Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.059530

ARTICLE

HRAM-VITON: High-Resolution Virtual Try-On with Attention Mechanism

Yue Chen[1], Xiaoman Liang[1,2,*], Mugang Lin[1,2], Fachao Zhang[1] and Huihuang Zhao[1,2]

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China

[2]Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, 421002, China

*Corresponding Author: Xiaoman Liang. Email: liangxm@hynu.edu.cn

Received: 10 October 2024 Accepted: 22 November 2024 Published: 17 February 2025

ABSTRACT

The objective of image-based virtual try-on is to seamlessly integrate clothing onto a target image, generating a realistic representation of the character in the specified attire. However, existing virtual try-on methods frequently encounter challenges, including misalignment between the body and clothing, noticeable artifacts, and the loss of intricate garment details. To overcome these challenges, we introduce a two-stage high-resolution virtual try-on framework that integrates an attention mechanism, comprising a garment warping stage and an image generation stage. During the garment warping stage, we incorporate a channel attention mechanism to effectively retain the critical features of the garment, addressing challenges such as the loss of patterns, colors, and other essential details commonly observed in virtual try-on images produced by existing methods. During the image generation stage, with the aim of maximizing the utilization of the information proffered by the input image, the input features undergo double sampling within the normalization procedure, thereby enhancing the detail fidelity and clothing alignment efficacy of the output image. Experimental evaluations conducted on high-resolution datasets validate the effectiveness of the proposed method. Results demonstrate significant improvements in preserving garment details, reducing artifacts, and achieving superior alignment between the clothing and body compared to baseline methods, establishing its advantage in generating realistic and high-quality virtual try-on images.

KEYWORDS

Virtual try-on; attention mechanism; high-resolution; image generation

1 Introduction

Numerous studies have highlighted the growing demand for virtual fitting technology in response to the increasing prevalence of online shopping. This area has attracted significant research attention due to its potential to enhance the online retail experience. Virtual try-on technology provides an innovative solution, enabling customers to visualize how garments purchased online would appear when worn, without the necessity of physical fitting. The primary objective of virtual try-on is to generate realistic and precise fitting images by seamlessly integrating a given garment with a reference image. The synthesized image should adhere to the following criteria to the greatest extent possible: (1) Preserve the garment's inherent attributes, including its color, texture, contour, and shape; (2) Maintain the original physical characteristics of the figure, such as body structure, posture, and proportions;

(3) Ensure the garment is naturally adjusted to align with the figure's position and posture; and (4) Accurately render the body regions covered by the garment. The primary challenge in virtual try-on lies in accurately aligning the target garment with the corresponding anatomical regions of the figure, a task that is inherently complex.

Prior studies have attempted to address the alignment challenge through explicit distortion models. While these algorithms effectively mitigate issues of clothing alignment and distortion, the limited resolution of the generated images (e.g., 256 × 192 pixels) often fails to meet the visual quality expectations of end users. To address this issue, methods such as VITON-HD [1] and HR-VITON [2] have been developed for virtual try-on using high-resolution datasets, producing high-quality fitted images. These approaches not only enhance visual impact but also effectively resolve issues related to clothing warping in high-resolution images, facilitating better alignment with the human body. However, challenges remain in accurately preserving critical clothing characteristics, such as color fidelity, intricate patterns, collar structure, and sleeve length integrity.

We propose a novel two-stage virtual try-on framework, consisting of a garment warping stage and an image generation stage, to facilitate the precise extraction of essential clothing features and address the aforementioned challenges effectively. In the garment warping stag stage, we introduce a channel attention mechanism, which effectively preserves important clothing details. During the image generation stage, to maximize the utilization of input image information, we design an upsampling normalization module (SNM) that collects two samples during the normalization process. The main contributions of this work are summarized as described subsequently:

- A channel attention method is incorporated into the garment warping step to evaluate the significance of each feature channel in the feature map, highlight important channels, suppress unimportant features, and address the loss of clothing details in producing fitting pictures.
- We address the issue of target information loss by designing an SNM module into the generator, which enhances subsequent model training and learning.
- The SmoothL1 loss function is used instead of the L1 loss function to train the clothing warping stage, guaranteeing the correctness of the warping clothing information.

2 Related Work

Virtual Try-On Based on Image: Research in virtual try-on generally focuses on two techniques: three-dimensional (3D) modeling [3–7] and two-dimensional (2D) image-based methods [8–12]. While 3D fitting technology shows promise, its adoption is currently limited by high costs, complex modeling requirements, and the need for specialized equipment. In contrast, 2D techniques are gaining popularity due to their simplicity, cost-effectiveness, and accurate results. Thin Plate Splines (TPS) are commonly used to warp clothing, aligning apparel images with the target object's posture. This technique, initially proposed by Han et al. [13] as part of VITON, employs a coarse-to-fine image generation approach. Several studies [10,14–17] have further utilized TPS for warping, establishing relationships between clothing and figure images through feature extraction and adjusting the clothing to fit relevant body sections. However, most two-stage approaches require training multiple networks. VITON-HD [1] and HR-VITON [2] improve conditional GAN performance for high-resolution images, but challenges such as detail loss remain. Our method addresses these limitations, producing images that better align with inputs while preserving finer details.

