LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On

Davide Morelli*
University of Modena and Reggio Emilia
Modena, Italy
davide.morelli@unimore.it

Alberto Baldrati*
University of Florence
Florence, Italy
alberto.baldrati@unifi.it

Giuseppe Cartella
University of Modena and Reggio Emilia
Modena, Italy
giuseppe.cartella@unimore.it

Marcella Cornia
University of Modena and Reggio Emilia
Modena, Italy
marcella.cornia@unimore.it

Marco Bertini
University of Florence
Florence, Italy
marco.bertini@unifi.it

Rita Cucchiara
University of Modena and Reggio Emilia
Modena, Italy
rita.cucchiara@unimore.it

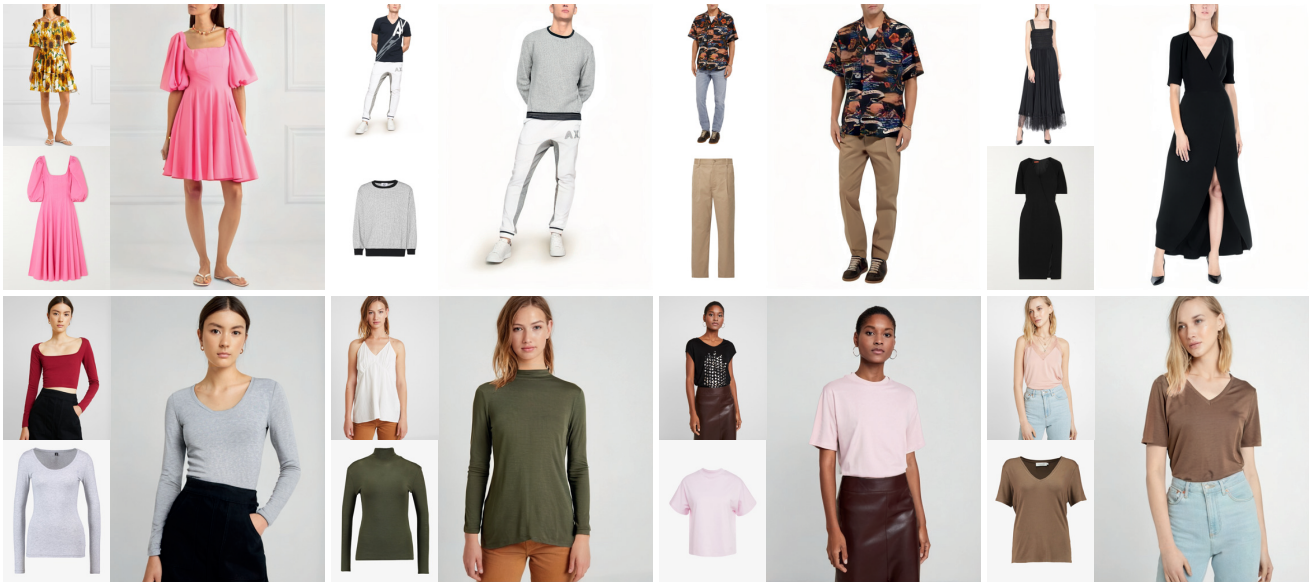

Figure 1: Images generated by the proposed LaDI-VTON model, given an input target model and a try-on clothing item from both Dress Code [44] (1st row) and VITON-HD [9] (2nd row) datasets.

ABSTRACT

The rapidly evolving fields of e-commerce and metaverse continue to seek innovative approaches to enhance the consumer experience. At the same time, recent advancements in the development of diffusion models have enabled generative networks to create remarkably realistic images. In this context, image-based virtual try-on, which consists in generating a novel image of a target model wearing a given in-shop garment, has yet to capitalize on the potential of these powerful generative solutions. This work introduces LaDI-VTON, the first Latent Diffusion textual Inversion-enhanced model for the Virtual Try-ON task. The proposed architecture relies on a latent diffusion model extended with a novel additional autoencoder module that exploits learnable skip connections to enhance the generation process preserving the model's characteristics. To effectively maintain the texture and details of the in-shop garment, we propose a textual inversion component that can map the visual features of the garment to the CLIP token embedding space and thus generate a set of pseudo-word token embeddings capable of conditioning the generation process. Experimental results on Dress Code and VITON-HD datasets demonstrate that our approach outperforms the competitors by a consistent margin, achieving a significant milestone for the task. Source code and trained models are publicly available at: https://github.com/miccunifi/ladi-vton.

*Both authors contributed equally to this research.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612137

CCS CONCEPTS

• **Computing methodologies** → *Computer vision tasks*; *Computer vision.*

KEYWORDS

Virtual Try-On, Latent Diffusion Models, Generative Architectures.

ACM Reference Format:
Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612137

1 INTRODUCTION

The disruptive success of e-commerce and online shopping is steadily demanding a more streamlined and enjoyable customer shopping experience, from personalized garment recommendation [10, 12, 31, 55] to visual product search [5, 24, 38, 43, 64]. Given the large availability of online images as accessories, garments, and other related products, Computer Vision and Multimedia research play a crucial role by offering valuable tools for a more personalized user experience. Among them, image-based virtual try-on has recently attracted significant interest in the research community with the introduction of several architectures [26, 44, 60] that, given an image of a person and a garment taken from a catalog, allow to dress the person with the given try-on garment.

The generation process carried out by current state-of-the-art methods for the task [3, 21, 36, 44] entirely relies on Generative Adversarial Networks (GANs) [22]. During the last years, a new family of generative architectures, namely diffusion models [29, 57], have shown superior image generation quality compared to GANs [14], also with a more stable training procedure. However, considering the high computational demand typical of diffusion models, Rombach *et al.* [50] have recently tackled the problem by introducing a latent-based version that works in the latent space of a pre-trained autoencoder, thus finding the best trade-off between computational load and image quality.

Motivated by the tremendous success of these generative models, in this work, we introduce and explore for the first time an image-based virtual try-on method based on Latent Diffusion Models (LDMs) [50], demonstrating their successful possible applications in this field. We design a novel diffusion-based architecture conditioned on the target in-shop garment and human keypoints to keep the model's body pose unchanged. To preserve the target garment texture in the generation process, we propose to augment LDMs with a textual inversion network able to map the visual features of the in-shop garment to the CLIP textual token embedding space [48]. We then condition the LDM generation through the cross-attention mechanism using the predicted tokens embeddings.

