10.330	1.74
Anydoor	0.821	0.099	10.850	2.46
StableVITON	0.864	0.070	9.465	1.40
Ours	**0.898**	**0.061**	<u>9.313</u>	<u>1.21</u>

Table 2: Quantitative comparison on VITON-HD dataset.

from the U-Net in PbE (Yang et al. 2023). During the training process, we utilize two NVIDIA RTX 4090 GPUs for a duration of 2 days. The AdamW optimizer (Loshchilov and Hutter 2017) is employed with a learning rate of 1×10^{-4}, and the batch size is set to 2 for training over 40 epochs. For inference, we adopt the pseudo linear multi-step (PLMS) sampling method (Liu et al. 2021b), setting the number of sampling steps to 50. The hyperparameter in the loss function is set as follows: $\lambda = 1 \times 10^{-4}$.

Evaluation Metrics. To facilitate quantitative evaluation of our method against baseline methods, we employ Structure Similarity (SSIM) (Seshadrinathan and Bovik 2008) to assess the similarity between generated and real images in terms of luminance, contrast, and structural similarity. We also use Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) to measure the semantic similarity between generated and real images. Furthermore, we utilize Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018) to evaluate the distribution similarity between generated and real images. Note that lower FID and KID scores indicate higher image quality in the generated images.

Baseline Methods. For qualitative and quantitative comparisons, we select state-of-the-art virtual try-on methods of different generative models as follows:

• GAN-based methods: ACGPN (Yang et al. 2020), LM-VTON (Liu et al. 2021a), DCTON (Ge et al. 2021a), PF-AFN (Ge et al. 2021b), ZFlow (Chopra et al. 2021), RT-VTON (Yang, Yu, and Liu 2022), SDAFN (Bai et al. 2022), DressCode (Morelli et al. 2022), PL-VTON (Zhang et al. 2023), POVNet (Li, Zhang, and Forsyth 2023), VITON-HD (Choi et al. 2021), HR-VITON (Lee et al. 2022), GP-VTON (Xie et al. 2023), and USC-PFN (Du et al. 2023).

• Diffusion-based methods: PbE (Yang et al. 2023), DCI-VTON (Gou et al. 2023), LADI-VTON (Morelli et al. 2023), KGI (Li et al. 2023), Anydoor (Chen et al. 2024), and StableVITON (Kim et al. 2024).

Figure 5: Visual ablation study for the clothing pre-alignment and the skin refinement loss.

Comparisons

Qualitative Results. We compare our method with state-of-the-art (SOTA) baseline methods, as shown in Figures 3 and 4. The results indicate that Style-Flow has some generalization issues; it struggles to retain certain identity information and clothing details. GP-VTON, due to the instability of GANs, produces a significant amount of noise in the results (1st row). Moreover, the synthesized clothing appears overly smooth, completely ignoring spatial structure and lighting/shadow issues (2nd and 3rd rows). While DCI-VTON successfully puts on the clothing, the generated clothing shape is constrained by the original clothing, such as in the collar area. StableVITON, which relies on cross-attention, is unreliable for some clothing samples; it can transform long sleeves into short sleeves (2nd row), be insensitive to the shape of rare samples (3rd row), and lose a lot of clothing information (1st row). Our method not only preserves identity information—such as the necklace in the 1st row—but also faithfully reproduces the clothing's shape, folds, and lighting/shadows in the try-on results.

Quantitative Results. We use the official weights of recent SOTA methods to re-measure their quantitative values. Specifically, we measure the SSIM and LPIPS scores for paired clothing and person images, as well as the FID and KID scores for unpaired clothing and person images. As shown in Tables 1 and 2, in the paired setting, our SSIM and LPIPS scores are higher than or equal to those of the SOTA GAN-based methods GP-VTON (Xie et al. 2023) and USC-PFN. Compared to the SOTA LDM-based methods StableVITON (Kim et al. 2024) and Anydoor (Chen et al. 2024), our SSIM scores completely surpassed theirs. In the unpaired setting, our FID and KID scores also demonstrate advantages in comparison to these baseline methods. The re-

Methods	Devices	Training time	Data	Total Params (M)	Memory Usage (G)	Inference Time (s)
TryonDiffusion	32 TPU-v4	9 Days	4M	$\sim$ 3261.8	$\sim$ 40.8	-
StableVITON	4 A100	4 Days	30K	$\sim$ 1607.9	$\sim$ 7.9	$\sim$ 16
Ours	2 4090	2 Day	30K	$\sim$ 1246.7	$\sim$ 11.2	$\sim$ 12

Table 3: Comparison of training overhead.

Methods	FID $\downarrow$	KID $\downarrow$
A: w/o Clothing Pre-alignment	10.984	2.29
B: w/o Noise Covering	10.097	1.88
C: w/o Skin Refinement	9.427	1.43
Full Ours	**9.313**	**1.21**

Table 4: Ablation study of our proposed components on the VITON-HD dataset.

sults objectively demonstrate the performance advantage of our proposed method.

Training Overhead. We compare the training overhead of our method with two well-known diffusion-based virtual try-on models, as shown in Table 3. The results indicate that TryonDiffusion (Zhu et al. 2023) consumed a significant amount of computational resources and training time to train on a 4-million-sized dataset. However, its performance in qualitative experiments was not ideal. The SOTA model, StableVITON (Kim et al. 2024), required only 4 A100 GPUs and achieved excellent performance on a 30,000-sized dataset within 4 days. Our method achieves superior results compared to StableVITON while using fewer computational resources and less training time, inference time, and parameter quantity on the same sample size, further demonstrating the superiority of our approach.

Ablation Study

We removed three components from the architecture and tested them separately to verify the effectiveness of each component, the results are shown in Table 4 and Figure 5.