Effective Attention Mechanism: In deep learning (DL), especially in Natural Language Processing (NLP), the attention mechanism is crucial for identifying key information and efficiently allocating processing resources. The core idea of the attention mechanism is to assign weights to different input

parts, adjusting the model's focus accordingly to the human visual system, which highlights important scene elements while ignoring minor ones. This selectivity helps reinforce key information, boosting model performance. Bahdanau et al. [18] applied the attention mechanism to machine translation by assigning different attention weights to each word in the source sentence. This enables the model to focus on relevant information, capture long-sequence context effectively, and enhance translation quality. Vaswani et al. [19] introduced the self-attention mechanism in the Transformer model, which assigns weights to each word based on its relationships with other words in the input sequence. This allows the model to focus on relevant information, effectively capturing context in long sequences and improving translation quality. Hu et al. [20] introduced a channel attention technique that assigns weights to each channel in the feature map. This allows the model to prioritize important information, optimize resource allocation, and improve expressiveness and performance. Woo et al. [21] introduced the Convolutional Block Attention Module (CBAM). This module combines spatial and channel attention mechanisms to enhance important features by weighting various spatial positions and channels in the feature map. This method improves feature selectivity and optimizes resource allocation, allowing the model to process crucial information more efficiently. Originally popular in machine translation, researchers now widely adopt attention mechanisms in image generation. In this study, a channel attention mechanism was incorporated into the garment warping stage. This addition captures local cross-channel interaction data, learns the critical features of each channel in the feature map, and highlights critical feature channels. This approach helps retain more detailed information about clothing and ensures the effectiveness and accuracy of the generated data.

Normalized Module: As deep neural networks grow deeper, the distribution of activated input values shifts, causing nonlinear functions like sigmoid to saturate, which results in vanishing gradients and slower convergence during backpropagation. To address this, data normalization is commonly used; however, it may erase semantic information and make features overly similar, negatively impacting image generation. Park et al. [22] proposed spatially adaptive normalization, which uses semantic layout as input to generate scaling coefficients through convolution. This approach preserves semantic information while enhancing image realism. Building upon this principle, the method proposed in this study is similar to spatially adaptive normalization, as it ensures spatial normalization of generator activations at multiple granularity levels. By operating at different resolutions, this method effectively utilizes input data, preserving critical information while enhancing output quality.

Appearance Flow: Appearance flow refers to 2D coordinate vectors indicating which pixels in the source can be used to synthesize the target. Originally introduced in image-based virtual try-on (VTON) by [8], appearance flow has since garnered considerable attention. Chopra et al. [23] applied 3D appearance flow to generate images of people in target poses, calculating appearance flow as supervision by fitting a 3D model-an approach not available for 2D try-on. PF-AFN [24] employed knowledge distillation to extract the appearance flow between human and clothing images, achieving an accurate, dense correspondence between the two. He et al. [17] proposed estimating the appearance flow by applying a global style vector through style modulation. Similarly, HR-VITON [2] incorporates appearance flow to enhance model performance and visual quality. Beyond VTON, appearance flow has proven useful in other tasks, such as novel view synthesis and feature mapping that distorts human pose transfer. In the garment warping stage of this paper, we use existing appearance flow technology as the sampling grid for clothing warping, which offers the advantages of lossless information transfer and detail preservation.

3 Method

In this section, the proposed high-resolution virtual try-on model with attention mechanism is introduced.

3.1 Architecture

Using a reference image of a person $I \in R^{3\times H\times W}$ and a garment image $G \in R^{3\times H\times W}$ (H and W denote the height and width of the images, respectively), the objective is to generate an image $I_C \in R^{3\times H\times W}$, depicting a person wearing garment G while preserving body structure and posture. The proposed framework achieves this goal through two main stages, as illustrated in Fig. 1: (a) the garment warping stage and (b) the image generation stage. In the garment warping stage, given the clothing image G and the human body segmentation image S, which shows clothing occlusion, the method simultaneously deforms G and generates the segmentation map $S^{\wedge}$ (refer to Section 3.2). The outputs $C^{\wedge}$ and $S^{\wedge}$ from the garment warping stage serve as inputs to the image generation stage, where the final try-on image I_C is synthesized (refer to Section 3.3).

