Article

FA-VTON: A Feature Alignment-Based Model for Virtual Try-On

Yan Wan, Ning Ding * and Li Yao

School of Computer Science and Technology, Donghua University, 2999 North Renmin Road, Shanghai 201620, China; winniewan@dhu.edu.cn (Y.W.); yaoli@dhu.edu.cn (L.Y.)
* Correspondence: dingning@mail.dhu.edu.cn

Abstract: The virtual try-on technology based on 2D images aims to seamlessly transfer provided garments onto target person images. Prior methods mainly concentrated on warping garments and generating images, overlooking the influence of feature alignment on the try-on results. In this study, we initially analyze the distortions present by existing methods and elucidate the critical role of feature alignment in the extraction stage. Building on this, we propose a novel feature alignment-based model (FA-VTON). Specifically, FA-VTON aligns the upsampled higher-level features from both person and garment images to acquire precise boundary information, which serves as guidance for subsequent garment warping. Concurrently, the Efficient Channel Attention mechanism (ECA) is introduced to generate the final result in the try-on generation module. This mechanism enables adaptive adjustment of channel feature weights to extract important features and reduce artifact generation. Furthermore, to make the student network focus on salient regions of each channel, we utilize channel-wise distillation (CWD) to minimize the Kullback–Leibler (KL) divergence between the channel probability maps of the two networks. The experiments show that our model achieves better results in both qualitative and quantitative analyses compared to current methods on the popular virtual try-on datasets.

Keywords: deep learning; virtual try-on; image generation; knowledge distillation

check for updates

Citation: Wan, Y.; Ding, N.; Yao, L. FA-VTON: A Feature Alignment-Based Model for Virtual Try-On. *Appl. Sci.* **2024**, *14*, 5255. https://doi.org/10.3390/app14125255

Academic Editor: Atsushi Mase

Received: 6 May 2024
Revised: 6 June 2024
Accepted: 12 June 2024
Published: 17 June 2024

1. Introduction

With the surge in online shopping, virtual try-on technology has emerged as a highly esteemed innovative solution, aiming to enhance users' try-on experience in response to the inherent challenge of not being able to physically try on garments while shopping online. Currently, research on virtual try-ons can be broadly categorized into two main types: based on 3D models [1–3] and based on 2D images [4–11].

Virtual try-on technology based on 3D methods requires 3D parametric human body models [12]. Acquiring data for 3D models presents challenges due to its difficulty and cost, which in turn restricts its accessibility and complicates its promotion and application. Conversely, image-based models offer a more straightforward and cost-effective means of obtaining image data, thereby expanding their potential applications. Consequently, they have garnered growing interest from researchers.

While 2D virtual try-on technology offers reduced complexity compared to its 3D counterpart and can yield satisfactory results, it still faces limitations. Many existing methods overlook the crucial aspect of feature alignment during the extraction process. Low-quality multi-scale features often lead to undesirable warping effects. For example, blurred edges in person features may cause garments to warp disproportionately, resulting in unclear try-on images or generating artifacts that adversely affect the try-on experience. As shown in Figure 1, we choose the current SOTA parser-free model FS-VTON [7] for comparison. In the second column of each sub-figure, we can see that the feature maps extracted by the FS-VTON model are very blurry, especially at the boundaries of objects. The third column shows the features extracted by our model, which are significantly higher in quality than those of FS-VTON, especially in terms of garment patterns and boundaries.

In virtual try-on tasks, we not only want warped garments to retain as much original information as possible but also need to remove parts that did not appear in the final try-on image. However, existing methods struggle to distinguish garment boundaries, remove necessary details, like complex texture structures, or retain unnecessary parts, such as the back fabric of garments.

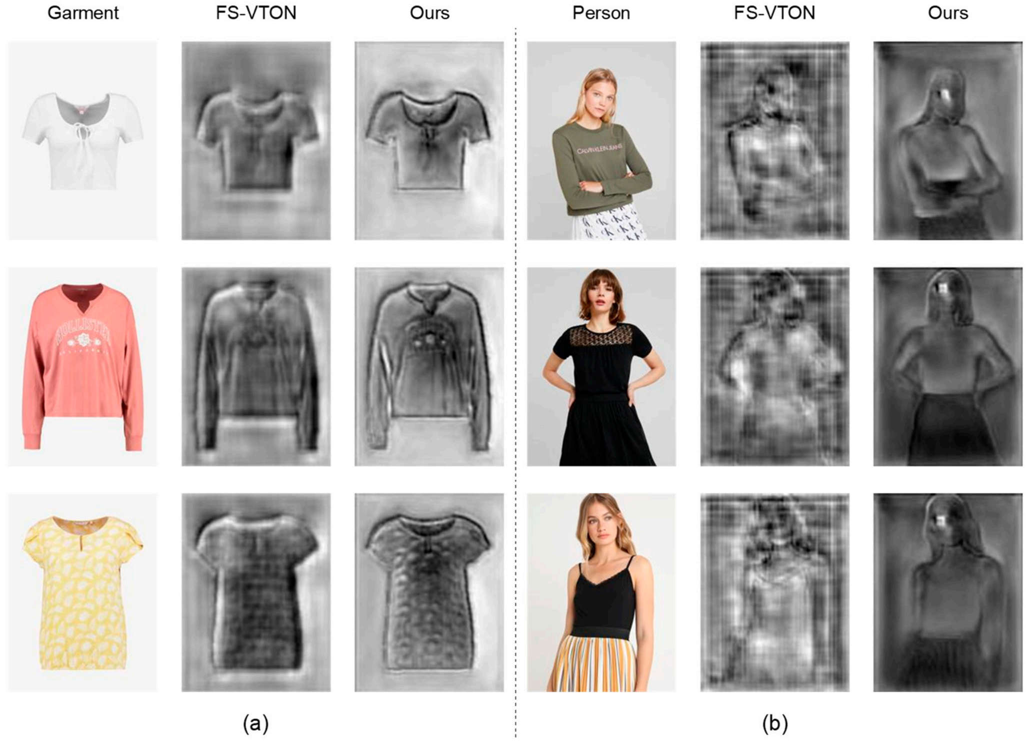

Figure 1. Comparison of feature maps generated by our method with FS-VTON. We visualize the feature maps by compressing the color channels. (**a**) Comparison of garment feature maps; (**b**) comparison of person feature maps.

