Mask Does Not Matter: A Unified Latent Diffusion-Enhanced Framework for Mask-Free Virtual Try-On

Chenghu Du[1] , **Junyin Wang**[1] , **Kai Liu**[1] , **Shengwu Xiong**[3,4*] and **Yi Rong**[1,2*]
[1]School of Computer Science and Artificial Intelligence, Wuhan University of Technology
[2]Sanya Science and Education Innovation Park, Wuhan University of Technology
[3]Shanghai Artificial Intelligence Laboratory
[4]Interdisciplinary Artificial Intelligence Research Institute, Wuhan College
{duch, wjy199708, liukai356, xiongsw, yrong}@whut.edu.cn

Abstract

A good virtual try-on model should introduce minimal redundant conditional information to avoid instability and increase inference efficiency. Existing methods rely on inpainting masks to guide the generation of the object, but the masks, generated by unstable human parsers, often produce unreliable results with fabric residues due to wrong segmentation. Moreover, large mask regions can lose spatial structure and identity information, requiring extra conditional inputs to compensate, which increases model instability and reduces efficiency. To tackle the problem, we present a novel **M**ask-**F**ree virtual **T**ry-**ON** (MFTON) framework. Specifically, we propose a mask-free strategy to eliminate all denoising conditions except for clothing and person images, thereby directly extracting spatial structure and identity information from the person image to improve efficiency and reduce instability. Additionally, to optimize the generated clothing regions, we propose a clothing texture-aware attention mechanism to enable the model to focus on texture generation with significant visual differences. We then introduce a geometric detail capture loss to further enable the model to capture more high-frequency information. Finally, we propose an appearance consistency inference method to reduce the initial randomness of the sampling process significantly. Extensive experiments on popular datasets demonstrate that our method outperforms state-of-the-art virtual try-on methods. Our source code will be available at: https://github.com/du-chenghu/MFTON.

1 Introduction

Image-based virtual try-ons (VTON) aim to transfer the target clothing from an in-shop clothing image onto the corresponding clothing area on a reference person within a user's photo [Gou *et al.*, 2023; Xie *et al.*, 2023; Du *et al.*, 2024; Kim *et al.*, 2024; Chen *et al.*, 2024]. With the advancement of generative artificial intelligence, VTON has gradually garnered widespread attention from researchers and consumers due to its significant economic and practical value. However, enhancing the realism of the clothing and limb skin areas in the generated results has been an ongoing issue.

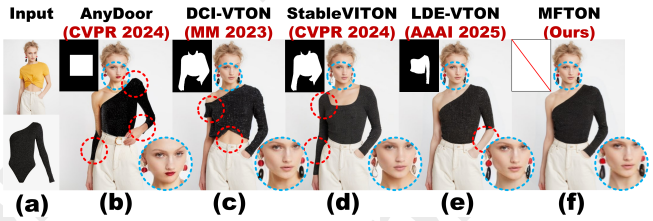

Figure 1: Comparison of negative impacts of different inpainting masks. (a) Box mask. (b,c) Clothing-agnostic mask. (d) Clothing-specific mask. Our method (MFTON) does not use any mask as the inpainting mask, achieving a more natural and realistic virtual try-on effect. **Dashed circles** highlight the limitations of each method.

Most of the previous works [Gou *et al.*, 2023; Kim *et al.*, 2024; Chen *et al.*, 2024; Wang *et al.*, 2024; Du *et al.*, 2025], following the inpainting paradigm, utilized an **inpainting mask** to focus the generative model on the clothing and limb skin areas that needed to be generated (inpainting object), thereby enhancing the realism, which has demonstrated outstanding performance. However, obtaining a good mask that accurately covers the inpainting object is challenging. First, the method based on a box mask (see Fig. 1(b)) is proposed [Chen *et al.*, 2024], which is easy to create and widely applicable to any clothing and human body, making it highly universal. However, the entire box area is not flexible enough, an overly large box can cover identity regions mistakenly ($e.g.$, face and hair), while an overly small box can leave parts of the inpainting object uncovered, resulting in residuals of original fabrics. Moreover, the human structure and original limb skin information within the entire box area are missing. It forces the model to rely on additional structure conditions as input for supplementation, which increases the model's instability and reduces efficiency.

To address these issues, the method based on a clothing-agnostic mask (see Fig. 1(c, d)) is proposed, which uses a human parser to outline the approximate shape of the inpainting object, thereby narrowing down the mask's contour [Gou

*Corresponding authors.

et al., 2023; Kim *et al.*, 2024] to minimize wrong segmentation. However, it merely alleviates the above issues to some extent and still requires the introduction of additional structure conditions to supplement the precise human structural information. In addition, since this mask also fails to capture the shape information of the target clothing, even if the same mask is used in (c, d), different models produce different inpainting results due to varying interpretations of the mask.

Recently, a pioneering method [Du *et al.*, 2025] based on a clothing-specific mask (see Fig. 1(e)) has been proposed. It uses a target clothing mask as the inpainting mask to specify the exact regions of the clothing and limbs to be generated, thereby eliminating the need for additional conditions. Nonetheless, this approach cannot avoid wrong segmentation caused by a specially trained parser, leading to mismatches between the target clothing and the mask shape.