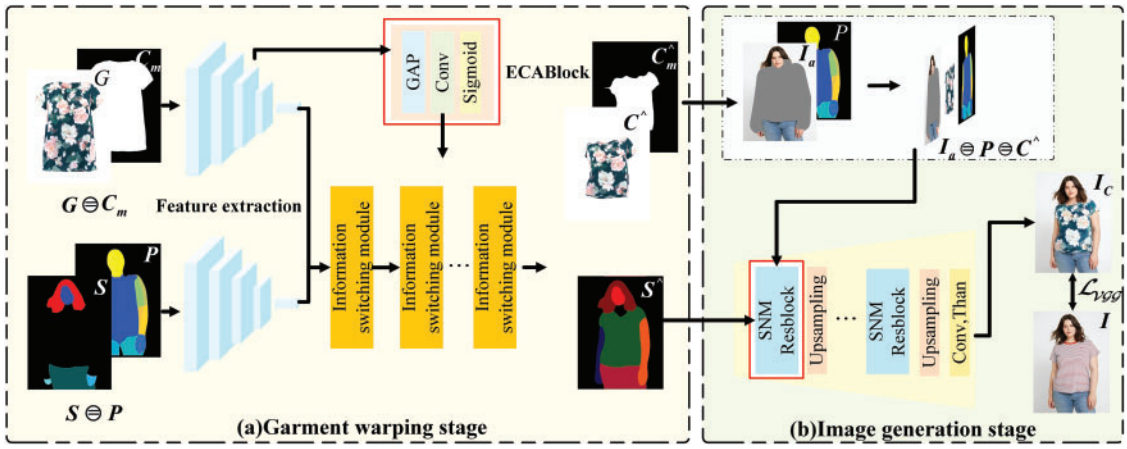

Figure 1: General frame diagram. The figure pose heatmap $P \in R^{3\times H\times W}$ and segmentation map $S \in L^{H\times W}$ are obtained via the figure image I processing, where L is a set of integers indicating the semantic labels and the clothing mask image C_m is obtained by clothing image G processing

3.2 Garment Warping Stage

The main objective in this stage is to generate the segmentation diagram $S^{\wedge}$ of the figure wearing the target costume G, warp the costume G to fit the human posture, and then employ the warped costume image $C^{\wedge}$ and the segmentation diagram $S^{\wedge}$ as inputs for the image generation stage. The basic structure of the proposed garment warping stage is illustrated in Fig. 2. It consists of a decoder and two identical feature encoders: a garment encoder E_C and a segmentation encoder E_S. The given (G, C_m) and (S, P) are processed through the feature encoders to extract the features. The features extracted by the garment encoder are then fed into the attention module to enhance them. Feature fusion of each layer from both encoders is achieved using an information fusion module. Consequently, the segmented image wearing the target clothes and the distorted clothing image's appearance flow are predicted using the fusion module. Ultimately, the segmented figure $S^{\wedge}$, the warped clothes G, and the warped clothing mask C_m are obtained.

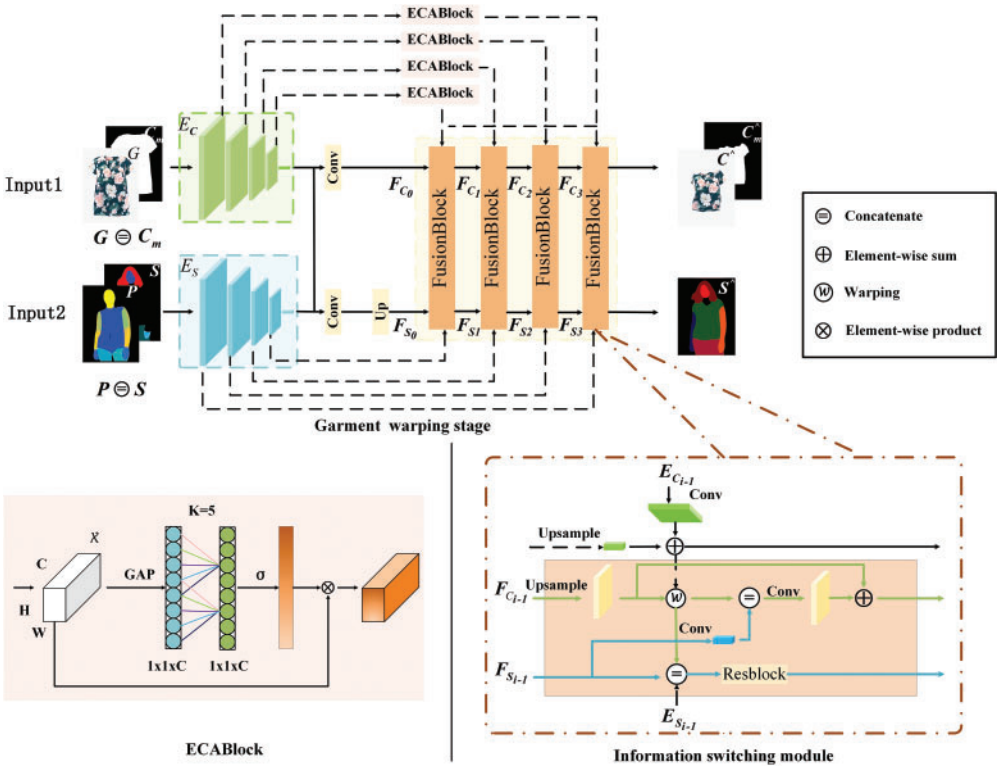

Figure 2: Garment warping stage. The ECABlock and Information switching module are described in detail

ECAAttention Module: It is not straightforward to determine which feature layers play a major role and which play a minor role when extracting features from feature maps. To address this, the Efficient Channel Attention (ECA) module adjusts channel weights based on the feature extraction process. This ensures that feature maps with greater significance exert a larger impact on the results, achieving more effective feature extraction compared to regular convolutional layers. Moreover, the ECAAttention [25] builds upon the SEAttention [20] by replacing the Multi-Layer Perceptron (MLP) module with one-dimensional (1D) convolution, as illustrated by the ECABlock in Fig. 2. This modification substantially reduces computational complexity. The comparison of the associated computational complexities is presented in Table 1, and the data presented herein is derived from ECA-Net [25].