While LDMs can generate highly realistic images, one of their drawbacks is that they struggle when dealing with high-frequency details in the pixel space. This problem stems from the spatial compression performed by the autoencoder, which gives access to a lower-dimensional latent space where high-frequency details may not be accurately represented [50]. In our setting, this can lead to details loss in the final generated images, especially when handling the model's hands, feet, and face. To address this issue, we introduce the Enhanced Mask-Aware Skip Connection (EMASC) module, a learnable skip connection that transfers the details from the encoding phase to the corresponding decoding one, improving the autoencoder reconstruction capabilities.

We extensively validate our architecture on two widely-used virtual try-on benchmarks (*i.e.*, Dress Code [44] and VITON-HD [9]), demonstrating superior quantitative and qualitative performance than state-of-the-art methods and showing that diffusion models applied to the virtual try-on field can achieve higher realism than GAN-based counterparts (Figure 1).

Contributions. To sum up, our contributions are as follows:

- We employ LDMs to solve the task of image-based virtual try-on, an approach that, to the best of our knowledge, has never been previously explored in this field.
- To reduce the reconstruction error of LDMs, we enhance the autoencoder with learnable skip connections, enabling the preservation of details outside the inpainting region.
- Additionally, to increase detail retention of the generation process, we define a forward-only textual inversion module to further condition the model on the input try-on garment without losing texture information.
- Extensive experiments validate the effectiveness of each component of our architecture, which achieves state-of-the-art results on two widely used benchmarks for the task. We believe our results can highlight how virtual try-on can strongly benefit from using LDMs and serve as a starting point for future research in the field.

2 RELATED WORK

Image-Based Virtual Try-On. Image-based virtual try-on [3, 9, 18, 26, 32, 44, 60] aims to transfer a desired garment onto the corresponding region of a target subject while preserving human pose and identity. One of the pioneering works in this field is VITON [26], a framework composed of an encoder-decoder generator that produces a coarse result further improved by a refinement network that exploits the warped clothing item obtained through a TPS transformation [15]. Some follow-up works have been oriented towards the enhancement of the warping module. Wang *et al.* [60] proposed a learnable TPS module to mitigate the problem of clothing details preservation, which has subsequently been improved either by combining TPS with affine transformations [19, 37] or taking into account generated semantic layouts [66] and body information [17].

Another research line focuses on the generation phase and refinement of the result [21, 32, 44]. Issenhuth *et al.* [32], for example, presented a distillation-based teacher-student architecture that does not leverage a predicted semantic layout during the generation. This idea has further been explored in [21] with the introduction of an additional tutor knowledge module to improve the generation quality. Differently, Morelli *et al.* [44] focused on the semantics of the generated results and proposed a semantic-aware discriminator working at the pixel level instead of the image or patch level. Lee *et al.* [36] solved the misalignment problem by designing a unified pipeline that combines the warping and segmentation stages to achieve better high-resolution results.

A common aspect linking all current methods is that the generation phase relies on GANs [22]. Driven by the enormous success of diffusion models [29] in different fields, we are the first, to the best of our knowledge, to propose an image-based virtual try-on architecture entirely relying on the aforesaid generative models.

Diffusion Models. A fundamental line of research in the image synthesis field is the one marked by diffusion models [14, 29, 30, 45, 57, 58]. Inspired by non-equilibrium statistical physics, Sohl-Dickstein *et al.* [57] defined a tractable generative model of data distribution by iteratively destroying the data structure through a forward diffusion process and then reconstructing with a learned reverse diffusion process. Some years later, Ho *et al.* [29] successfully demonstrated that this process is applicable to generate high-quality images. Nichol *et al.* [45] further improved the work presented in [29] by learning the variance parameter of the reverse diffusion process and generating the output with fewer forward passes without sacrificing sample quality. While these methods work in the pixel space, Rombach *et al.* [50] proposed a variant working in the latent space of a pre-trained autoencoder, enabling higher computational efficiency.

The impact of diffusion models has rapidly become disruptive in diverse tasks such as text-to-image synthesis [23, 46, 49, 54], image-to-image translation [53, 61, 68], image editing [2, 41, 65], and inpainting [40, 46]. Strictly related to virtual try-on is the task of human image generation, where pose preservation is often a strict constraint. On this line, Jiang *et al.* [33] focused on synthesizing full-body images given human pose and textual descriptions of shapes and textures of clothes, generating the output via sampling from a learned texture-aware codebook. Bhunia *et al.* [7] tackled the task of pose-guided human generation by developing a texture diffusion block based on cross attention and conditioned on multi-scale texture patterns from the encoded source image. Baldrati *et al.* [6], instead, proposed to guide the generation process constraining a latent diffusion model with the model pose, the garment sketch, and a textual description of the garment itself.

Textual Inversion. Textual inversion is a recent technique proposed in [20] to learn a pseudo word in the embedding space of the text encoder starting from visual concepts. Following [20], several promising methods [11, 25, 42, 52] have been designed to enable personalized image generation and editing. Ruiz *et al.* [52] presented a fine-tuning technique to bind an identifier with a subject represented by a few images and adopted a class-specific prior preservation loss to mitigate language drift. Similarly, Kumari *et al.* [35] proposed a different fine-tuning method to enable multi-concept composition and showed that updating only a small subset of model weights is sufficient to integrate new concepts. On a different line, Han *et al.* [25] decomposed the CLIP embedding space [48] based on semantics and enabled image manipulation without requiring any additional fine-tuning.

3 PROPOSED METHOD

While most of the existing virtual try-on approaches leverage generative adversarial networks [26, 32, 44, 60], we propose a novel solution based, for the first time, on Latent Diffusion Models (LDMs). In particular, our work employs the Stable Diffusion architecture [50] as a starting point to perform the virtual try-on task. To augment the text-to-image model with try-on capabilities, we modify the architecture to take as input both the try-on garment and the pose information of the target model. In addition, to better preserve the input clothing item details, we propose to add a novel forward-only textual inversion technique during the generation process. Finally, we enhance the image reconstruction autoencoder of Stable Diffusion with masked skip connections, thus improving the quality of generated images and better preserving the fine-grained details of the original model image. Figure 2 depicts an overview of the proposed model.