Furthermore, existing models fail to consider the weighted processing of channel information during try-on image generation. The generated results may not adequately enhance retained garment features and fail to suppress incorrectly warped garment features, resulting in artifacts. Finally, virtual try-on models typically combine various person or garment parsing results as prior knowledge inputs. Most parser-free models simply minimize point-wise differences, neglecting the impact of semantic information from different channels in teacher models.

To handle the above issues, we propose a feature alignment-based model (FA-VTON) without relying on people or garment parsing. Firstly, in the feature extraction stage, we introduce a Feature Alignment Selection Module (FASM), which consists of two parts: a Feature Alignment Module (FAM) and a Feature Selection Module (FSM). The FAM effectively learns pixel transformation offsets to contextually align upsampled higher-level features, allowing the model to learn more precise boundaries of both the person and garment image. It helps guide accurate garment warping. The FSM selectively retains important features and reduces irrelevant feature interference. Secondly, we utilize the ECA mechanism in the try-on image generation stage. It learns effective channel attention and avoids losses caused by dimensionality reduction. The try-on image generation module effectively mitigates errors resulting from incorrect parsing or excessive garment deformation, improving the quality of generated images. Lastly, we adopt a knowledge distillation framework to train our parser-free virtual try-on model and utilize a channel-wise knowledge distillation method to learn significant feature information from the teacher network.

In summary, our contributions are as follows. Firstly, we analyze the twisted try-on results by existing methods and explain the importance of the feature alignment in the extraction module in garment deformation and image generation. Secondly, we propose a feature alignment-based model (FA-VTON) that learns accurate person and garment boundaries from upsampled features without relying on human or cloth parsing. This method enables more precise garment warping by aligning upsampled features. Then, we adopt the ECA mechanism in the try-on image generation module to generate more realistic try-on images and reduce the impact of incorrect deformations. Finally, based on the training framework of the parser-free model, we introduce a channel-wise knowledge distillation method to learn rich channel information from the teacher model, enhancing the feature extraction capability of student models and thereby generating high-quality try-on results.

2. Related Works

2.1. Image-Based Virtual Try-On

Image-based virtual try-on technology aims to seamlessly transfer selected garments onto target person images to produce realistic and natural-looking results. CAGAN [13] made the first attempt to address the limitations of training triplet data (original person image, chosen garment, and image post-try-on) using a cycle consistency structure, but the quality of the generated images fell short of expectations. VITON [5] broke down the virtual try-on process into three key stages: parsing the person image, warping the garment, and generating and refining the result. It introduced a person representation method that did not rely on the original garments but retained identity characteristics to compensate for the lack of supervised training data. However, even with these improvements, images produced by the VITON method still suffered from imperfections and bad quality. Presently, image-based virtual try-on methods follow the VITON framework, typically involving three main steps: parsing and extracting features from person and garment images, warping garments, and generating the final try-on results. To improve the quality of try-on images and eliminate imperfections, most approaches focus on refining the garment warping and try-on generation stages. For instance, CP-VTON [10] employs convolutional neural networks (CNNs) to learn Thin Plate Spline (TPS) [14] parameters for garment warping, resulting in better-fitting garments and more realistic try-on effects. However, CP-VTON struggles with complex garment deformations in different poses. ClothFlow [4] introduced appearance flow [15] to enhance garment adjustments. To tackle the challenge of complex poses, FS-VTON [7] adopts a StyleGAN [16]-based architecture to estimate a global appearance flow. RMGN-VTON [17] improves the generation network, which uses a regional mask to fuse the features of garment and person images. HR-VTON [9] proposed a try-on condition generator to warp garments and a try-on image generator guided by a segmentation map. Some methods utilize image inpainting and reconstruction techniques [18–23] for virtual try-on tasks, such as TWD [24–26]. Despite these advancements in garment warping and try-on image generation, they ignore the impact of feature alignment. Since the garment warping module must refine features at different scales, the quality of extracted features directly influences the accuracy of garment warping. Poor quality features can lead to misaligned images, reducing the effectiveness of the warping module. Therefore, the design of a robust feature extraction module is critical in virtual try-on models. In this paper, we analyze existing shortcomings and explain the significance of the feature extraction module for garment warping and try-on image generation.

Previous methods typically utilize two simple Feature Pyramid Networks (FPNs) [27] to extract multi-scale person and garment features, guiding garment warp based on the extracted features. FPNs are commonly used to address multi-scale information processing issues in tasks, such as object detection and semantic segmentation, improving the model's sensitivity to image details. Since being introduced into virtual try-on tasks after PF-AFN [6], they have been widely used [7,28–30]. However, the FPN overlooks the issue of feature alignment, and directly adding pixels between upsampling and local features can cause a misalignment of contextual feature mappings, leading to incorrect garment warping and unrealistic results. In this paper, we introduce the Feature Alignment Selection Module (FASM) [31] into the feature extractor to address the problem of misaligned contextual feature mappings. Additionally, since existing methods overlook the issue of selecting different feature information in the try-on image generation module, which can result in images retaining incorrect garment deformation results, we introduce the ECA mechanism [32] into the model to weigh different channel information, mitigating adverse results caused by excessive deformation.

2.2. Parser-Free VTON

To reduce the model's dependence on additional parsers during inference, some research attempts to train networks using parser-free methods. Parser-free methods for virtual try-on allow the model to infer without relying on other person or garment parser models, only requiring the input of garment and person images. WUTON [8] is a pioneering parser-free method in the virtual try-on field but produces significant artifacts and fails to achieve the desired results. PF-AFN [6] introduces a teacher-assistant training pipeline based on knowledge distillation, reducing the influence of erroneous teacher results on the student model, and becoming the standard for subsequent parser-free methods. SDAFN [30] can generate try-on images in a single stage, but it still requires additional human pose key points. DM-VTON [28] adopts a new knowledge distillation framework based on FS-VTON and introduces virtual guiding poses to improve the model's ability to generate complex poses, achieving real-time parser-free virtual try-on while maintaining quality. In addition to human pose information, GP-VTON [29] also utilizes a garment parsing map. In this paper, we propose an FA-VTON model trained using a parser-free method, requiring no parsers during inference. Additionally, we utilize channel-wise knowledge distillation loss to help the student model learn rich semantic information from the teacher network.