Table 1: Comparison of the ECAAttention and SEAttention models on ImageNet, focusing on network parameters (Param), floating point operations per second (FLOPs), and inference speed (frames per second, FPS)

Model	Param	FLOPs	Inference
SEAttention(MLP)	63.68M	10.85G	**761FPS**
ECAAttention(Conv1D)	**57.40M**	**10.83G**	785FPS

Given the aggregate features generated by average pooling $[1, 1, C]$, the ECABlock computes channel weights using 1D convolution with a kernel size of K, where K is adaptively determined by mapping the channel dimension C through a nonlinear function. In addition, the number of convolution kernels is set to 2^K, as the typical channel size is 2. The corresponding expressions are defined in Eqs. (1) and (2):

$$C = 2^{(\gamma * k - b)} \tag{1}$$

$$k = \psi(C) = |\frac{log_2(c)}{\gamma} + \frac{b}{\gamma}|_{odd} \tag{2}$$

where $||_{odd}$ indicates that k must be odd, γ and b are set to 2 and 1, respectively, to adjust the ratio between the channel count C and the convolution kernel size. Nonlinear mapping is applied in lower-dimensional channels to produce shorter interactions, while linear mapping achieves this in higher-dimensional channels.

Before injecting the relevant clothing features into the information switching module, we first pass them through the ECABlock for feature enhancement. ECA Attention assigns weights to these channels, selectively emphasizing key features to improve the network's ability to discriminate important characteristics. Initially, each channel of the input feature map undergoes global average pooling to capture comprehensive information. Next, fully connected layers are used to scale the features, generating channel-specific weights through an activation function. Finally, these weights are applied across the channels of the original feature map, strengthening important features while suppressing less significant ones. This approach effectively preserves critical attributes, such as color and texture, as confirmed by subsequent experiments.

Information Switching Module: The information switching module consists of two branches: the garment appearance flow branch F_{C_i} and the segmentation flow branch F_{S_i}, as shown in Fig. 2. The garment appearance flow branch F_{C_i} extracts clothing appearance features, while the segmentation flow branch F_{S_i} processes features obtained from the segmentation encoder. These features are passed into the information exchange module, and the outputs are forwarded to the next stage. The two branches interact to compute the appearance flow and segment the image.

3.3 Image Generation Stage

In this stage, the final fitting image I_C is generated by integrating the unknown clothing image I_a, the warped clothing image $C^\wedge$, the segmented image $S^\wedge$, and the pose image P guided by $S^\wedge$. Fig. 3 shows the generator's network framework, composed of residual blocks and upsampling layers. Each residual block, except the first layer, receives additional input, ensuring progressively more information is captured.

The generator employs a set of residual blocks with an upsampling layer (referred to as SNM Resblocks). Each SNM Resblock consists of three convolution layers and three normalization layers, as shown in Fig. 3b SNM Block. The operation of SNM is similar to that of SPADE [22]. SNM samples the segmented map (Seg) twice. In Fig. 3b, Seg is first upsampled and downsampled before interpolation, which we refer to as double sampling. This process reduces high-frequency noise while preserving the image's approximate structure and size, thereby efficiently utilizing the input data and enhancing the model's ability to recognize objects of various sizes and shapes, ultimately producing higher-quality images.

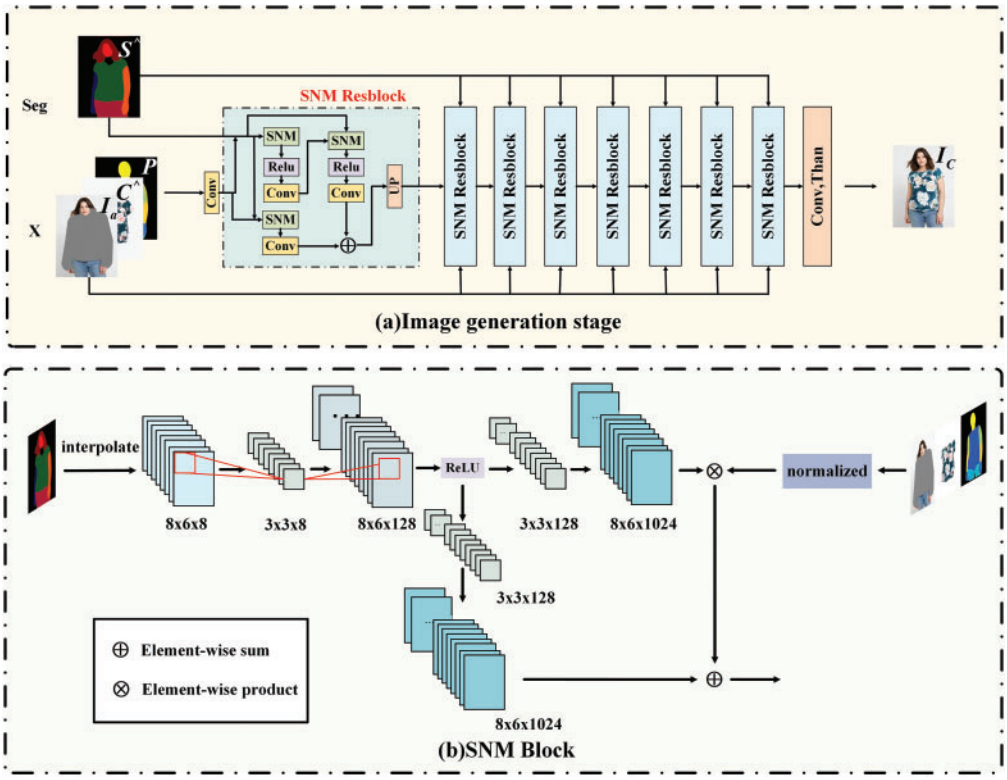

Figure 3: Image generation stage. The composition of Image generation stage and SNM Block are described in detail