To address these problems, we attempt to eliminate the negative impacts of these incorrect masks. To this end, we present a novel **M**ask-**F**ree **V**irtual **T**ry-**ON** framework (MFTON), which produces highly photo-realistic results without using any inpainting masks as the input condition. For this purpose, we transform the traditional approach of first removing the original clothing and limb regions and then inpainting the target clothing and limbs by the model, into an approach that first generates a temporary try-on result (TTR) and then adaptively eliminates the original fabric residue. On the basis of method [Du *et al.*, 2025], we overlay the warped target clothing onto the original clothing region of the reference person (RP) to form TTR, which is then used to transfer back to the RP. Thus, two scenarios may arise: If the original clothing region is fully covered by the target clothing, the model refines the clothing area while retaining the areas outside the clothing. Otherwise, if the original clothing region is not completely covered by the target clothing, the model must learn to eliminate the uncovered areas of the original clothing. However, at this point, the target clothing and the original fabric residue in TTR may be perceived by the model as a complete single garment in the absence of additional cues, due to the frequent occurrence of the first scenario. To address this issue, we use RP as the denoising condition, and the complete original clothing on RP allows the model to capture where the residual area in TTR is located due to the residual area originating from RP. As a result, the model can rely solely on the RP and TTR, eliminating the need for the mask and the additional structure condition, to achieve a robust VTON model.

Furthermore, the lack of region guidance in mask-free frameworks can reduce the model's focus on clothing areas. To address this, we propose a clothing texture-aware attention mechanism and a geometric detail capture loss to compel the model to focus on clothing generation from both texture and spatial detail perspectives. Finally, to realistically reconstruct the details of the inpainting object, we design an appearance consistency inference method that initiates the inference process from a posterior Gaussian noise, significantly reducing the initial randomness of the sampling process.

In general, the contributions of this work are as follows:

- We present a novel mask-free framework for virtual try-on, to produce highly photo-realistic results without using any masks as the denoising condition, providing a new perspective for mask-free virtual try-on strategy.
- We propose a clothing texture-aware attention mechanism to enable the model to focus on texture generation with significant visual differences.
- We propose a geometric detail capture loss to further enable the model to capture high-frequency information.
- We propose an appearance consistency inference to reduce the initial randomness of the sampling process.

2 Preliminary: Diffusion Models

Diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Nichol and Dhariwal, 2021], as probabilistic generative models, encompass a two-step process: diffusion and its reverse. The diffusion phase adheres to a Markov chain defined by $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I})$, spanning T iterations with a noise schedule $\{\beta_t\}_{t=1}^T$. This schedule incrementally corrupts the initial data, $\boldsymbol{z}_0 \sim q(\boldsymbol{z}_0)$, with Gaussian noise. Each noisy latent state $\boldsymbol{z}_t$ at any timestep t can be sampled directly through a closed-form sampling function:

$$\boldsymbol{z}_t := \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where t is uniformly sampled from $\{1, \ldots, T\}$. The noise level is determined by $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The reverse process starts with a noisy data $\boldsymbol{z}_T \sim \mathcal{N}(0, \mathbf{I})$ at step T and gradually denoises it using known real distributions $q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ for each step:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \mu_\theta(\boldsymbol{z}_t, t), \Sigma_\theta(\boldsymbol{z}_t, t)). \tag{2}$$

To achieve this, a denoising autoencoder $\epsilon_\theta(\cdot)$ is trained to remove noise ϵ from $\boldsymbol{z}_t$ to reconstruct $\boldsymbol{z}_0$ by optimizing the following objective:

$$\min_\theta \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{\epsilon}, t} \|\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) - \boldsymbol{\epsilon}\|_2^2. \tag{3}$$

3 Proposed Approach: MFTON

Problem Statement. Given an arbitrary clothing image $\boldsymbol{C}_{un} \in \mathbb{R}^{H\times W\times 3}$ and a reference person image $\boldsymbol{P} \in \mathbb{R}^{H\times W\times 3}$, the VTON task aims to generate a try-on result $\boldsymbol{T}_{un} \in \mathbb{R}^{H\times W\times 3}$, where the clothing worn by the person in $\boldsymbol{P}$ is replaced with the target clothing from $\boldsymbol{C}_{un}$. Here, H, W, and 3 represent the height, width, and number of channels of the image, respectively.

Framework. Fig. 2 illustrates the overview of our proposed method. It consists of three modules: a *Mask-Free Strategy* used to present the mask-free pipeline, a *Clothing Texture-aware Attention Module* for injecting clothing texture information, and a *Geometric Detail Capture Module* for supervising the geometric details of clothing.

Clothing Pre-processing. To mimic the interaction between the clothing and the human body in reality, the target clothing $\boldsymbol{C}$ needs to be non-rigidly warped to align with the human body posture of $\boldsymbol{P}$ naturally. The off-the-shelf warping network $\mathcal{W}$ [He *et al.*, 2022; Xie *et al.*, 2023] is directly adopted to generate the warped clothing $\boldsymbol{C}^w$:

$$\mathcal{F} = \mathcal{W}(\boldsymbol{C} \odot \boldsymbol{C}^M, \boldsymbol{P}), \quad \boldsymbol{C}^w = \mathcal{B}(\boldsymbol{C} \odot \boldsymbol{C}^M, \mathcal{F}), \tag{4}$$

where $\mathcal{B}(\cdot,\cdot)$ is the bi-linear interpolation based on generated deformation field $\mathcal{F} \in \mathbb{R}^{H\times W\times 2}$. $\boldsymbol{C}^M$ is shape parsing of $\boldsymbol{C}$.