3.1 Preliminaries

Stable Diffusion. It consists of an autoencoder $\mathcal{A}$ with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, a text time-conditional U-Net denoising model ϵ_θ, and a CLIP text encoder T_E, which takes text Y as input. The encoder $\mathcal{E}$ compresses an image $I \in \mathbb{R}^{3\times H\times W}$ into a lower-dimensional latent space in $\mathbb{R}^{4\times h\times w}$, where $h = \frac{H}{8}$ and $w = \frac{W}{8}$, while the decoder $\mathcal{D}$ performs the inverse operation and decodes a latent variable into the pixel space. For clarity, we refer to the ϵ_θ convolutional input as the spatial input γ (*e.g.*, z_t) since convolutions preserve the spatial structure, and to the attention conditioning input as ψ (*e.g.*, $[t, T_E(Y)]$). The training of the denoising network ϵ_θ is performed by minimizing the following loss function:

$$L = \mathbb{E}_{\mathcal{E}(I),Y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\gamma,\psi)\|_2^2\right], \tag{1}$$

where t represents the diffusing time step, $\gamma = z_t$, z_t is the encoded image $\mathcal{E}(I)$ where we stochastically add Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$, and $\psi = [t; T_E(Y)]$.

We aim to generate a new image $\tilde{I}$ that replaces a target garment in the model input image I with an in-shop garment C provided by the user while retaining the model's physical characteristics, pose, and identity. This task can be seen as a particular type of inpainting, specialized in replacing garment information in human-based images according to a target garment image provided by the user. For this reason, we use the Stable Diffusion inpainting pipeline as the starting point of our approach. It takes as spatial input γ the channel-wise concatenation of an encoded masked image $\mathcal{E}(I_M)$, a resized binary inpainting mask $m \in \{0,1\}^{1\times h\times w}$, and the denoising network input z_t. Specifically, I_M is the model image I masked according to the inpainting mask $M \in \{0,1\}^{1\times H\times W}$, and the binary inpainting mask m is the resized version according to the latent space spatial dimension of the original inpainting mask M. To summarize, the spatial input of the inpainting denoising network is $\gamma = [z_t; m; \mathcal{E}(I_M)] \in \mathbb{R}^{(4+1+4)\times h\times w}$.

CLIP. It is a vision-language model [48] which aligns visual and textual inputs in a shared embedding space. In particular, CLIP consists of a visual encoder V_E and a text encoder T_E that extract feature representations $V_E(I) \in \mathbb{R}^d$ and $T_E(E_L(Y)) \in \mathbb{R}^d$ for an input image I and its corresponding text caption Y, respectively. Here, d is the size of the CLIP embedding space, and E_L is the embedding lookup layer which maps each Y tokenized word to the token embedding space $\mathcal{W}$.

The proposed approach introduces a novel textual inversion technique to generate a representation of the in-shop garment C. We feed this representation to the CLIP text encoder and use it to condition the diffusion process. It consists in mapping the visual

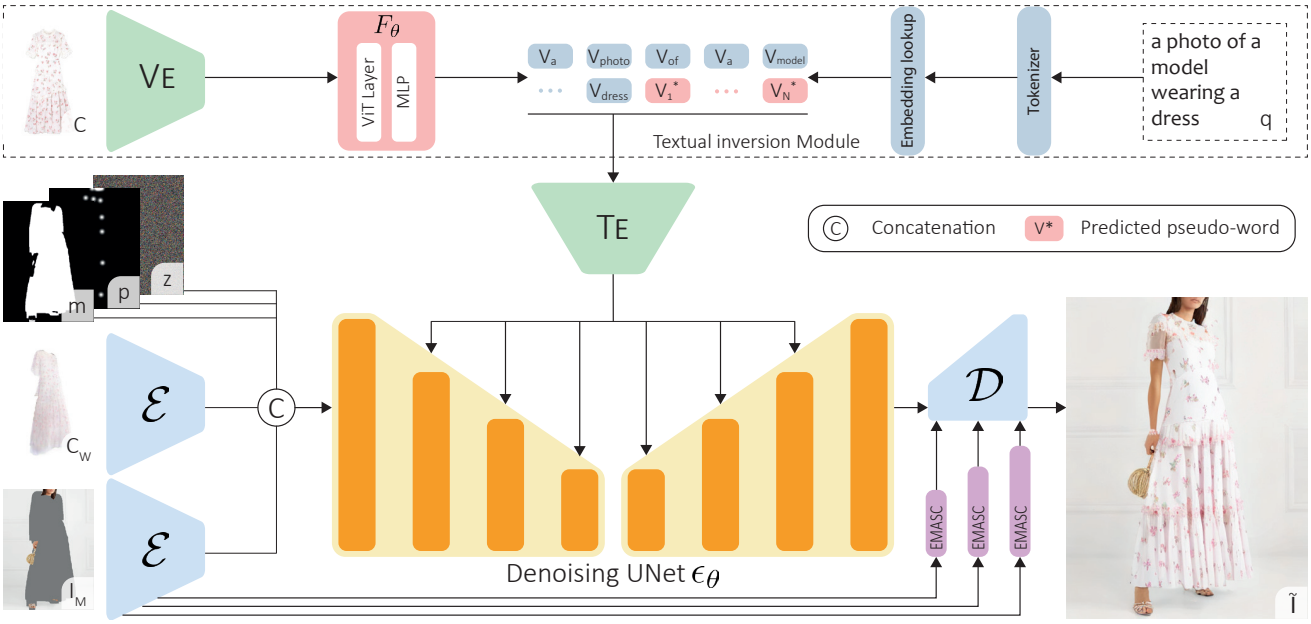

Figure 2: Overview of the proposed LaDI-VTON model. On the top, the textual inversion module generates a representation of the in-shop garment. This information conditions the Stable Diffusion model along with other convolutional inputs. The decoder $\mathcal{D}$ is enriched with the Enhanced Mask-Aware Skip Connection (EMASC) modules to reduce the reconstruction error, improving the high-frequency details in the final image.

features of C into a set of N new token embeddings $V_n^* \in \mathcal{W}, n = \{1, \ldots, N\}$. Following the terminology introduced in [4], we refer to these embeddings as Pseudo-word Tokens Embeddings (PTEs) since they do not correspond to any linguistically meaningful entity but rather are a representation of the in-shop garment visual features in the token embedding space $\mathcal{W}$.