The SNM ResBlock operates at multiple resolutions, so inputs are resized before normalization and fed into each layer. Similarly, the generator's input values are adjusted to different resolutions. Before each SNM ResBlock, the convolutional layer connects the resized inputs to the previous layer's activation, and each SNM ResBlock uses these connected inputs to enhance activations.

The input resolution for each subsequent SNM ResBlock is 2^{n-1} times the initial input resolution, where n denotes the number of SNM ResBlocks, and the input resolution for the first SNM ResBlock is set to the lowest resolution (8×6 pixels). This approach enables multi-scale feature-level refinement, which better preserves garment details compared to single-scale pixel-level refinement.

3.4 Training Losses

To train the garment warping stage, pixel-level cross-entropy loss $£_{CE}$ is applied between S (the one-hot encoding of the analytical graph) and the projected segmentation graph $\hat{S}$. To ensure accurate clothing deformation aligned with human posture, SmoothL1 loss and perceptual loss are also employed. These loss functions regulate the intermediate flow, preventing vanishing gradients and improving training efficiency. Formally, the losses $£_{Smooth_{L1}}$ and $£_{vgg}$ are defined in Eqs. (4) and (5):

$$Smooth_{L1}(x) = \begin{cases} 0.5 * (x)/beat, & if|x| < beat \\ |x| - 0.5 * beat, & otherwise \end{cases} \tag{3}$$

$$£_{Smooth_{L1}} = \sum_{i=0}^{3} w_i \cdot Smooth_{L1}(W(C_m, F_{C_i}), S_c) + Smooth_L(S^{\wedge}, S_c) \tag{4}$$

$$£_{vgg} = \sum_{i=0}^{3} w_i \cdot \phi(W(G, F_{C_i}), I_c) + \phi(C^{\wedge}, I_c) \tag{5}$$

where w_i denotes the relative significance of each component, W represents warping, as proposed in the warping method of HR-VIION [2], ϕ denotes the VGG loss function, and finally, S_c and I_c, respectively represent the parsed cloth and its mask.

Moreover, $£_{tv}$ denotes the total variation loss, which promotes the smoothness of the appearance flow, as formulated in Eq. (6):

$$£_{tv} = ||\nabla F_{C_4}||_1 \tag{6}$$

For end-to-end training at the warping garment stage, the objective function is shown in Eq. (7):

$$£_w = \lambda_{CE}£_{CE} + £_{cGAN} + \lambda_{SmoothL1}£_1 + £_{vgg} + \lambda_{tv}£_{tv} \tag{7}$$

where $£_{GAN}$ is the conditional GAN loss between S and $S^{\wedge}$, λ_{CE}, $\lambda_{SmoothL1}$, and λ_{tv} represent the hyperparameters of relative importance between different losses.

To train the image generator, the same loss used in SPADE and pix2pixHD [26] is employed in this algorithm. Conditional antagonistic loss, perceptual loss, and feature matching loss represent the different parts of our objective loss function. Therefore, the target loss function is expressed by Eq. (8):

$$£_G = £_{CG} + \lambda_{vgg}£_{vgg} + \lambda_F£_F \tag{8}$$

where $£_{CG}$, $£_{vgg}$, and $£_F$ denote the antagonistic loss, perceptual loss, and feature matching loss, respectively.

4 Experiment

4.1 Dataset

In this experiment, we employed the high-resolution virtual try-on dataset provided by VITON-HD [1]. The dataset includes 13,679 positive samples of women wearing various garments, with each image originally sized at 1024×768 pixels. Images were down-sampled as necessary to meet resolution requirements. A training set of 11,647 pairs and a test set of 2032 pairs were created from the dataset.

4.2 Implementation Details

All experiments were implemented using PyTorch on an NVIDIA RTX 3090 GPU. The batch sizes for the garment warping and image generation stages were set to 8 and 1, respectively. The generator learning rate was fixed at 0.0002, and the discriminator learning rates were set to 0.0002 and 0.0004, respectively. The AdamW optimizer was adopted with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for the garment warping stage, and $\beta_1 = 0.0$ and $\beta_2 = 0.999$ for the image generation stage. Training was conducted over 100,000 and 200,000 iterations for the garment warping and image generation stages, respectively, with the model being saved every 10,000 iterations.