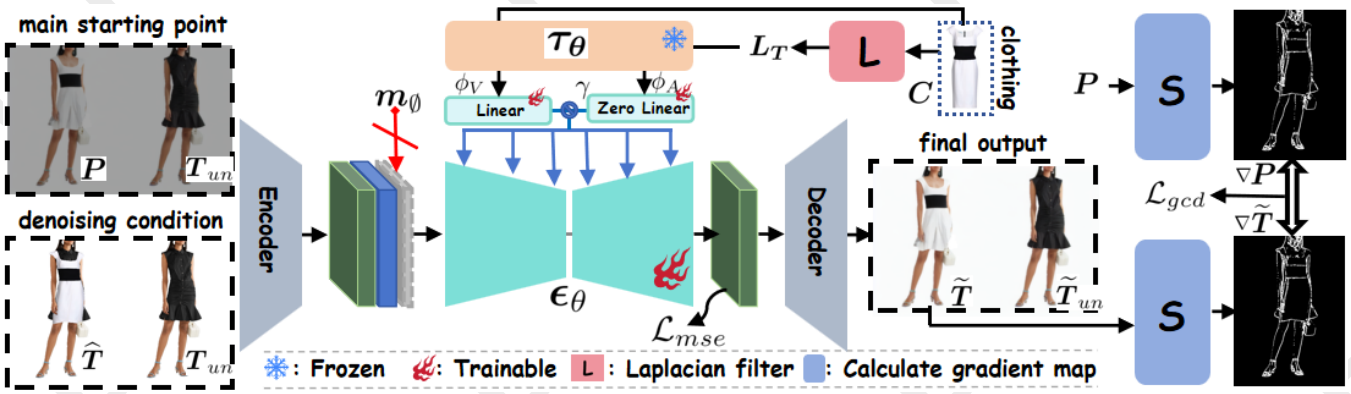

Figure 2: Overview architecture of our proposed MFTON, containing a diffusion-based generator $\boldsymbol{\epsilon}_\theta$, a geometric detail capture module, and a clothing texture-aware attention module. The system is optimized using a mean squared error loss $\mathcal{L}_{mse}$ and a gradient consistency loss $\mathcal{L}_{gcd}$. The encoder and decoder come from KL-regularized autoencoder, respectively.

3.1 Proposed Framework

Mask-Free Strategy. The role of the inpainting mask is to indicate the masked human body regions (the inpainting image) that need to be inpainted, but it simultaneously loses the spatial structure and identity information of those regions. Instead of utilizing additional human parsing to supplement this information, we use the random try-on result $\boldsymbol{T}_{un}$ of $\boldsymbol{P}$, generated by the off-the-shelf model [Kim *et al.*, 2024], as a prior guiding condition to replace the mask and human parsing. This is because $\boldsymbol{T}_{un}$ contains the complete spatial structure and identity information. Specifically, we first overlay the warped clothing $\boldsymbol{C}^w$ onto the corresponding area of $\boldsymbol{T}_{un}$ to generate the temporary try-on result $\widehat{\boldsymbol{T}} = \boldsymbol{T}_{un} + \boldsymbol{C}^w$. However, $\widehat{\boldsymbol{T}}$ is coarse, with the clothing area appearing overly smooth and lacking natural wrinkles. Moreover, the original clothing area still retains significant fabric residues, as shown in Fig. 2. To eliminate these residues, we use $\boldsymbol{T}_{un}$ as another condition to inform the model of the locations of the residual original clothing areas in $\widehat{\boldsymbol{T}}$ and directly inject all identity information through the conditioning effect.

Thus, by setting the ground truths of $\widehat{\boldsymbol{T}}$ and $\boldsymbol{T}_{un}$ as $\boldsymbol{P}$ and $\boldsymbol{T}_{un}$, respectively, the model learns to transform $\widehat{\boldsymbol{T}}$ into a result $\widetilde{\boldsymbol{T}}$ whose appearance is infinitely close to $\boldsymbol{P}$ through the process of adding noise to $\boldsymbol{P}$ and $\boldsymbol{T}_{un}$ and then denoising them with conditions $\widehat{\boldsymbol{T}}$ and $\boldsymbol{T}_{un}$. The noise addition process is represented as Eq. (5):

$$\boldsymbol{z}_t^p = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^p + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{z}_t^{up} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^{up} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{5}$$

where, $\boldsymbol{z}_0^p, \boldsymbol{z}_0^{up} \in \mathbb{R}^{(H/f)\times(W/f)\times 4} = \mathcal{E}(\boldsymbol{P}), \mathcal{E}(\boldsymbol{T}_{un})$, $\mathcal{E}$ is the encoder of KL-regularized autoencoder with its default latent-space downsampling factor $f = 8$. Then, the latent feature $\boldsymbol{z}_0^{up}$, $\boldsymbol{z}_0^p$ and the denoising condition are concatenated along the channel dimension, represented as Eq. (6):

$$\boldsymbol{\psi}_t^{total} = \left[[\boldsymbol{z}_t^p, \boldsymbol{z}_t^{up}]_S; [\mathcal{E}(\widehat{\boldsymbol{T}}), \mathcal{E}(\boldsymbol{T}_{un})]_S; [\boldsymbol{m}_\emptyset, \boldsymbol{m}_\emptyset]_S\right]_C, \tag{6}$$

where $[,]_C$ and $[,]_S$ denotes the concatenation operation along the channel and spatial dimension, respectively. $\boldsymbol{m}_\emptyset$ represents not inputting any mask. However, it is presented here because there is another option: to prevent catastrophic forgetting of pre-trained $\boldsymbol{\epsilon}_\theta$'s weights due to changes in the number of channels, thereby increasing the convergence burden, $\boldsymbol{m}_\emptyset$ can be set to all ones or all zeros in $\boldsymbol{\epsilon}_\theta$. $\boldsymbol{\psi}_t^{total}$ is severed as the input to train $\boldsymbol{\epsilon}_\theta$, represented as Eq. (7):

$$\mathcal{L}_{mse} = \mathbb{E}_{\boldsymbol{z}_0^p, \boldsymbol{z}_0^{up}, \boldsymbol{\epsilon}, t}\left\|\boldsymbol{\epsilon}_\theta\left(\boldsymbol{\psi}_t^{total}, t\right) - \boldsymbol{\epsilon}\right\|_2^2. \tag{7}$$