3.2 Textual-Inversion Enhanced Virtual Try-On

To tackle the virtual try-on task, we propose injecting in the Stable Diffusion textual conditioning branch additional information from the target garment C extracted through textual inversion. In particular, starting from the features of the in-shop garment C extracted from the CLIP visual encoder, we learn a textual inversion adapter F_θ to predict a set of fine-grained PTEs describing the in-shop garment C itself. These PTEs lie in the CLIP token embedding space $\mathcal{W}$ and thus can be used as an additional conditioning signal.

We also propose to extend the Stable Diffusion inpainting pipeline to accept the model pose map $P \in \mathbb{R}^{18\times H\times W}$, where each channel is associated with a human keypoint, and the warped in-shop garment $C_W \in \mathbb{R}^{3\times H\times W}$, representing the target garment C warped according to the model body pose. While the pose map P enables the method to preserve the original human pose of the model I, the warped garment C_W helps the generation process to properly fit the garment onto the model.

Data Preparation. The warped garment C_W is obtained by training a module that warps the in-shop garment C fitting the model body shape in I. We employ the geometric matching module proposed in [60] and refine the results with a U-Net-based component [51]. The virtual try-on task involves replacing one or more garments the target model is wearing. With this aim, we define the inpainting area determined by the mask M to fully encompass the target garment. We adopt the method proposed in previous works such as [32, 44] to ensure the mask completely covers the target garment.

Textual Inversion. Given the in-shop image C, the aim of the textual inversion adapter F_θ is to predict a set of pseudo-word token embeddings $\{V_1^*, \ldots, V_N^*\}$ able to well represent the image C in the CLIP token embedding space $\mathcal{W}$. We then use the predicted PTEs to condition the Stable Diffusion denoising network ϵ_θ and obtain the final image $\tilde{I}$ where the model in I is wearing the garment in C. For clarity, we intend that a set of PTEs represent well a target image if a Stable Diffusion model conditioned on the concatenation of a generic prompt and the predicted pseudo-words can reconstruct the target image itself.

We first build a textual prompt q that guides the diffusion process to perform the virtual try-on task, tokenize it and map each token into the token embedding space using the CLIP embedding lookup module, obtaining V_q. Then, we encode the image C using the CLIP visual encoder V_E and feed the features extracted from the last hidden layer to the textual inversion adapter F_θ, which maps the input visual features to the CLIP token embedding space $\mathcal{W}$. We then concatenate the prompt embedding vectors with the predicted pseudo-word token embeddings as follows:

$$\hat{Y} = \text{Concat}(V_q, F_\theta(V_E(C))). \tag{2}$$

We feed the embedded concatenation $\hat{Y}$ to the CLIP text encoder T_E and use the output to condition the denoising network ϵ_θ leveraging the existing Stable Diffusion textual cross-attention.

To train the textual inversion adapter F_θ, we use the inpainting pipeline of the out-of-the-box Stable Diffusion model as ϵ_θ. Specifically, it takes as input the encoded masked target model $\mathcal{E}(I_M)$, the

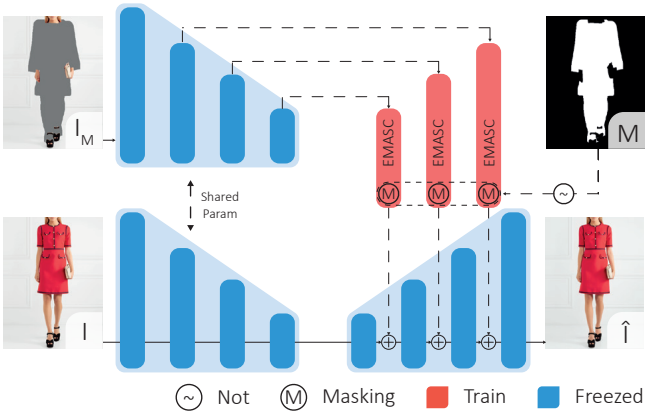

Figure 3: Overview of the proposed autoencoder with Enhanced Mask-Aware Skip Connection (EMASC) modules.

inpainting mask M, and the latent variable z. When training the adapter F_θ, we freeze all the other model parameters.

To the best of our knowledge, this study marks the first instance in which a textual inversion approach has been employed in the domain of virtual try-on. As shown in the experimental section, this innovative conditioning methodology can significantly strengthen the final results and contribute to preserving the details and texture of the original in-shop garment. Note that our proposed approach differs from traditional textual inversion techniques [20, 35, 52]. Rather than directly optimizing the pseudo-word token embeddings through iterative methods, in our solution, the adapter F_θ is trained to generate these embeddings in a single forward pass.

Diffusion Virtual Try-On Model. To perform the complete virtual try-on task, we employ the additional inputs described above (*i.e.*, textual-inverted information $\hat{Y}$ of the in-shop garment, the pose map P, and the garment fitted to the model body shape C_W) to condition the Stable Diffusion inpainting pipeline. In particular, we extend the spatial input $\gamma \in \mathbb{R}^{9\times h\times w}$ of the denoising network ϵ_θ concatenating it with the resized pose map $p \in \mathbb{R}^{18\times h\times w}$ and the encoded warped garment $\mathcal{E}(C_W) \in \mathbb{R}^{4\times h\times w}$. The final spatial input results in $\gamma = [z_t; m; \mathcal{E}(I_M); p; \mathcal{E}(C_W)] \in \mathbb{R}^{(9+18+4)\times h\times w}$.

To enrich the input capacity of the denoising network ϵ_θ without needing to retrain it from scratch [6, 50], we propose to extend the kernel channels of the first convolutional layer by adding zero-initialized weights to match the new input channel dimension. In such a way, we can retain the knowledge embedded in the original denoising network while allowing the model to deal with the newly proposed inputs. Since the warped garment C_W is not always able to properly represent the contextualization of the in-shop garment with the target model information, we also modify the Stable Diffusion textual input by using $\hat{Y}$ obtained from the output of the trained textual inversion adapter F_θ as described in Eq. 2.