3. Twisted Analysis

Current models have made various improvements in both garment warping and try-on image generation stages, but they have overlooked the importance of feature alignment in the extraction stage. We analyze the issues present in current virtual try-on results and explain the impact of feature alignment on garment warping and the final try-on result. Parser-based models, unlike parser-free ones, require multiple people or garment parsing results as inputs. Person representation types include pose key points, semantic segmentation, and DensePose segmentation [33]. Garment representation types include semantic segmentation and corresponding landmarks [34,35]. The previous models simply concatenated the inputs from these different channel-wise modalities, without considering aligning them. To illustrate the importance of feature alignment in virtual try-on tasks, we take the results of FS-VTON as an example. We can visualize the final layer feature maps and obtain grayscale images by compressing the color channels.

Unlike 3D virtual try-on, where garments can be modeled based on the body's spatial structure, 2D virtual try-on requires our model to correctly retain the front part of the garment shown after trying it on and remove the back fabric that will not appear in the result. As shown in Figure 2, in the first two rows, the try-on result retains the back fabric of the garment at the neckline, while in the third row, the detailed part of the neckline that should be retained is removed. In the feature maps in the second column, we can see that the model can only extract the rough outline of the garment and cannot distinguish complex clothing details. The garment warping module calculates the appearance flow based just on the extracted features to generate the warped garment. Consequently, misaligned areas in the feature extraction stage result in the erroneous retention or removal of areas in the warped garment in the third column, thereby affecting the final try-on result.

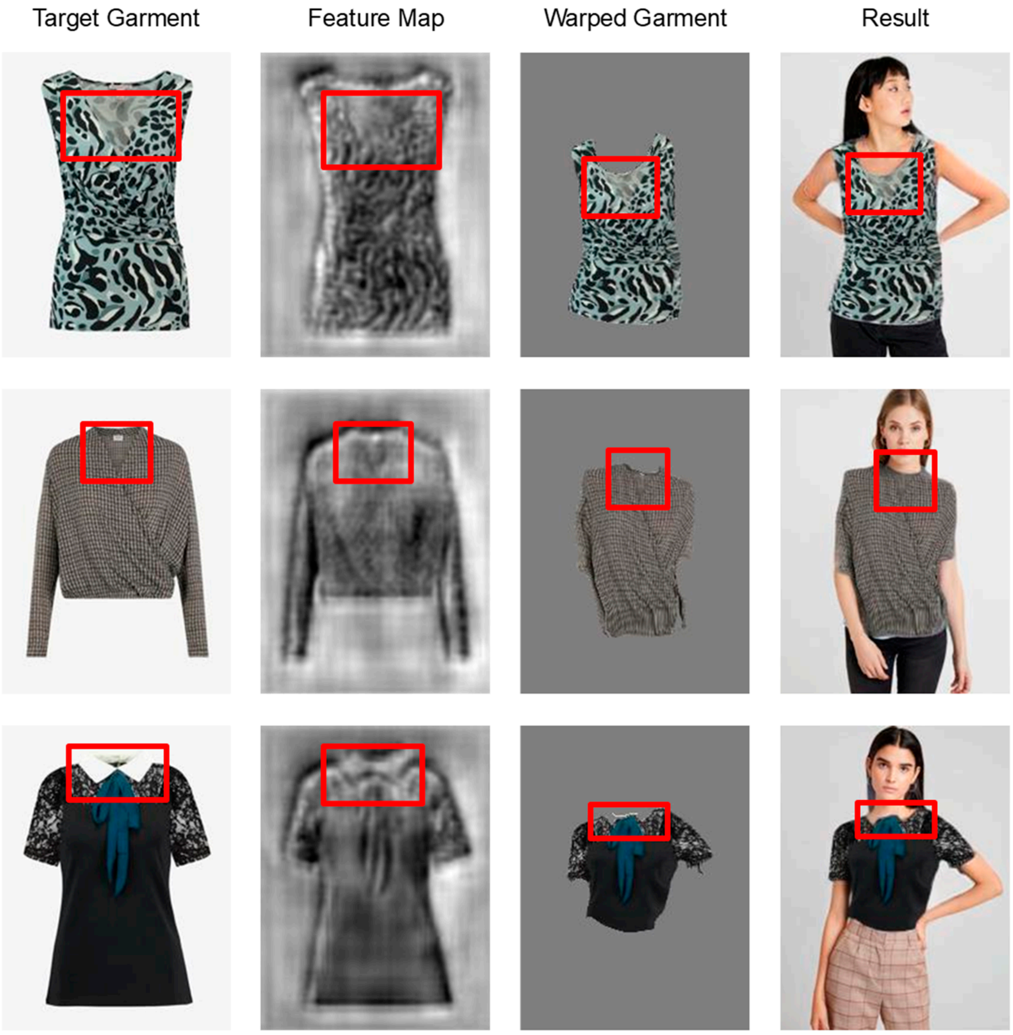

Figure 2. The influence of garment feature extraction on try-on results.

Then, we analyze the influence of the person feature extractor on the try-on results, as shown in Figure 3. We classify the upper body, including the arms and neck, as the deformable garment regions. Features like the face and other areas that do not require deformation, such as the hands and lower body, are identified as non-deformable regions. We expect the person feature extractor to accurately identify the deformable garment regions. This should exclude the influence of existing clothing from the original person image and guide the subsequent garment deformation. Parser-free methods struggle to exclude interference from existing garments during input. Therefore, the person feature extractor must accurately identify real human body boundaries to determine regions to be retained and warped.