Clothing Texture-aware Attention Module. Clothing images typically contain abundant global information, such as text, shape, color, and pattern, which must be preserved. Although previous methods [Yang *et al.*, 2023; Gou *et al.*, 2023; Kim *et al.*, 2024] using Language-Image Pre-Training models [Radford *et al.*, 2021] could easily inject the global attributes of the inpainting object into the attention module, they were insensitive to **texture** attributes with significant visual differences. To enable the model to focus on texture generation, we design a clothing texture-aware attention module to additionally inject texture information into cross-attention.

Specifically, we use an isotropic Laplacian filter to process the target clothing $\boldsymbol{C} \in \mathbb{R}^{H'\times W'\times 3}$. This filter can detect texture variations comprehensively in any direction without omission, thereby obtaining a texture-aware map $\boldsymbol{L_T} \in \mathbb{R}^{H'\times W'\times 3}$. Subsequently, we utilize a domain-specific encoder $\boldsymbol{\tau_\theta}$ to project $\boldsymbol{C}$ and $\boldsymbol{L_T}$ into intermediate representations $\boldsymbol{\tau_\theta}(\boldsymbol{C})$ and $\boldsymbol{\tau_\theta}(\boldsymbol{L_T})$, which are then passed through linear layers and combined using a residual connection $\oplus$ with a scale parameter γ, denoted as Eq. (8):

$$\boldsymbol{B} = \phi_V\left(\boldsymbol{\tau_\theta}\left(\boldsymbol{C}\right)\right) \oplus \gamma \cdot \phi_A\left(\boldsymbol{\tau_\theta}\left(\boldsymbol{L_T}\right)\right), \tag{8}$$

where ϕ_A is zero linear layer, a standard 1×1 linear layer with both weight and bias initialized to zero. It keeps the gradient of ϕ_A small enough in the early stage of training so that ϕ_A can focus on learning to provide a high-level texture understanding compatible with the ϕ_V. γ balances the influence of general and texture-specific conditions on the generation. $\boldsymbol{B}$ is then mapped to the intermediate layers of the UNet $\boldsymbol{\epsilon_\theta}$ via the cross-attention layer, represented as Eq. (9):

$$\boldsymbol{Q} = \boldsymbol{W_q^{(i)}}\boldsymbol{F_{in}^{(i)}}, \quad \boldsymbol{K}^\dagger = \boldsymbol{W_k^{(i)}}\boldsymbol{B}, \quad \boldsymbol{V}^\dagger = \boldsymbol{W_v^{(i)}}\boldsymbol{B}, \tag{9}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}^{\dagger}$, and $\boldsymbol{V}^{\dagger}$ are the query, key, and value matrices of the attention module, respectively. $\boldsymbol{W}_{\boldsymbol{q}}^{(i)}$, $\boldsymbol{W}_{\boldsymbol{k}}^{(i)}$, and $\boldsymbol{W}_{\boldsymbol{v}}^{(i)}$ are the projection matrices for the (i)-th scale block. $\boldsymbol{F}_{\boldsymbol{in}}^{(i)}$ is noise feature in (i)-th layer of $\boldsymbol{\epsilon}_\theta$. The output $\boldsymbol{F}_{\boldsymbol{out}}^{(i)}$ can be represented as Eq. (10):

$$\boldsymbol{F}_{\boldsymbol{out}}^{(i)} = \texttt{Att}\left(\boldsymbol{Q}, \boldsymbol{K}^{\dagger}, \boldsymbol{V}^{\dagger}\right) = \texttt{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\dagger T}}{\sqrt{d}}\right)\boldsymbol{V}^{\dagger}. \quad (10)$$

Thus, Eq. (7) is transformed as Eq. (11):

$$\mathcal{L}_{mse} = \mathbb{E}_{\boldsymbol{z}_0^p, \boldsymbol{z}_0^{up}, \boldsymbol{\epsilon}, \boldsymbol{C}, t}\left\|\boldsymbol{\epsilon}_\theta\left(\boldsymbol{\psi}_t^{total}, \boldsymbol{B}, t\right) - \boldsymbol{\epsilon}\right\|_2^2. \quad (11)$$

Geometric Detail Capture Module. Although the texture-aware attention module can already enhance the model's focus on texture details to faithfully reconstruct the appearance of clothing, preserving complex **high-frequency** details in clothing images—such as text, patterns, badges, and stripes—remains challenging, as the injected information is global and concise. To further enable the model to capture more high-frequency information, such as the arrangement and layout of stripes, we design a new gradient loss function. Specifically, since the generated try-on result $\widetilde{\boldsymbol{T}}$ is rich in noise in the early stages of training, we abandon the previously employed Laplacian filter, which uses a second-order derivative that amplifies the impact of noise. Instead, we adopt the Sobel operator, which has a noise suppression effect, to obtain the gradient images of the ground truth $\boldsymbol{P}$ and $\widetilde{\boldsymbol{T}}$. Then, we replace the vanilla L1 loss with a Log-L1 loss to penalize abnormal gradient points more heavily [Zhao *et al.*, 2021], thereby guiding the model to focus on intricate local details, formulated as Eq. (12):

$$\mathcal{L}_{gcd} = \frac{1}{2HW}\left[\ln\left(\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left\|\nabla_x \boldsymbol{P}_{ij} - \nabla_x \widetilde{\boldsymbol{T}}_{ij}\right\|_1 + 1\right)\right. \quad (12)$$
$$\left. + \ln\left(\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left\|\nabla_y \boldsymbol{P}_{ij} - \nabla_y \widetilde{\boldsymbol{T}}_{ij}\right\|_1 + 1\right)\right],$$

where ∇ denotes the Sobel operator in $x-$ or $y-$ direction.