As in standard LDMs, we train the proposed denoising network to predict the noise stochastically added to an encoded input $z_t = \mathcal{E}(I)$. We specify the corresponding objective function as:

$$L = \mathbb{E}_{\mathcal{E}(I),\hat{Y},\epsilon\sim\mathcal{N}(0,1),t,\mathcal{E}(I_M),M,p,\mathcal{E}(C_W)}\left[\|\epsilon - \epsilon_\theta(\gamma,\psi)\|_2^2\right], \quad (3)$$

where $\psi = \left[t; T_E(\hat{Y})\right]$.

3.3 Enhanced Mask-Aware Skip Connections

The autoencoder $\mathcal{A}$ of LDMs enables the denoising network ϵ_θ to work within a latent space smaller than the pixel space. Compared to standard diffusion networks, this behavior is essential to reduce the parameters ϵ_θ of the latent diffusion denoising network allowing it to reach the best trade-off between image quality and computational load [50]. We remind that given an image $I \in \mathbb{R}^{3\times H\times W}$, the Stable Diffusion encoder $\mathcal{E}$ compresses it in a latent space $Z \in \mathbb{R}^{4\times\frac{H}{8}\times\frac{W}{8}}$, resulting in a total compression of 48×. However, this trade-off comes at a cost especially when dealing with human images and small high-frequency details such as hands, feet, and faces. We argue that the autoencoder reconstruction error partially depends on the data loss deriving from the latent space compression.

To address the problem, we propose to extend the autoencoder architecture with an Enhanced Mask-Aware Skip Connection (EMASC) module whose aim is to learn to propagate relevant information from different layers of the encoder $\mathcal{E}$ to corresponding ones of the decoder $\mathcal{D}$. In particular, instead of skipping the information of the encoded image I to reconstruct, we pass to the EMASC modules the intermediate features of the masked image I_M encoding process, using the encoder $\mathcal{E}$. This procedure allows only the features not modified in the inpainting task to percolate, keeping the process cloth agnostic. We implement EMASC employing additive non-linear learned skip connections in which we mask the output according to the inverted inpainting mask. Since the EMASC inputs are the intermediate features of the masked model I_M encoding process, masking the EMASC output features helps avoid propagating the masked regions through the skip connections. Formally, the EMASC module is defined as follows:

$$\begin{aligned} EMASC_i &= f(E_i) * NOT(m_i) \\ D_i &= D_{i-1} + EMASC_i \end{aligned} \quad (4)$$

where f is a learned non-linear function, E_i is the i-th feature map coming from the encoder $\mathcal{E}$, D_i is the corresponding i-th decoder feature map, and m_i is obtained by resizing the mask M according to the E_i spatial dimension. An overview of the proposed autoencoder enhanced with EMASC modules is reported in Figure 3.

Notice that the EMASC modules only depend on the Stable Diffusion denoising autoencoder, and once trained, they can be easily added to the standard Stable Diffusion pipeline in a plug-and-play manner without requiring additional training. We show that this simple proposed modification can reduce the compression information loss in the inpainting task, resulting in better high-frequency human-related reconstructed details.

4 EXPERIMENTAL EVALUATION

4.1 Datasets and Evaluation Metrics

We perform experiments on two virtual try-on datasets, namely Dress Code [44] and VITON-HD [9], that feature high-resolution image pairs of in-shop garments and model images in both paired and unpaired settings. While in the paired setting the in-shop garment is the same as the model is wearing, in the unpaired one, a different garment is selected for the virtual try-on task.

The Dress Code dataset [44] features over 53,000 image pairs of clothes and human models wearing them. The dataset includes high-resolution images (*i.e.*, 1024×768) and garments belonging to

Table 1: Quantitative results on the Dress Code dataset [44]. The * marker indicates results reported in previous works, which may differ in terms of metric implementation. Best results are reported in bold.

	Upper-body		Lower-body		Dresses		All					
Model	**FID$_u$** ↓	**KID$_u$** ↓	**FID$_u$** ↓	**KID$_u$** ↓	**FID$_u$** ↓	**KID$_u$** ↓	**LPIPS** ↓	**SSIM** ↑	**FID$_p$** ↓	**KID$_p$** ↓	**FID$_u$** ↓	**KID$_u$** ↓
PF-AFN* [21]	14.32	-	18.32	-	13.59	-	-	-	-	-	-	-
HR-VITON* [36]	16.86	-	22.81	-	16.12	-	-	-	-	-	-	-
CP-VTON [60]	48.31	35.25	51.29	38.48	25.94	15.81	0.186	0.842	28.44	21.96	31.19	25.17
CP-VTON† [60]	22.18	12.09	18.85	10.24	21.83	12.31	0.095	0.898	12.90	9.81	13.77	10.12
PSAD [44]	17.51	7.15	19.68	8.90	17.07	6.66	**0.058**	**0.918**	8.01	4.90	10.61	6.17
LaDI-VTON	**13.26**	**2.67**	**14.80**	**3.13**	**13.40**	**2.50**	0.064	0.906	**4.14**	**1.21**	**6.48**	**2.20**

different macro-categories, such as upper-body clothes, lower-body clothes, and dresses. In our experiments, we employ the original splits of the dataset where 5,400 image pairs (1,800 for each category) compose the test set and the rest the training one. The VITON-HD dataset [9] instead comprises 13,679 image pairs, each composed of a frontal-view woman and an upper-body clothing item with a resolution equal to 1024×768. The dataset is divided into training and test sets of 11,647 and 2,032 pairs, respectively.

To quantitatively evaluate our model, we employ evaluation metrics to estimate the coherence and realism of the generation. In particular, we use the Learned Perceptual Image Patch Similarity (LPIPS) [67] and the Structural Similarity (SSIM) [62] to evaluate the coherence of the generated image compared to the ground-truth. We compute these metrics on the paired setting of both datasets. To measure the realism, we instead employ the Fréchet Inception Distance [28] and the Kernel Inception Distance [8] in both paired (*i.e.*, FID$_p$ and KID$_p$) and unpaired (*i.e.*, FID$_u$ and KID$_u$) settings. For the LPIPS and SSIM implementation, we use the torch-metrics Python package [13], while for the FID and KID scores, we employ the implementation in [47].