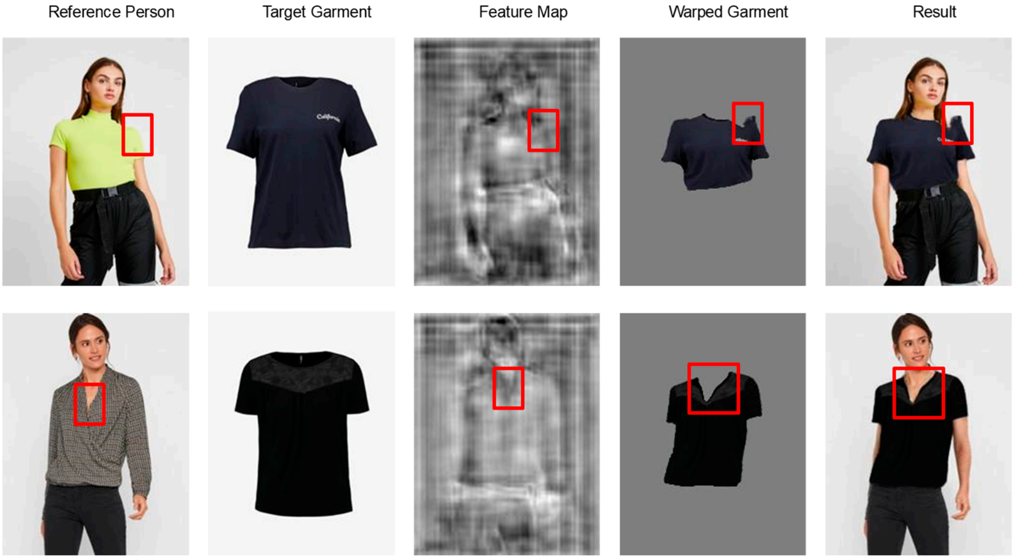

Figure 3. The influence of person feature extraction on try-on results.

We classify the twists caused by existing garment feature extraction into two categories. The first category is twists from unclear edges of human images. In the first row in the trial results, excessive twisting occurs at the shoulders. This happens because the person feature extractor fails to learn clear edge features, leading to an ambiguous deformable garment region. The garment warping module requires guidance from the human feature map for deformation. Thus, the appearance flow struggles to learn the accurate position of corresponding pixels. Under multiple spatial transformations, this deviation is magnified, causing twists at the edges. The second category is twists from the inability to eliminate the influence of existing garments. In the second row, the try-on image retains the V-neck design of the original garment, while the target garment has a round neck. The feature extractor is disturbed by the existing clothing, incorrectly retaining the skin at the neckline of the original image. As a result, this leads the garment to deform incorrectly, ultimately impacting the try-on outcome.

After the above analysis, we can observe the influence of the alignment in feature extraction. The garment warping module requires clear boundaries of the garment and the region to be warped to calculate the correct appearance flow, ensuring that the garment properly conforms to the human body. The try-on module directly generates missing body parts, such as the neck and arms, based on the warped garment. Incorrectly warped garments are directly reflected in the try-on results. Therefore, aligning features is essential for virtual try-on tasks.

4. Methods

Our aim is to seamlessly transfer clothing onto human images while preserving individual identity features. The proposed FA-VTON comprises three components: the Feature Extraction Module (FEM), the Coarse-to-Fine Warping Module (CFWM), and the try-on image generation module (TGM), detailed in Sections 4.1–4.3, respectively. To achieve parser-free virtual try-on, we adopt the training strategy employed by existing non-parsing models [6]. Initially, we pre-train a parser-based model and then employ knowledge distillation to utilize it as a teacher model to facilitate the training of the final parser-free model. Additionally, we introduce channel-wise knowledge distillation to help the student network learn rich features from the teacher network, as elaborated in Section 4.4. The main architecture of FA-VTON is illustrated in Figure 4. FA-VTON maintains an identical structure to the teacher network within the student network but with different inputs.

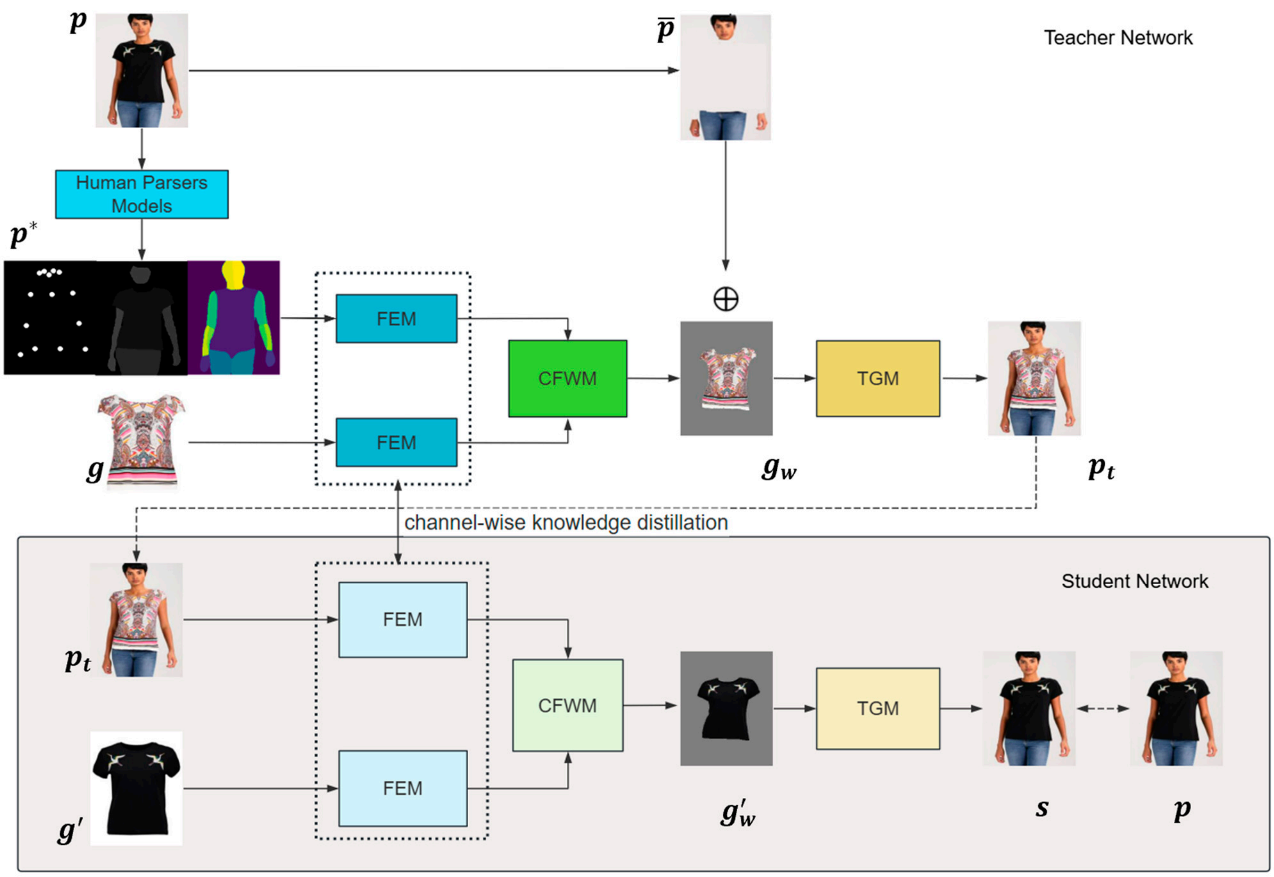

Figure 4. The overall architecture framework of FA-VTON.