4.3 Qualitative Results

Comparison with Baselines: To qualitatively evaluate the proposed strategy, we compared it with several State-of-the-Art (SOTA) baseline methods. Due to limitations in experimental resources, we excluded diffusion models from the comparison. The selected methods include VITON-HD [1], HR-VITON [2], PF-AFN [24], and SD-VITON [27]. Fig. 4 presents the results of our approach compared to these baselines, using publicly available code implementations for the baseline methods. It is evident that the previous approaches struggle to generate satisfactory images. While recent methods have improved clothing fit to the body, they often fail to retain critical details, such as the original color, pattern, and neckline shape of the garments. In contrast, the proposed method ensures that the outfit conforms to the character's physique while preserving the original garment's features.

Figure 4: (Continued)

Figure 4: Qualitative comparison with baselines. Compared with Copyright 2021, IEEE. VITON-HD: Reprinted with permission from Reference [1]. Copyright 2021 IEEE. HR-VITON: Reprinted with permission from Reference [2]. PF-AFN: Reprinted with permission from Reference [22]. Copyright 2022 ECCV. SD-VITON: Reprinted with permission from Reference [27]. Copyright 2024 AAAI

Referring to the first row of Fig. 4, the baseline method fails to preserve the original neckline shape, sleeve length, and other details after warping the garment. The fifth row presents results using a side view, where the baseline approach struggles with significant pose deformations, leading to a poor fit. In contrast, the proposed approach successfully adapts the garment to the character's body shape and performs well even when parts of the body are occluded.

Effectiveness of the Attention Mechanism: We incorporated the attention mechanism to retain more information during garment warping, as illustrated in Fig. 5 (first row, fifth column). The red and yellow dotted lines highlight regions where the baseline method loses details, such as the pattern near the arm (red box). In contrast, the proposed method preserves more visual information (yellow box). Furthermore, in the second row (third and fourth columns), the proposed approach generates a garment color that more closely matches the original.

Figure 5: Effectiveness of attention mechanisms

Fig. 6 shows that the attention mechanism improves the alignment of warped clothing with body parts. In the first row, our method matches better the segmap clothing regions, with sleeve width aligning more accurately with the arm. The second row highlights our approach that warps the clothing to fit the body's curvature more precisely, validating its effectiveness.

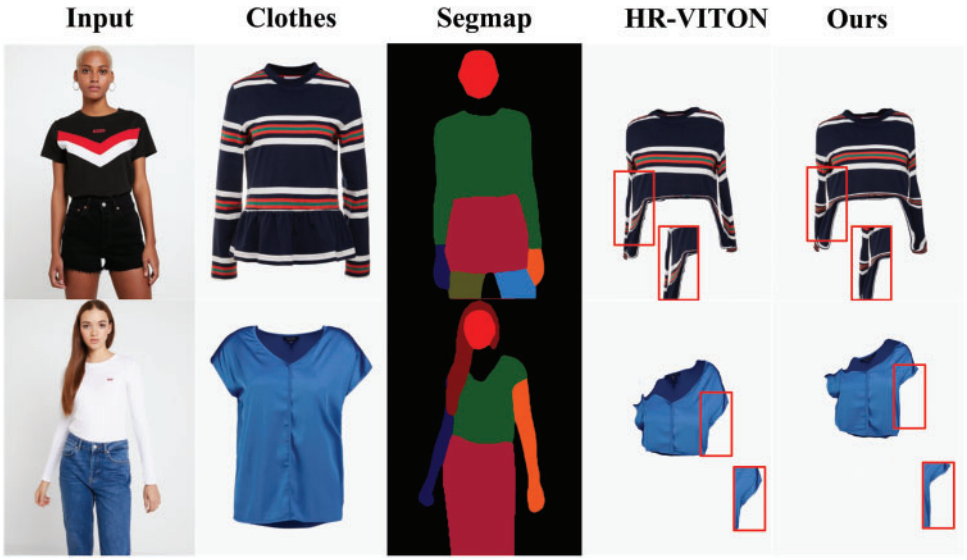

Figure 6: Demonstration of the effectiveness of attention mechanism on garment warping

4.4 Quantitative Results

We quantitatively compare our method to baseline approaches under both paired and unpaired settings. In the paired setting, the task is to reconstruct the person's image using the original clothing image, while in the unpaired setting, the objective is to swap the clothing item of the person's image. It is important to note that variations in experimental equipment and environmental conditions may result in discrepancies between the data reported here and those in the original paper.

Compared With the Baseline Method: As shown in Table 2, we compare our approach with industry-standard methods, including VITON-HD [1], HR-VITON [2], FS-VITON [17], PF-AFN [24], and SD-VITON [27]. In the paired settings, we evaluate the quality of the generated composite images using metrics such as the Structural Similarity Index Measure (SSIM) [28], Peak Signal-to-Noise Ratio (PSNR) [29], and Learned Perceptual Image Patch Similarity (LPIPS) [30], comparing the composite images to the target (ground truth) images. For unpaired settings, we assess performance using the Fréchet Inception Distance (FID) [31].

Utilize The 512 × 384 Dataset: Additionally, we evaluated our method on a dataset with a resolution of 512 × 384 pixels and compared its performance against several state-of-the-art approaches, including VITON-HD [1], HR-VITON [2], FS-VITON [17], DAFLow [32], PF-AFN [24], and SD-VITON [27]. The results of this quantitative comparison are summarized in Table 3. Our method demonstrates robust performance, even with lower-resolution images.