4.2 Implementation Details

We first train the EMASC modules, the textual-inversion adapter, and the warping component. Then, we freeze the weights of all modules except for the textual inversion adapter and train the proposed enhanced Stable Diffusion pipeline*. In all our experiments, we generate images at 512×384 resolution.

Textual Inversion. The textual inversion network F_θ consists of a single ViT layer followed by a multi-layer perception composed of three fully-connected layers separated by a GELU non-linearity [27] and a dropout layer [59]. We set the number of PTEs generated by F_θ to 16. We train F_θ for 200k steps, with batch size 16, learning rate 1e-5 with 500 warm-up steps using a linear schedule, AdamW [39] as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay equal to 1e-2. As the visual encoder V_E, we leverage the OpenCLIP ViT-H/14 model [63] pre-trained on LAION-2B [56].

Diffusion Virtual Try-On Model. We train the proposed virtual try-on pipeline for 200k iterations, with batch size 16 and the same optimizer and scheduling strategy used to train the textual inversion network. At training time, we randomly mask the text, the warped garment, and the pose map input with a probability of 0.2 for each condition. This allows the later use of the classifier-free guidance technique [30] at inference time. Following [1], we use the fast variant of the multi-conditional classifier-free guidance, which allows computing the final result with a computational complexity independent from the amount of the input constraints.

*https://huggingface.co/stabilityai/stable-diffusion-2-inpainting

Table 2: Quantitative results on the VITON-HD dataset [9]. The * marker indicates results reported in previous works.

Model	**LPIPS** ↓	**SSIM** ↑	**FID$_p$** ↓	**KID$_p$** ↓	**FID$_u$** ↓	**KID$_u$** ↓
CP-VTON* [60]	-	0.791	-	-	30.25	40.12
ACGPN* [66]	-	0.858	-	-	14.43	5.87
VITON-HD [9]	0.116	0.863	11.01	3.71	12.96	4.09
HR-VITON [36]	0.097	**0.878**	10.88	4.48	13.06	4.72
LaDI-VTON	**0.091**	0.876	**6.66**	**1.08**	**9.41**	**1.60**

Autoencoder with EMASC. We apply the proposed EMASC modules to the variational autoencoder of the Stable Diffusion model. In particular, each EMASC module consists of two convolutional layers, where a SiLU non-linearity [16] activates the first one. We apply the EMASC modules to the `conv_in` layer output and the feature before the `down_block` connecting each encoder layer to its corresponding decoder one. The convolutional layers have a kernel size of 3, padding of 1, and stride of 1. The first convolutional layer maintains the number of channels constant, while the second one adapts the channel axis dimension to the decoder features. Finally, we sum the EMASC output to the corresponding decoder features. We train the EMASC modules for 40k steps with batch size 16, learning rate equal to 1e-5, AdamW as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1e-2. Also, in this case, we perform 500 warm-up steps with a linear schedule. We employ a combination of the L1 and VGG [34] loss functions, scaling the perceptual VGG loss term by a factor of 0.5. In our setting, we found the VGG loss essential to avoid blurriness in the reconstructed images. During training, the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are frozen (see Figure 3), and only the EMASC modules are learned.

4.3 Experimental Results

Comparison with State-of-the-Art Models. We compare our method with several state-of-the-art competitors. For the Dress Code dataset, we compare our method with CP-VTON [60] and PSAD [44], retrained from scratch using the same image resolution of our model (*i.e.*, 512×384) using the source codes when available or otherwise implementing them. Following [44], we also

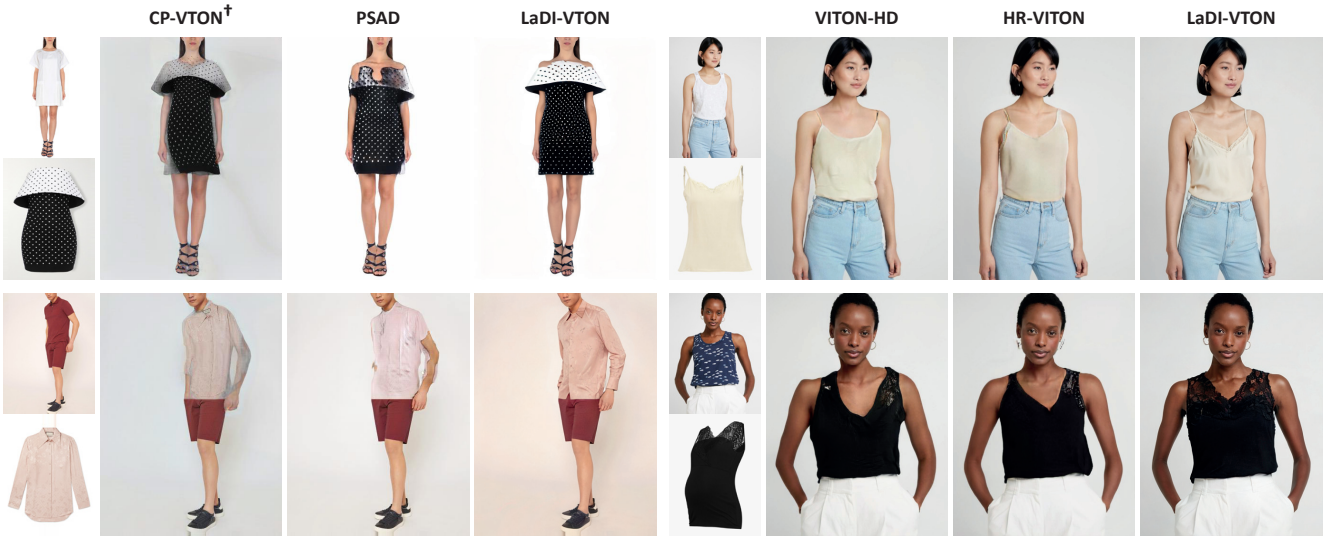

Figure 4: Qualitative results generated by LaDI-VTON and competitors on Dress Code [44] (left) and VITON-HD [9] (right).

include an improved version of CP-VTON (*i.e.*, CP-VTON†) where we add as additional input the masked image I_M. For the VITON-HD dataset, instead, we compare our model with VITON-HD [9] and HR-VITON [36] using source codes and checkpoints released by the authors to extract the results. Given that some evaluation scores (*e.g.*, LPIPS and FID) are very sensitive to different implementations, to ensure a fair comparison, we compute the quantitative results of these methods using the same metric implementation of our model. For completeness, we also include in the comparison some additional virtual try-on methods for which the results are from previous works and, therefore, may have been obtained using different evaluation source codes.