4.1. Overview of FA-VTON Framework

Figure 4 shows a teacher network that requires parsers and a student network that does not. Prior to training the student network, we must first train the teacher network. We use Human Parsers Models to obtain the parsing results of the human image p, including human pose estimation, human body segmentation, and DensePose results. These parsing results are used as a representation p^* and garment image g, respectively, which serve as inputs to the FEMs. This approach helps reduce interference from the original clothing and eliminates the need for triplet data (i.e., the original human image, target clothing, and human image wearing the target clothing). The CFWM predicts the appearance flow and generates the warped garment g_w with it. Subsequently, we use the preserved parts of the person image $\overline{p}$ together with the warped garment image g_w as inputs to the GEM module supervised by the real image p.

During the student model training, we utilize the generated results p_t of the teacher model as inputs for the student model. In this case, we do not need a human parser model but can simply input p_t together with clothing images g' into the student model. The student model predicts the appearance flow and warps the clothing g'_w. The generation module synthesizes the images p_t with the warped garment image g'_w as inputs and generates the final result image s under the supervision of p. During the training process for the student network, we also employ the knowledge distillation function to help the student network learn the rich features extracted by the teacher network.

4.2. Feature Extraction Module (FEM)

In the architecture of virtual try-on models, especially in teacher models, various human and garment parsing results are often utilized as prior knowledge. They are concatenated along the channel dimension and collectively serve as the input to the network. This requires our feature extraction module not only to effectively learn from additional parsing information but also to ignore features that could interfere with the final image generation, reducing the generation of artifacts.

We have devised a feature extraction module to address the issue of feature alignment and fusion of multimodal information during the feature extraction process for virtual try-on. Building on the traditional feature pyramid, we introduced a Feature Alignment Module (FAM) in its top-down portion and a Feature Selection Module (FSM) in its lateral connection section. We term this structure the Feature Alignment Selection Module (FASM). It aligns person and garment features, enhancing the feature fusion capability of the multi-scale network to better guide subsequent garment deformation. The network structure of the person FEM is illustrated in Figure 5.

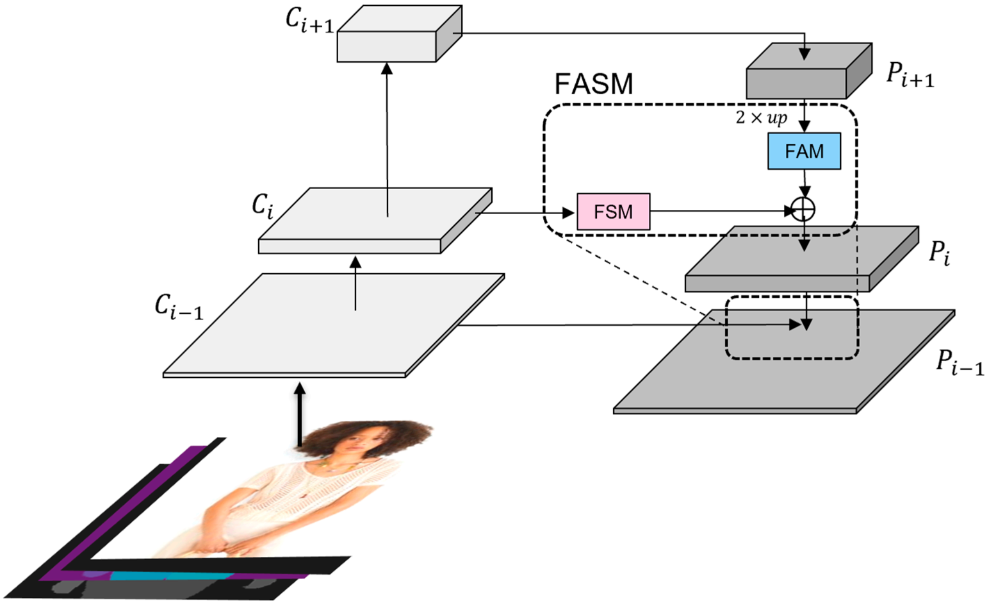

Figure 5. The structure of the person FEM. The garment FEM has the same structure with different inputs. The dashed box represents the structure of the FASM, consisting of the FAM and FSM.

In the figure, the image in the bottom left corner represents the input image for training. During the training phase, various human parsing results are often input to the model for reference. The multi-scale feature maps output by the residual blocks are shown above the input image, while the fused multi-scale feature maps in the pyramid network are displayed on the right. In our model, we set the FEMs with five layers. In the bottom-up process, the number of output channels is [64, 128, 256, 256, 256], and the feature size is halved with each downsampling operation. In the top-down process, the feature size changes correspondingly to the bottom-up process, while the channel dimension remains fixed at 256.

4.2.1. Feature Alignment Module (FAM)

The recursive downsampling operations operated by the FPN often have spatial misalignments between the upsampled high-dimensional feature map and the corresponding low-dimensional feature map. Consequently, the simplistic feature fusion methods used by the traditional FPN, such as element-wise addition or channel-wise concatenation, can adversely affect the boundary perception of the target object.

However, regardless of whether it is for person or garment feature extraction, errors in boundary perception can significantly impact the subsequent warping stage. Errors in extracting the edges of a person can lead to incorrect shaping of garments. Similarly, boundary perception errors in garment feature extraction can result in the loss of garment details or erroneous retention of back fabric.

In this paper, we use a Feature Alignment Module to adjust the upsampled feature map by learning offsets to align various sampling positions. Specifically, we adjust the upsampled feature P_i^{up} based on spatial positional information provided by $\hat{C}_{i-1}$. We represent spatial positional information using a 2D feature map, where each offset value can be viewed as the two-dimensional spatial displacement distance between each point in P_i^{up} and its corresponding point in $\hat{C}_{i-1}$.