3.2 Training and Inference

Full Objective. The full objective functions to optimize $\boldsymbol{\epsilon}_\theta$ as Eq. (13):

$$\min_{\theta}\ \mathcal{L}_{mse} + \lambda \cdot \mathcal{L}_{gdc}, \quad (13)$$

where λ is a hyper-parameter controlling relative importance between different losses.

Appearance Consistency Inference. During the inference stage, previous methods [Yang *et al.*, 2023; Gou *et al.*, 2023; Kim *et al.*, 2024] initiate the inference process by sampling random noise $\boldsymbol{z}_T$ from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. However, $\boldsymbol{z}_T$ has significant initial randomness [Wang *et al.*, 2024], which is not conducive to realistically reconstructing the details of the inpainting object. To address this issue, we design an appearance consistency inference, which initiates the inference process from a posterior Gaussian noise $\boldsymbol{z}_T^p$ and $\boldsymbol{z}_T^{up}$. Thus, during inference, $\boldsymbol{z}_t^p$ and $\boldsymbol{z}_t^{up}$ in Eq. (5) are rewritten as $\boldsymbol{z}_T^p$ and $\boldsymbol{z}_T^{up}$ in Eq. (14):

$$\boldsymbol{z}_T^p = \sqrt{\bar{\alpha}_T}\boldsymbol{z}_0^v + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\ \boldsymbol{z}_T^{up} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^p + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}, \quad (14)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{z}_0^v \in \mathbb{R}^{(H/f)\times(W/f)\times 4} = \mathcal{E}(\boldsymbol{P}_{agn} + \boldsymbol{C}_{un}^w)$, where $\boldsymbol{P}_{agn}$ is clothing-agnostic (inpainting) person image [Lee *et al.*, 2022]. By doing so, the initial randomness of the sampling process can be significantly reduced, thereby further enhancing the model's performance.

4 Experiments

Datasets. Our experiments use VITON-HD [Choi *et al.*, 2021], VITON [Han *et al.*, 2018], and DressCode [Morelli *et al.*, 2022], which are **three** challenging datasets in VTON. **VITON-HD** is a high-resolution dataset with a resolution of 512×384. It comprises 13,679 image groups and is split into a training set with 11,647 groups and a testing set with 2,032 groups. Each group includes a frontal-view woman image, a top clothing image, a semantic map, and a pose heatmap. **VITON** consists of 16,253 image groups with a resolution of 256×192. VITON is split into a training set with 14,221 groups and a testing set with 2,032 groups. **DressCode** is another high-resolution dataset with a resolution of 512×384. It comprises 15,363 image groups and is split into a training set with 12,863 groups and a testing set with 2,500 groups.

Implementation Details. All experiments are performed on a single NVIDIA A100 GPU through PyTorch. For the diffusion model, we follow the configuration of [Gou *et al.*, 2023]. During training, the AdamW optimizer [Loshchilov and Hutter, 2017] is employed with a learning rate of $1e^{-4}$, and the batch size is set to 2 for training over 40 epochs. During inference, we adopt the PLMS sampling method, setting the number of sampling steps to 20 for qualitative analysis. The hyper-parameters are configured as follows: $\gamma = 0.1$ in Eq. (8) and $\lambda = 0.5$ in Eq. (13).

Evaluation Metrics. To facilitate quantitative evaluation, we take paired $(\boldsymbol{P}, \boldsymbol{C})$ in the testing set as inputs, then we employ Structure Similarity (**SSIM**) [Seshadrinathan and Bovik, 2008] and Learned Perceptual Image Patch Similarity (**LPIPS**) [Zhang *et al.*, 2018] to evaluate the structural and perceptual similarity between real and generated images in terms of brightness, contrast, and structure. In addition, we take unpaired $(\boldsymbol{P}, \boldsymbol{C}_{un})$ in the testing set as inputs, then we use Fréchet Inception Distance (**FID**) [Heusel *et al.*, 2017] and Kernel Inception Distance (**KID**) [Bińkowski *et al.*, 2018] to measure distribution discrepancy between real and generated images.

Baseline Methods. To conduct qualitative experiments, we employ **30** state-of-the-art (SOTA) methods, including GAN-based methods: CP-VTON [Wang *et al.*, 2018], Clothflow [Han *et al.*, 2019], CP-VTON+ [Minar *et al.*, 2020], SieveNet [Jandial *et al.*, 2020], VTNFP [Yu *et al.*, 2019], ACGPN [Yang *et al.*, 2020], DCTON [Ge *et al.*, 2021a], PF-AFN [Ge *et al.*, 2021b], ZFlow [Chopra *et al.*, 2021], OVNet [Li *et al.*, 2021], LM-VTON [Liu *et al.*, 2021], DAFlow [Bai *et al.*, 2022], Style-Flow [He *et al.*, 2022], RT-VTON [Yang *et al.*, 2022], Dress Code [Morelli *et al.*, 2022], VITON-HD [Choi *et al.*, 2021], HR-VITON [Lee *et al.*, 2022], CIT [Ren *et al.*, 2023], PL-VTON [Zhang *et al.*, 2023], POVNet [Li *et al.*, 2023], GP-VTON [Xie *et al.*, 2023], USC-PFN [Du *et al.*, 2024], TPD [Yang *et al.*, 2024], and diffusion-based methods:

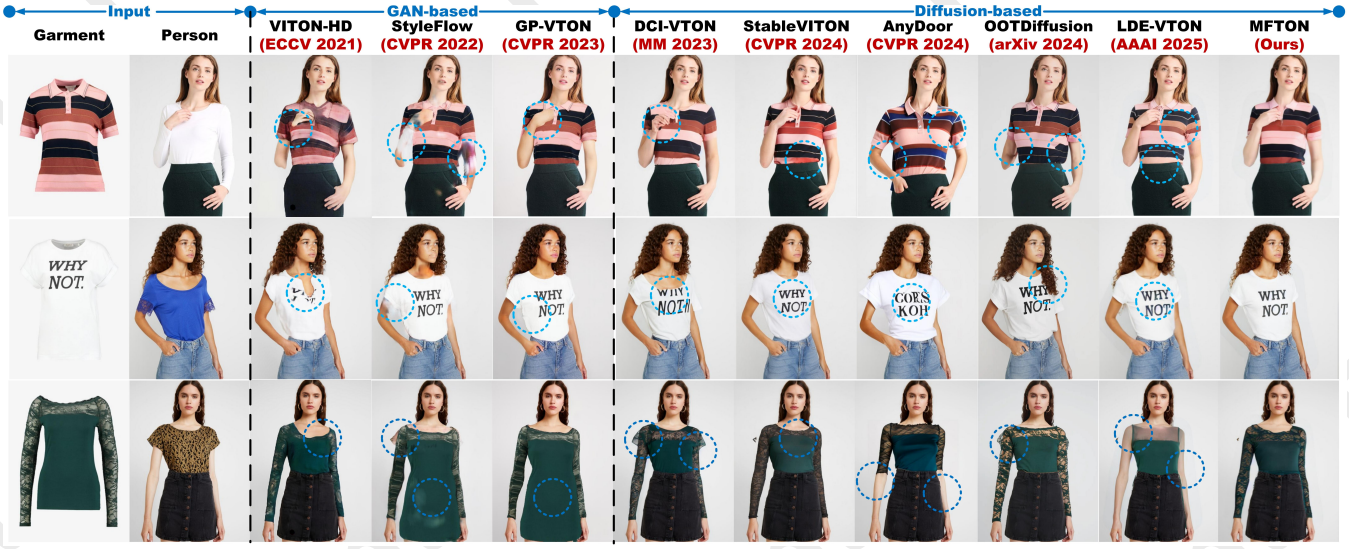

Figure 3: Qualitative results with different baseline methods on the VITON-HD dataset. The baseline methods consist of GAN-based methods and diffusion-based methods. Cyan dashed circles highlight the limitations of each method.

Train / Test				VITON-HD				DressCode Upper			
Methods	**Public.**	**M-F**	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{FID}_{up} \downarrow$	$\text{KID}_{up} \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{FID}_{up} \downarrow$	$\text{KID}_{up} \downarrow$
VITON-HD [Choi *et al.*, 2021]	CVPR'21	✗	0.862	0.117	12.117	3.23	n/a	n/a	n/a	n/a
HR-VITON [Lee *et al.*, 2022]	ECCV'22	✗	0.878	0.105	11.265	2.73	0.936	0.065	13.82	2.71
GP-VTON [Xie *et al.*, 2023]	CVPR'23	✗	0.884	0.081	9.701	1.26	0.769	0.270	20.11	8.17
LaDI-VTON [Morelli *et al.*, 2023]	MM'23	✗	0.864	0.096	9.480	1.99	0.915	0.063	14.26	3.33
PbE [Yang *et al.*, 2023]	CVPR'23	✗	0.802	0.143	11.939	3.85	0.897	0.078	15.33	4.64
DCI-VTON [Gou *et al.*, 2023]	MM'23	✗	0.880	<u>0.080</u>	**8.754**	**1.10**	<u>0.937</u>	<u>0.042</u>	11.92	1.89
StableVITON [Kim *et al.*, 2024]	CVPR'24	✗	0.864	0.084	9.465	1.40	0.911	0.050	<u>11.27</u>	<u>0.72</u>
Anydoor [Chen *et al.*, 2024]	CVPR'24	✗	0.821	0.099	10.850	2.46	0.899	0.119	14.83	3.05
LDE-VTON [Du *et al.*, 2025]	AAAI'25	✗	<u>0.898</u>	0.081	9.640	1.21	n/a	n/a	n/a	n/a
MFTON (Ours)	**This Work**	✓	**0.902**	**0.079**	<u>9.382</u>	<u>1.19</u>	**0.939**	**0.040**	**11.15**	**0.71**

Table 1: Quantitative comparisons on the VITON-HD and DressCode datasets. For LPIPS, FID, and KID, the lower the better. For SSIM, the higher the better. "M-F" denotes whether the mask is used during inference. **Bold** denotes the best result. <u>Underline</u> denotes the second best.

LaDI-VTON [Morelli *et al.*, 2023], PbE [Yang *et al.*, 2023], DCI-VTON [Gou *et al.*, 2023], StableVITON [Kim *et al.*, 2024], OOTDiffusion [Xu *et al.*, 2025], Anydoor [Chen *et al.*, 2024], and LDE-VTON [Du *et al.*, 2025], as baselines for quantitative evaluation and select several publicly available methods for qualitative evaluation.