Table 1 reports the quantitative results on the Dress Code dataset. As can be seen, LaDI-VTON achieves comparable results to PSAD [44] in terms of coherence with the inputs (*i.e.*, LPIPS and SSIM), while significantly outperforming all competitors in terms of realism in both paired and unpaired settings. In particular, on the Dress Code test set, our model reaches a FID score of 4.14 and 6.48 for the paired and unpaired settings, respectively. These results are considerably lower than the best-performing competitor (*i.e.*, PSAD). In Table 2, we instead show the quantitative analysis of the VITON-HD dataset. Also, in this case, LaDI-VTON surpasses all other competitors by a large margin in terms of FID and KID, demonstrating its effectiveness in this setting.

To qualitatively evaluate our results, we report in Figure 4 sample images generated by our model and by the competitors. Notably, our solution can generate high-realistic images and preserve the texture and details of the original in-shop garments, as well as the physical characteristics of target models.

Human Evaluation. To further evaluate the generation quality of our model, we conduct a user study to measure both the realism of generated images and their coherence with the inputs given to the virtual try-on model. Overall, we collect around 2,000 evaluations for each test, involving more than 50 unique users. In Table 3, we report the percentage of times in which an image generated by our model is preferred against a competitor. As can be seen, LaDI-VTON is always selected more than 60% of the time, further confirming the progress over previous methods.

Table 3: User study results on the unpaired test set of both datasets. We report the percentage of times an image from LaDI-VTON is preferred against a competitor.

Dataset	Model	Realism	Coherence
Dress Code	CP-VTON [60]	93.10	89.68
	CP-VTON† [60]	80.21	75.69
	PSAD [44]	74.14	70.83
VITON-HD	VITON-HD [9]	79.19	71.48
	HR-VITON [36]	77.95	60.98

Table 4: Quantitative results on the entire Dress Code test set [44] using different model configurations.

Model	LPIPS ↓	SSIM ↑	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓
w/o text	0.071	0.902	4.99	1.61	8.50	3.70
w/ retrieved text	0.070	0.903	4.85	1.61	7.49	2.93
w/ F_θ and standard SD	0.105	0.876	5.42	1.87	7.50	2.83
w/o warped garment	0.068	0.904	4.50	1.44	**6.30**	**1.99**
LaDI-VTON	**0.064**	**0.906**	**4.14**	**1.21**	6.48	2.20

Configuration Analysis. In Table 4, we study the model performance by varying its configuration. We conduct this analysis on the Dress Code test set. In particular, the experiment in the first row replaces the Stable Diffusion textual input $\hat{Y}$ with an empty string. The one in the second row replaces the Stable Diffusion textual input $\hat{Y}$ with textual elements retrieved using the in-shop garment image C as the query for a CLIP-based model [6]. The results show that the proposed textual inversion adapter outperforms the other textual input alternatives. The third experiment regards the textual inversion adapter condition abilities, in particular, we can see that it is possible to obtain excellent results by using the proposed textual inversion adapter to condition an out-of-the-box Stable Diffusion model. Finally, we test the warped garment C_W input in the

Table 5: Quantitative analysis changing the number of predicted V^*. Results are reported on the Dress Code test set [44] using the out-of-the-box Stable Diffusion as backbone.

# V^*	LPIPS ↓	SSIM ↑	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓
1	0.115	0.867	6.14	2.24	8.19	3.14
4	0.108	0.873	5.87	2.15	8.17	3.10
16	0.105	0.876	5.42	1.87	7.50	2.83
32	0.103	0.878	5.37	1.80	7.66	2.92

Table 6: Analysis on the effectiveness of the proposed Enhanced Mask Aware Skip Connection modules. Results are reported on Dress Code [44] and VITON-HD [9].

	Model	EMASC	Masked	LPIPS ↓	SSIM ↑
Dress Code	SD VAE	None	-	0.0214	0.9538
	SD VAE	Linear	✓	0.0196	0.9636
	SD VAE	Non-Linear	✗	0.0183	0.9646
	SD VAE	Non-Linear	✓	**0.0181**	**0.9652**
	LaDI-VTON	None	-	0.0642	0.8985
	LaDI-VTON	Non-Linear	✓	**0.0640**	**0.9060**
VITON-HD	SD VAE	None	-	0.0260	0.9336
	SD VAE	Linear	✓	0.0220	0.9545
	SD VAE	Non-Linear	✗	0.0203	0.9560
	SD VAE	Non-Linear	✓	**0.0200**	**0.9561**
	LaDI-VTON	None	-	0.0960	0.8491
	LaDI-VTON	Non-Linear	✓	**0.0907**	**0.8758**

overall pipeline by removing it. In this case, we can see that this additional input helps in the paired setting, but interestingly does not appreciably contribute to the unpaired one.

Analysis on V^*. In Table 5, we show the results of the out-of-the-box Stable Diffusion inpainting model conditioned using the textual inversion module when varying the number of PTEs generated by F_θ. Overall, we obtain the best scores in terms of FID and KID on the unpaired setting using 16 pseudo-word tokens embeddings, while for the metrics on the paired setting employing 32 PTEs leads to slightly better results. Since increasing the number of PTEs can increase memory usage, the best trade-off between computational load and performance is reached when using 16 PTEs.

Effectiveness of EMASC modules. We test the proposed EMASC modules on the paired settings of Dress Code and VITON-HD. In particular, we test the EMASC performance on both the autoencoder $\mathcal{A}$ for image reconstruction and the final model (*i.e.*, LaDI-VTON) for the complete virtual try-on task. In the first experiment, we simply encode and then decode the model image I obtaining the reconstructed image $\hat{I}$ (*i.e.*, $\hat{I} = \mathcal{D}(\mathcal{E}(I))$). In the second experiment, we compare the performance of our complete model with and without the EMASC modules. While for the first experiment, LPIPS and SSIM are computed by comparing the model image I with its reconstruction $\hat{I}$, in the second experiment, we evaluate the metrics by comparing the model image I with its reconstruction $\tilde{I}$, where we define $\tilde{I}$ as the output of the virtual try-on pipeline. Results reported in Table 6 show that the proposed method can enhance both the reconstruction capabilities of the Stable Diffusion autoencoder and the output performance of the final virtual try-on pipeline leading to better evaluation scores.