The specific structure of the FAM is illustrated in Figure 6a.

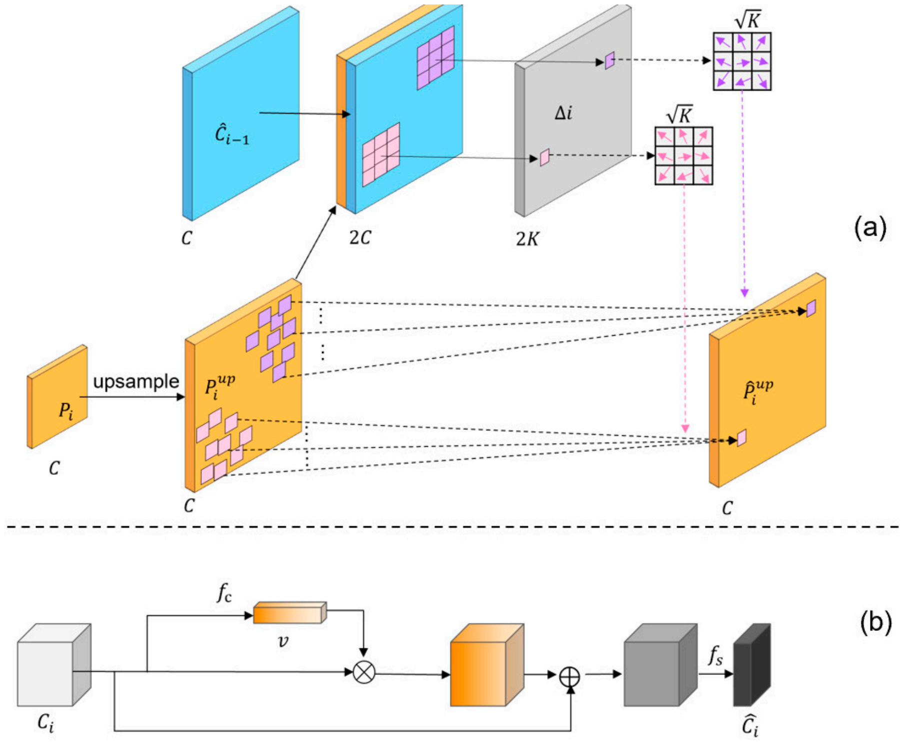

Figure 6. The structure of the FAM and FSM. (**a**) The workflow of the Feature Alignment Module (FAM). Aligning P_i^{up} by learning the spatial position offsets Δ_i of $\hat{C}_{i-1}$ and P_i^{up}; (**b**) the workflow of the Feature Selection Module (FSM).

Figure 6a shows the workflow of the FAM. Prior to feature fusion, the upsampled feature map P_i^{up} is aligned with its reference feature map $\hat{C}_{i-1}$. This alignment involves normalizing the upsampled feature P_i^{up} based on the spatial position information provided by $\hat{C}_{i-1}$. Additionally, K denotes convolutional kernels at K sampled positions, and C represents feature channel numbers. Δ_i represents the offset to be learned for the convolutional kernels. Feature alignment can be mathematically described.

Learn offsets from upsampled and downsampled feature maps as follows:

$$\Delta_i = f_{offset}\left(\left[\hat{C}_{i-1}, P_i^{up}\right]\right), \tag{1}$$

then, apply offsets to the downsampled feature map P_i^{up} for alignment as follows:

$$\hat{P}_i^{up} = f_{align}\left(P_i^{up}, \triangle_i\right), \tag{2}$$

where $\left[\hat{C}_{i-1}, P_i^{up}\right]$ denotes the concatenation of $\hat{C}_{i-1}$ and P_i^{up}, providing spatial disparities between the upsampled and corresponding bottom-up features. $f_{offset}(\cdot)$ represents learning offset from spatial disparities Δ_i, and $f_{align}(\cdot)$ is the function aligning features with learned offsets. $f_{offset}(\cdot)$ and $f_{align}(\cdot)$ are 3 × 3 deformable convolutions [36]. Deformable convolutions adjust convolutional sampling positions based on offsets, aligning features P_i^{up} according to spatial distances between $\hat{C}_{i-1}$ and P_i^{up}.

4.2.2. Feature Selection Module (FSM)

In the traditional FPN, a simple 1 × 1 convolution operation is performed to unify the channel numbers of high-dimensional and low-dimensional features. However, this approach overlooks the significance of different channels, resulting in the loss of important spatial details during channel compression. Moreover, due to the specific nature of virtual try-on tasks, information from different modalities is often concatenated as the input. Effectively utilizing important features while suppressing irrelevant features is a problem that the feature extraction module needs to solve.

We adopt the FSM, which models important features during the feature mapping process. It utilizes global max-pooling to calculate the maximum value for each channel, extracting the most relevant information from each channel. This simultaneously suppresses and recalibrates redundant feature mappings. Figure 6b illustrates the data flow of the FSM.

First, global information is extracted from the input feature map using max-pooling operations to minimize loss. After extracting global information, the important feature construction layer $f_c(\cdot)$ learns the weights of each channel in the input feature map. These weights are represented as a feature importance vector v, indicating the importance of each feature map. The initial input feature map is scaled according to the importance vector, and the scaled feature map is then concatenated with the original feature map in order to generate the resized feature map. This process $f_s(\cdot)$ preserves selectively retains features. The specific workflow of the Feature Selection Module is as follows:

$$v = f_{sig}(f_c(z_{max})), \tag{3}$$

$$\hat{C}_i = f_s(C_i + v * C_i), \tag{4}$$

where v represents the saliency vector of the input feature map after activation; z_{max} is the feature vector obtained after the global max-pooling operation; $f_c(\cdot)$ includes a 1 × 1 convolutional layer; $f_{sig}(\cdot)$ represents the feature activation layer constituted by the sigmoid function; $f_s(\cdot)$ is composed of a 1 × 1 convolutional layer; and C_i and $\hat{C}_i$ refer to the input and output feature maps, respectively.