Table 2: Quantitative comparison with baseline methods on a dataset with a resolution of 1024 × 768

Methods	SSIM ↑	LPIPS ↓	PSRN ↑	FID ↓
PF-AFN	0.8079	0.2134	15.0392	32.2668
VITON-HD	0.8716	0.0806	20.8346	12.3131
FS-VITON	0.8430	0.1043	19.7314	37.0364
HR-VITON	0.8840	0.0654	22.0112	11.8212
SD-VITON	0.8956	**0.0627**	22.4036	12.5387
Ours	**0.8960**	0.0630	**23.2814**	**10.6912**

Table 3: Quantitative comparison with baseline methods on a dataset with a resolution of 512 × 384

Methods	SSIM ↑	LPIPS ↓	PSRN ↑	FID ↓
PF-AFN	0.8540	0.0935	21.0184	27.8159
VITON-HD	0.8568	0.0813	21.2007	11.8384
DAFLow	0.7623	0.2500	16.2416	92.5041
FS-VITON	0.8476	0.1035	20.1014	29.3052
HR-VITON	0.9090	0.0720	21.8201	26.9497
SD-VITON	0.9167	0.0673	22.0065	**11.6040**
Ours	**0.9223**	**0.05987**	**22.7936**	23.0787

4.5 Ablation Study

Fig. 7 presents the results of sub-optimization experiments, where we excluded the attention mechanism in the clothing warping stage, removed SNM from the image generation stage, and omitted SmoothL1 loss during model training. Compared to the results generated by our approach, these results show notable issues such as loss of garment detail and poor generation quality, demonstrating the effectiveness of our proposed method.

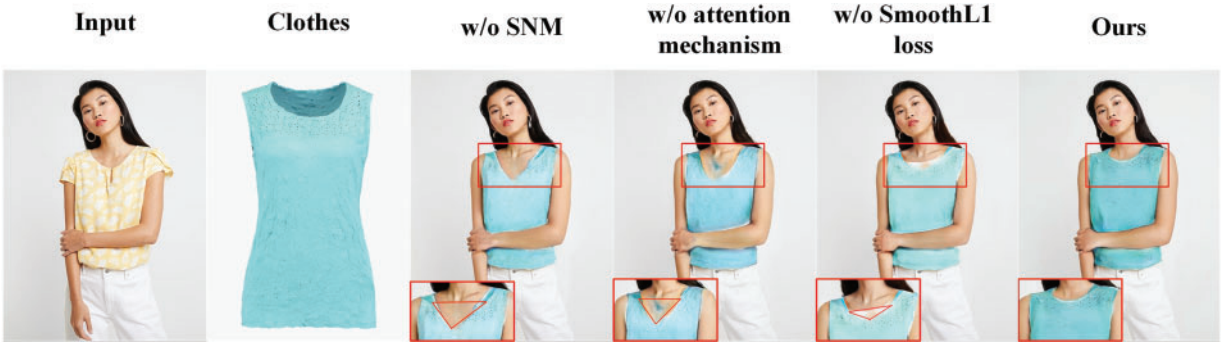

Figure 7: (Continued)

Figure 7: Ablation diagram. It shows the effects of different methods and works best when all methods are used together

Additionally, the line plot in Fig. 8 shows that training the model with SmoothL1 loss improves the overlap between the predicted results and the actual combinations, thereby enhancing the accuracy of the generated images.

Table 4 presents the results of the ablation study, which evaluated suboptimal configurations where the attention mechanism was excluded from the garment warping step, the model was not trained with SmoothL1 loss, and SNM was not used in the image generation stage. The comparison demonstrates that incorporating the attention mechanism and SmoothL1 loss in the clothing warping stage, as well as using SNM in the image generation stage, significantly improves the model's accuracy and performance, thereby highlighting the effectiveness of the proposed approach.

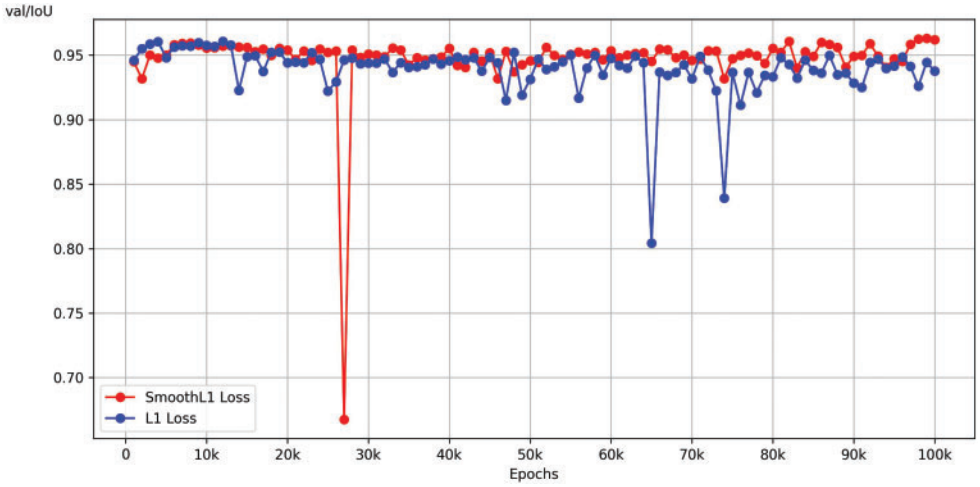

Figure 8: The model trained with SmoothL1 loss and L1 loss achieved an Intersection over Union (IoU) score on the validation set. The IoU measures the overlap between the region predicted by the model and the corresponding ground truth region. The values range from 0 to 1, with higher values indicating greater overlap