4.1 Comparison with SOTA Methods

We qualitatively and quantitatively compare our proposed MFTON with several SOTA baseline methods.

Qualitative Results. The qualitative comparisons are illustrated in Figs. 3, 4, and 5. It can be observed that SOTA GAN-based methods such as GP-VTON [Xie *et al.*, 2023] and USC-PFN [Du *et al.*, 2024] are only capable of narrowly fulfilling the requirements of try-on under simple poses. Nevertheless, the garment regions in their results are consistently overly smooth, lacking natural clothing wrinkles, and exhibit severe occlusion and distortion issues in complex poses. Our adoption of a diffusion-based generator effectively addresses the issue of fabric wrinkles. However, diffusion-based methods tend to lose some details of the target clothing and partial identity information to varying degrees, due to insufficient structural supervision of the clothing and the lack of rigorous body structure cues from the unreliable inpainting mask. As shown in the 1-*st* row of Fig. 3, DCI-VTON [Gou *et al.*, 2023], AnyDoor [Chen *et al.*, 2024], and OOTDiffusion [Xu *et al.*, 2025] almost lose the information of the right arm. In the 2-*nd* row, the alignment of the text is not natural, and AnyDoor suffers from text information loss. Unlike these, our method initially eliminates the potential negative impact of the inpainting mask by using the designed mask-free strategy, thus ensuring that information such as arms and hair are completely preserved. Furthermore, we enhance the supervision of clothing texture and structural information through the proposed clothing texture-aware attention and geometric detail capture loss $\mathcal{L}_{gdc}$, to guarantee that the global and local key details of the clothing, including color, texture, patterns, and characters, are faithfully reconstructed. Overall, compared to the baseline method, our approach yields more realistic visual results, which holds significant implications for practical applications.

Figure 4: Qualitative results with different baseline methods on the VITON dataset. The results consist of easy samples and hard samples. Cyan dashed circles highlight the limitations of each method.

Methods	Publication	Mask-Free	$SSIM_p \uparrow$	$FID_{up} \downarrow$
CP-VTON	ECCV'18	✗	0.72	24.45
VTNFP	ICCV'19	✗	0.80	n/a
Cloth-flow	CVPR'19	✗	0.84	14.43
CP-VTON+	CVPRW'20	✗	0.75	21.04
SieveNet	WACV'20	✗	0.77	n/a
ACGPN	CVPR'20	✗	0.84	16.64
LM-VTON	AAAI'21	✗	0.85	17.18
DCTON	CVPR'21	✗	0.83	14.82
ZFlow	ICCV'21	✗	0.88	15.17
OVNet	CVPR'21	✗	0.85	15.78
PF-AFN	CVPR'21	✓	0.89	10.21
RT-VTON	CVPR'22	✗	n/a	11.66
DAFlow	ECCV'22	✗	0.88	12.05
Dress Code	CVPR'22	✗	0.89	13.71
CIT	TMM'23	✗	0.83	13.97
PL-VTON	TMM'23	✗	0.87	10.96
POVNet	TPAMI'23	✗	0.89	13.37
PbE	CVPR'23	✗	0.83	12.56
USC-PFN	NeurIPS'23	✓	<u>0.91</u>	10.47
TPD	CVPR'24	✗	0.89	<u>9.58</u>
LDE-VTON	AAAI'25	✗	<u>0.91</u>	9.86
MFTON	**This Work**	✓	**0.91**	**9.42**

Table 2: Quantitative comparisons on the VITON dataset. "Mask-Free" denotes whether the mask is used during inference.

Figure 5: Visualization results with different baseline methods on the DressCode dataset. Zooming in for more details.

Different Configuration	$SSIM_p \uparrow$	$FID_{up} \downarrow$
(B) Conventional Inpainting-based LDM	0.802	11.939
+ **(P)** Consistency Inference	0.857	9.912
+ **(M)** Mask-Free Strategy	0.882	9.694
+ **(G)** Geometric Detail Capture	0.895	9.570
+ **(T)** Texture-aware Attention	**0.902**	**9.382**

Table 3: Ablation studies of the proposed different components.

Quantitative Results. Tab. 1 (note that the majority of the methods listed in the Tab. have also utilized only SSIM and FID metrics for comparison in their paper) and Tab. 2 present the quantitative results on the VITON-HD, VITON, and Dresscode datasets, respectively. On the VITON dataset, our method achieves the same SSIM as the SOTA diffusion-based LDE-VTON [Du *et al.*, 2025], a plateau arises because the diffusion process inherently leads to some loss of original information. However, our method excels in FID, outperforming all other approaches. On the VITON-HD and Dresscode, our method outperforms the SOTA GAN-based methods, GP-VTON [Xie *et al.*, 2023] and USC-PFN [Du *et al.*, 2024], and the SOTA diffusion-based methods, Anydoor [Chen *et al.*, 2024] and LDE-VTON [Du *et al.*, 2025]. However, in terms of FID and KID metrics, our performance is slightly inferior to DCI-VTON [Gou *et al.*, 2023]. This is attributed to the fact that DCI-VTON retains excessive identity and fabric residues that significantly deviate from the realistic effect of VTON. For instance, the 2-nd row and 6-th column of Fig. 3, completely replicates the neck area from the original person image, which substantially reduces the FID and KID scores. Nevertheless, our method outperforms other methods in all four metrics.

4.2 Ablation Studies

We perform ablation experiments to verify the effectiveness of the proposed different components.