4.3. Coarse-to-Fine Warping Module (CFWM)

The purpose of the CFWM is to warp the garment to adapt to body posture while preserving garment details. Due to the superior flexibility of flow-based networks compared to the TPS algorithm, they can adapt to complex posture warping. Following the work of FS-VTON [7], we adopt a coarse-to-fine warping structure, composed of subnetworks with different-sized convolutional layers, to utilize multi-scale features extracted by the FEM for appearance flow estimation. Flow-based networks effectively capture distant correspondences between garment and person images, thereby reducing the problem of garment warping misalignment. A specific structure is shown in Figure 7.

The CFWM is composed of n stacked Warping Blocks, and the specific structure of the Warping Block is shown in Figure 7b; each of our Warping Blocks consists of two parts. Specifically, we first use the garment and person features extracted from the last layer of the previous module to generate the global style vector s through the connected operation, which includes the global information, such as the position and structure of the two, and guides the generation of the initial rough appearance flow f'_{k-1}, which is formulated as follows:

$$f'_{k-1} = MConv(W(g_k, U(f_{k-1})), s), \tag{5}$$

where $U(\cdot)$ is the upsampling operation and $W(\cdot)$ is the sampling operation. According to the spatial transformation of the corresponding layer g_k after the upsampling of the appearance flow, we follow the design of the modulation convolution operation $Mconv(\cdot)$

in StyleGan [16], which is able to modulate with the help of global style vectors, and it can deal with the bias of the large distance between the garment and the person to generate the rough appearance flow.

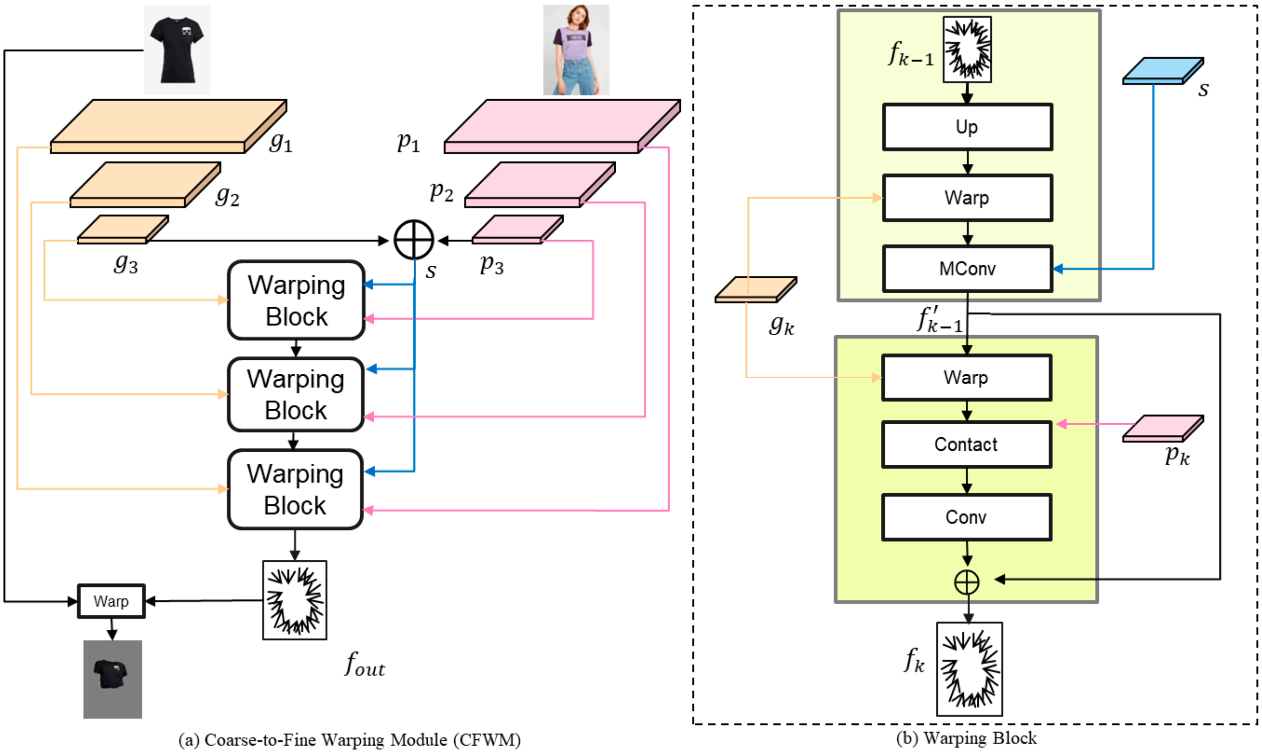

Figure 7. Overview and details of our Coarse-Fine Warping Module (CFWM). (**a**) The structure of the CFWM; (**b**) The specific structure of the Warping Block.

Then, we refine the rough appearance flow with the help of both person and garment features in order to further improve the correspondence between the warped garment and the person. The specific formula is as follows:

$$r_k = W(g_k, f'_{k-1}), \tag{6}$$

$$f_k = Conv\left(r_k \bigoplus p_k\right) \bigoplus f'_{k-1}, \tag{7}$$

where $W(\cdot)$ denotes the sampling operation. After performing a spatial transformation on g_k, it is concatenated with the person feature to obtain r_k. $Conv(\cdot)$ represents convolutional blocks composed of convolutions and Leaky ReLU functions. By inputting r_k together with the person feature p_k into the convolutional blocks, the network is better able to learn the correspondence between the garment and the person. Finally, the refined appearance flow, concatenated with the coarse appearance flow f'_{k-1}, serves as the final output f_k of this Warping Block. Throughout the entire CFWM, we utilize the appearance flow from the final module to perform the warp operation, thereby generating the final deformed garment.

Additionally, to enhance the preservation of garment features, we also optimize this module with second-order smoothness loss [6]. We aim to achieve pixel-to-pixel matching between the source and target garment regions to better estimate geometric transformations and generate realistic results. By implementing pixel-level matching, we can capture subtle details, such as wrinkles and fabric textures, thus rendering a more lifelike garment appearance in the synthesized images. However, due to the density and high degree of freedom of the appearance flow, some undesirable phenomena often occur, such as significant artifacts and incoherent textures. To address these issues, we reference the work of [6,7,28] and introduce a second-order smoothness loss, which helps regularize the estimated flow field, making it spatially smoother and more continuous. This approach reduces unnatural jagged edges and fragmented textures in the images, thereby enhancing the visual quality and realism of the synthesized results. The calculation process is as follows:

$$L_{second} = \sum_{i=1}^{N} \sum_{p} \sum_{\pi \in N_p} Char\left(f_i^{p-\pi} + f_i^{p+\pi} - 2f_i^{p}\right), \tag{8}$$

where f_i^p represents the p-th point on the i-th scale flowchart; N_p is the collection of horizontal, vertical, and diagonal neighborhoods around the p-th point; and $Char(x) = (x^2 + \epsilon^2)^a$ denotes the generalized Charbonnier loss [37]. The generalized Charbonnier loss function is widely used in optical flow estimation [38,39]. The introduction of the power operation reduces its sensitivity to large discrepancies, thereby enhancing the model's robustness to outliers.