Effectiveness of Clothing Pre-alignment. As we mentioned, the clothing pre-alignment component is crucial for preserving garment details. It helps the diffusion process to quickly reconstruct the shape of the clothing. Removing it, or even using cross-attention (Vaswani et al. 2017; Kim et al. 2024), results in only a coarse alignment of the clothing shape. This limitation is due to the nature of CNNs, as convolutions cannot handle large spatial deformations. As shown in Table 4 and Figure 5, when we removed the clothing pre-alignment, the FID increased by approximately 17.98%, and the KID increased by approximately 89.26%. The clothing deformation becomes very unnatural, significantly violating the fabric's physical properties and interaction states. These results confirm the effectiveness of clothing pre-alignment.

Effectiveness of Noise Covering. Noise covering can further prevent the incompetence of t_{un} as an input, meaning that the issue of distortion in the generated clothing region can be resolved. This would positively impact the training process in downstream tasks. Replacing t_{un} with the clothing-agnostic person r^u would result in the loss of significant human body details. Therefore, we use the warped target clothing to cover most of the clothing region of t_{un} to conduct ablation research. When we replace t_{un} with the clothing-agnostic person r^u, as indicated in Table 4, the FID increased by approximately 8.42%, and the KID increased by approximately 55.37%. This clearly demonstrates that the noise covering is more effective than using the clothing-agnostic person, as it can also preserve the original human body details.

Effectiveness of Skin Refinement. To preserve more skin details, we introduce the skin refinement loss, as given in Equation (11), to emphasize the importance of the skin region during the training process. When we exclude Equation (11), as shown in Figure 5, although the overall results are satisfactory, the skin region remains relatively rough. However, when we incorporate Equation (11), highly realistic skin can be generated. This demonstrates the effectiveness of our skin refinement loss.

Conclusions and Limitations

In this paper, we propose a latent diffusion model-based virtual try-on network designed to achieve fully supervised learning through the concept of the cycle-consistency and knowledge distillation. Specifically, we divide our work into two phases: pretext and downstream tasks. To refine pseudo-person images and address the loss of structural coherence caused by the clothing-agnostic mask, we propose a noise-covering strategy, which aims at fully optimizing the pseudo-label to eliminate the impact of the incorrect inpainting mask as much as possible. Additionally, we propose a skin refinement loss to further enhance the generation of details in the skin region. Extended experiments demonstrate that our proposed method is superior to SOTA methods.

While our proposed method improves results, its practical application is actually limited by the dataset. Specifically, the diversity of clothing available on the market is vast, with fashion designers continuously introducing new shapes and styles. Consequently, the scarcity of training samples greatly restricts the effectiveness of current virtual try-on algorithms. Additionally, performance limitations also arise from the variability and unpredictability of the human body poses. In the future, we aim to collect more diverse and extensive datasets and design more effective network architectures to enhance the generalization capabilities of virtual try-on algorithms.

Acknowledgments

This work was in part supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160604) and NSFC (Grant No. 62176194), and the Key Research and Development Program of Hubei Province (Grant No. 2023BAB083), the Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ-JYRC-2022-76, SKJC-2022-PTDX-031), the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031), and the Huawei Kunpeng-Ascend Innovation Incentive Programme.

References

Bai, S.; Zhou, H.; Li, Z.; Zhou, C.; and Yang, H. 2022. 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*, 409–425. Springer.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Chopra, A.; Jain, R.; Hemani, M.; and Krishnamurthy, B. 2021. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5433–5442.

Du, C.; Liu, S.; Xiong, S.; et al. 2023. Greatness in Simplicity: Unified Self-Cycle Consistency for Parser-Free Virtual Try-On. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Du, C.; Wang, J.; Rong, Y.; Liu, S.; Liu, K.; and Xiong, S. 2024. CycleVTON: A Cycle Mapping Framework for Parser-Free Virtual Try-On. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1618–1625.

Ge, C.; Song, Y.; Ge, Y.; Yang, H.; Liu, W.; and Luo, P. 2021a. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16928–16937.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021b. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. 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*, 7599–7607.

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.

He, S.; Song, Y.-Z.; Xiang, T.; and Xiang, T. 2022. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3470–3479.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Chen, S.; and Cao, L. 2024. Diffusion Model-Based Image Editing: A Survey. *arXiv preprint arXiv:2402.17525*.

Issenhuth, T.; Mary, J.; and Calauzenes, C. 2020. 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*, 619–635. Springer.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8176–8185.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, S.; Gu, G.; Park, S.; Choi, S.; and Choo, J. 2022. 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*, 204–219. Springer.

Li, K.; Zhang, J.; and Forsyth, D. 2023. POVNet: Image-Based Virtual Try-On Through Accurate Warping and Residual. *IEEE transactions on pattern analysis and machine intelligence*.

Li, Z.; Wei, P.; Yin, X.; Ma, Z.; and Kot, A. C. 2023. Virtual Try-On with Pose-Garment Keypoints Guided Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22788–22797.

Liu, G.; Song, D.; Tong, R.; and Tang, M. 2021a. Toward realistic virtual try-on through landmark guided shape matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2118–2126.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2021b. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. LaDI-VTON: latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2231–2235.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Seshadrinathan, K.; and Bovik, A. C. 2008. Unifying analysis of full reference image quality assessment. In *2008 15th IEEE International Conference on Image Processing*, 1200–1203. IEEE.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; and Liang, X. 2023. 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 (CVPR)*, 23550–23559.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, H.; Yu, X.; and Liu, Z. 2022. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3460–3469.

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.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Han, X.; Zhang, W.; Lan, X.; Yao, H.; and Huang, Q. 2023. Limb-Aware Virtual Try-On Network with Progressive Clothing Warping. *IEEE Transactions on Multimedia*, 1–16.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 286–301. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.