Table 4: Ablation experiment

Methods	SSIM ↑	PSRN ↑
w/o attenion mechanism	0.8474	21.1594
w/o SNM	0.8646	21.2292
w/o smoothL1 loss	0.8850	22.2733
Our method	**0.8960**	**23.2814**

5 Conclusion

This study has significant implications for the field of virtual try-ons, offering a wide range of potential applications. We propose a virtual try-on framework that utilizes an attention mechanism to improve feature transmission, resulting in better preservation of clothing texture, pattern, color, and sleeve lengths. Experimental results demonstrate that our approach outperforms existing techniques at a resolution of 1024×768 pixels. In future research, we aim to develop more refined methods and further advance virtual fitting technology. Our future work will focus on enhancing the realism of fitting images and expanding the framework's applicability to a broader range of apparel categories. We anticipate continued progress and advancements in subsequent studies.

Acknowledgement: None.

Funding Statement: This work was supported by the National Natural Science Foundation of China (61772179), Hunan Provincial Natural Science Foundation of China (2022JJ50016, 2023JJ50095), and the Science and Technology Plan Project of Hunan Province (2016TP1020), Double First-Class University Project of Hunan Province (Xiangjiaotong [2018]469, [2020]248).

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Yue Chen, Xiaoman Liang, Mugang Lin, Fachao Zhang, and Huihuang Zhao; data collection: Yue Chen; analysis and interpretation of results: Yue Chen, Xiaoman Liang, Mugang Lin, Fachao Zhang, and Huihuang Zhao; draft manuscript preparation: Yue Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest toreport regarding the present study.

References

[1] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2021, pp. 14131–14140.

[2] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," in *Comput. Vis.–ECCV 2022*, Israel, Springer, 2022, pp. 204–219.

[3] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "DRAPE: Dressing any person," *ACM Trans. Grap.*, vol. 31, no. 4, pp. 1–10, 2012. doi: 10.1145/2185520.2185531.

[4] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in *Int. Conf. on 3D Body Scan. Technol.*, Lugano, Switzerland, 2014, pp. 406–413.

[5] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, 2017. doi: 10.1145/3072959.3073711.

[6] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape, and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7365–7375.

[7] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3D: clothed 3D humans," in *Comput. Vis.–ECCV 2020*, UK, Springer, vol. 2020, pp. 344–359.

[8] X. Han, X. Hu, W. Huang, and M. R. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10471–10480.

[9] H. Dong *et al.*, "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9026–9035.

[10] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 7850–7859.

[11] R. Yu, X. Wang, and X. Xie, "VTNFP: An image-based virtual try-on network with body and clothing feature preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10511–10520.

[12] D. Song, T. Li, Z. Mao, and A. -A. Liu, "SP-VITON: Shape-preserving image-based virtual try-on network," *Multimed. Tools Appl.*, vol. 79, pp. 33757–33769, 2020. doi: 10.1007/s11042-019-08363-w.

[13] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 7543–7552.

[14] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.

[15] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu and P. Luo, "Disentangled cycle consistency for highly-realistic virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2021, pp. 16928–16937.

[16] T. Issenhuth, J. Mary, and C. Calauzenes, "Do not mask what you do not need to mask: a parser-free virtual try-on," in *Comput. Vis.–ECCV 2020*, UK, Springer, 2020, pp. 619–635.

[17] S. He, Y. -Z. Song, and T. Xiang, "Style-based global appearance flow for virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2022, pp. 3470–3479.
[18] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.
[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 7132–7141.
[21] S. Woo, J. Park, J. -Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[22] T. Park, M. -Y. Liu, T. -C. Wang, and J. -Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2019, pp. 2337–2346.
[23] A. Chopra, R. Jain, M. Hemani, and B. Krishnamurthy, "ZFlow: Gated appearance flow-based virtual try-on with 3D priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5433–5442.
[24] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2021, pp. 8485–8493.
[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 11534–11542.
[26] T. -C. Wang, M. -Y. Liu, J. -Y. Zhu, A. Tao, J. Kautz and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 8798–8807.
[27] S. -H. Shim, J. Chung, and J. -P. Heo, "Towards squeezing-averse virtual try-on via sequential deformation," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, pp. 4856–4863, 2024. doi: 10.1609/aaai.v38i5.28288.
[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. doi: 10.1109/TIP.2003.819861.
[29] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *2010 20th Int. Conf. Pattern Recognit.*, IEEE, 2010, pp. 2366–2369.
[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inform. Process. Syst.*, vol. 30, pp. 6626–6637, 2017.
[32] S. Bai, H. Zhou, Z. Li, C. Zhou, and H. Yang, "Single stage virtual try-on via deformable attention flows," in *Computer Vision-ECCV 2022*, Israel, Springer, 2022, pp. 409–425.