Effectiveness of Appearance Consistency Inference. In Fig. 6, **(B)** and **(P)** demonstrate the effectiveness of our proposed clothing appearance consistency inference. When it is removed from the baseline method **(B)**, the clothing loses a significant amount of detail, retaining only the global information. Upon its reintroduction **(P)**, a plethora of clothing

Figure 6: Visual ablation studies of different components in our approach. Zooming in for more details.

Figure 7: Analysis of sample steps. We set the total sample steps as 20 to balance quality and speed for conventional visualization.

details are restored, a fact corroborated by the quantitative results in Tab. 3. Additionally, by substituting flat clothing for warped clothing as input (U), it is apparent that within our designed framework, the warped clothing serves as essential conditional information.

Effectiveness of Mask-Free Strategy. In Fig. 6, (C) and (M) demonstrate the effectiveness of our proposed mask-free strategy. When the inpainting mask is reintroduced into the input conditions (C), the generated body layout is adversely affected by the mask, resulting in unrealistic outcomes. Tab. 3 substantiates the positive contribution of mask removal.

Effectiveness of Geometric Detail Capture. In Fig. 6, (G) demonstrates the effectiveness of our proposed geometric detail capture loss $\mathcal{L}_{gdc}$. This loss effectively captures geometric information on the clothing, such as character and stripe patterns, to guide the generator in faithfully reconstructing the clothing's signature characteristics. Tab. 3 highlights the improvements attributed to this loss function.

Effectiveness of Clothing Texture-aware Attention. In Fig. 6, (T) demonstrates the effectiveness of our proposed clothing texture-aware attention. When we introduce the proposed clothing texture-aware attention mechanism, the texture information of the clothing is more thoroughly complemented, resulting in visual outcomes that are more authentic and natural. This is amply demonstrated in Tab. 3.

Analysis of Sampling Steps. In addition, we analyze the sampling steps, as detailed in Fig. 7. We set the total sample steps as 20 to balance quality and speed for conventional visualization (not for quantitative experimental results).

5 Conclusion

In this work, we present a novel mask-free virtual try-on framework, MFTON, which can produce photo-realistic results without using any inpainting masks and information as the denoising condition by using the proposed mask-free strategy. Moreover, for clothing regions, we propose a clothing texture-aware attention mechanism to enable the model to focus on texture generation with significant visual differences. We then introduce a geometric detail capture loss to further enable the model to capture more high-frequency information. Finally, we propose an appearance consistency inference method to reduce the initial randomness of the sampling process significantly. Extensive experiments show that our method outperforms the existing virtual try-on methods.

Limitations. Although our mask-free method has achieved good visualization results, however, currently available datasets still have inevitable limitations. In the real world, preferences for clothing among users from different regions vary with local customs, yet the diversity of currently popular datasets is limited. Therefore, acquiring a model capable of virtual try-ons for various styles of clothing based on such datasets poses a significant challenge.

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), the Huawei Kunpeng-Ascend Innovation Incentive Programme.

References

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

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[Chen *et al.*, 2024] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024.

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

[Chopra *et al.*, 2021] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442, 2021.

[Du *et al.*, 2024] Chenghu Du, Shuqing Liu, Shengwu Xiong, et al. Greatness in simplicity: Unified self-cycle consistency for parser-free virtual try-on. *Advances in Neural Information Processing Systems*, 36, 2024.

[Du *et al.*, 2025] Chenghu Du, Junyin Wang, Feng Yu, and Shengwu Xiong. Latent diffusion-enhanced virtual try-on via optimized pseudo-label generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2780–2788, 2025.

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

[Ge *et al.*, 2021b] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.

[Gou *et al.*, 2023] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 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*, pages 7599–7607, 2023.

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

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

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

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

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jandial *et al.*, 2020] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. Sievenet: A unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2182–2190, 2020.

[Kim *et al.*, 2024] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 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*, pages 8176–8185, 2024.

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

[Li *et al.*, 2021] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15546–15555, 2021.

[Li *et al.*, 2023] Kedan Li, Jeffrey Zhang, and David Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12222–12235, 2023.

[Liu *et al.*, 2021] Guoqiang Liu, Dan Song, Ruofeng Tong, and Min Tang. Toward realistic virtual try-on through landmark guided shape matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2118–2126, 2021.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Minar *et al.*, 2020] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, volume 3, pages 10–14, 2020.

[Morelli *et al.*, 2022] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022.

[Morelli *et al.*, 2023] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ren *et al.*, 2023] Bin Ren, Hao Tang, Fanyang Meng, Ding Runwei, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–20, 2023.

[Seshadrinathan and Bovik, 2008] Kalpana Seshadrinathan and Alan C Bovik. Unifying analysis of full reference image quality assessment. In *2008 15th IEEE International Conference on Image Processing*, pages 1200–1203. IEEE, 2008.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

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

[Wang *et al.*, 2024] Chenhui Wang, Tao Chen, Zhihao Chen, Zhizhong Huang, Taoran Jiang, Qi Wang, and Hongming Shan. Fldm-vton: Faithful latent diffusion model for virtual try-on. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1362–1370. International Joint Conferences on Artificial Intelligence Organization, 8 2024.

[Xie *et al.*, 2023] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 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)*, pages 23550–23559, June 2023.

[Xu *et al.*, 2025] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8996–9004, 2025.

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

[Yang *et al.*, 2022] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2022.

[Yang *et al.*, 2023] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[Yang *et al.*, 2024] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024.

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

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

[Zhang *et al.*, 2023] Shengping Zhang, Xiaoyu Han, Weigang Zhang, Xiangyuan Lan, Hongxun Yao, and Qingming Huang. Limb-aware virtual try-on network with progressive clothing warping. *IEEE Transactions on Multimedia*, pages 1–16, 2023.

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