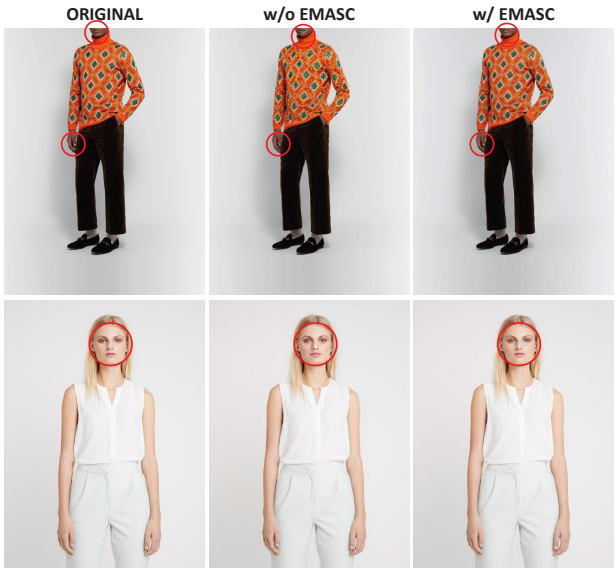

Figure 5: Image reconstruction results from the Stable Diffusion autoencoder with and without the EMASC modules.

To better assess the contribution of the EMASC modules in the autoencoder analysis, we compare the proposed EMASC method with two of its variants. The first variant involves removing the feature masking after the final convolutional layer, while in the second variant, we use only one convolutional layer without any non-linear activation. We can notice that the masked non-linear EMASC modules achieve better results in all metrics on both datasets. In Figure 5, we also show sample qualitative results of the Stable Diffusion autoencoder with and without EMASC modules. As it is possible to see, the proposed learnable mask-aware skip-connections reduce the reconstruction loss resulting in better faces, hands, and feet. Note that we achieve such results without retraining or fine-tuning the autoencoder.

5 CONCLUSION

In this work, we propose the first latent diffusion-based approach for virtual try-on. To increase the detail retention of the input in-shop garment, we exploit the textual inversion technique for the first time in this task, demonstrating its capability in conditioning the generation process. Moreover, we introduce the EMASC modules that can enhance the inpainting output image quality reducing the autoencoder compression loss of LDMs. This advancement notably improves the human perceived quality of high-frequency human body details such as hands, faces, and feet. Results show that the proposed LaDI-VTON model outperforms by a large margin the competitors in terms of realism on both Dress Code and VITON-HD datasets, two widely used benchmarks for the task.

ACKNOWLEDGMENTS

This work has partially been supported by the European Horizon 2020 Programme (grant number 101004545 - ReInHerit) and by the PRIN project "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content" (CUP B87G22000460001), co-funded by the Italian Ministry of University.

REFERENCES

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[3] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. 2022. Single stage virtual try-on via deformable attention flows. In *Proceedings of the European Conference on Computer Vision.*

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[6] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[7] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. 2023. Person Image Synthesis via Denoising Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[8] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations.*

[9] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 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.*

[10] Guillem Cucurull, Perouz Taslakian, and David Vazquez. 2019. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[11] Giannis Daras and Alexandros G Dimakis. 2022. Multiresolution Textual Inversion. In *Advances in Neural Information Processing Systems Workshops.*

[12] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2023. Disentangling features for fashion recommendation. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 1s (2023), 1–21.

[13] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. TorchMetrics-Measuring Reproducibility in PyTorch. *Journal of Open Source Software* 7, 70 (2022), 4101.

[14] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems.*

[15] Jean Duchon. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables.*

[16] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* 107 (2018), 3–11.

[17] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. 2022. C-VTON: Context-Driven Image-Based Virtual Try-On Network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision.*

[18] Emanuele Fenocchi, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Fabio Cesari, and Rita Cucchiara. 2022. Dual-Branch Collaborative Transformer for Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[19] Matteo Fincato, Federico Landi, Marcella Cornia, Fabio Cesari, and Rita Cucchiara. 2021. VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In *Proceedings of the International Conference on Pattern Recognition.*

[20] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proceedings of the International Conference on Learning Representations.*

[21] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. 2021. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems.*

[23] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[24] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[25] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. 2023. Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion. *arXiv preprint arXiv:2303.08767* (2023).

[26] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[27] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415* (2016).

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *Advances in Neural Information Processing Systems.*

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems.*

[30] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *Advances in Neural Information Processing Systems Workshops.*

[31] Wei-Lin Hsiao and Kristen Grauman. 2018. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[32] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. 2020. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Proceedings of the European Conference on Computer Vision.*

[33] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics* 41, 4 (2022), 1–11.

[34] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision.*

[35] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[36] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *Proceedings of the European Conference on Computer Vision.*

[37] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. 2021. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[38] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[39] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations.*

[40] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[41] Chenlin Meng, Yutong He adnd Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Proceedings of the International Conference on Learning Representations.*

[42] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[43] Davide Morelli, Marcella Cornia, Rita Cucchiara, et al. 2021. FashionSearch++: Improving consumer-to-shop clothes retrieval with hard negatives. In *CEUR Workshop Proceedings.*

[44] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision.*

[45] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning.*

[46] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning.*

[47] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on*

Machine Learning.

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*.

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[53] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *Proceedings the ACM SIGGRAPH Conference*.

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.

[55] Rohan Sarkar, Navaneeth Bodla, Mariya I Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. 2023. OutfitTransformer: Learning Outfit Representations for Fashion Recommendation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*.

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations*.

[59] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958.

[60] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision*.

[61] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. 2022. Pretraining is All You Need for Image-to-Image Translation. *arXiv preprint arXiv:2205.12952* (2022).

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

[63] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[64] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[65] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[66] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 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*.

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[68] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. 2022. EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. In *Advances in Neural Information Processing Systems*.