4.4. Try-On Generation Module (TGM)

The TGM takes the human image and the warped garment as inputs, fusing them to generate the final try-on result. The goal of the module is to produce a realistic image of the person wearing the target garment. In order to achieve try-on without parsing, the model must generate try-on images without any human or garment parsing. In this method, the try-on image generation module cannot use additional human parsing to create a human-agnostic image that reduces the influence of the original image's garment, which requires the generation module to be able to filter out important features and reduce irrelevant information, thereby minimizing artifact generation. The overall structure of the module is illustrated in Figure 8.

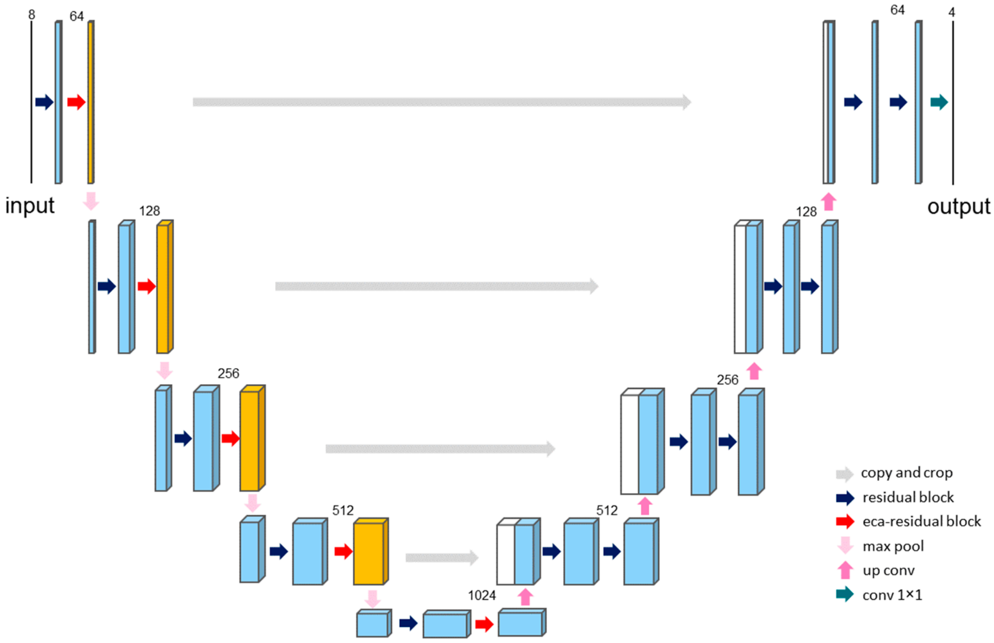

Figure 8. Structure of the TGM.

In this module, we utilize an enhanced U-Net [40] as the backbone architecture for the TGM. We introduce residual structures and the ECA attention mechanism [32] on it. The addition of the ECA mechanism helps alleviate incorrect results due to prior parsing errors or excessive garment deformation, improving the quality of the generated image. It not only preserves the details of the warped garment but also maintains key human body parts, reducing the impact of irrelevant information.

Many studies have shown that attention mechanisms can enhance the overall performance of deep learning methods. The SE module [41] proposed an effective mechanism for attention learning, which first learned channel attention and achieved outstanding performance. However, the SE attention mechanism compresses the input feature maps along the channel dimension, which has a detrimental effect on learning dependencies between channels.

In order to avoid dimensionality reduction and facilitate proper cross-channel interaction, ECA proposes a local cross-channel interaction strategy without dimensionality reduction and a method for adaptive selection of the size of one-dimensional convolutional kernels. With just a few extra parameters, the ECA module achieves significant performance improvements. The structure of the ECA module is shown in Figure 9.

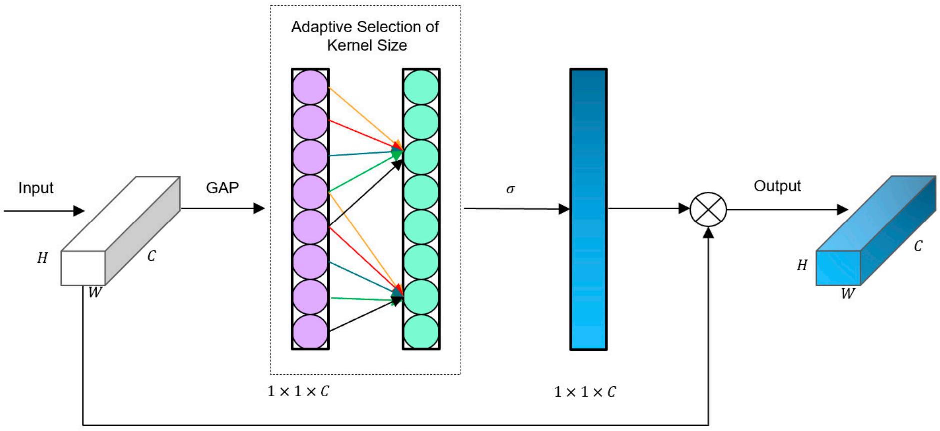

Figure 9. The overall architecture of the ECA module.

ECA first performs global pooling on the input feature layer. Then, ECA captures local cross-channel interactions and obtains the weight for each channel by sigmoid function. Finally, the weights are multiplied by the corresponding elements of the input feature layer to obtain the final output feature layer. The local cross-channel interaction strategy can be implemented through a one-dimensional convolution with a kernel size of k, calculated as follows:

$$w = \sigma(L_k(y)), \tag{9}$$

where w represents the channel weights, σ is the sigmoid function, $L_k(\cdot)$ denotes the one-dimensional convolution, k is the kernel size of L, and y represents the aggregated features. This method involves only k parameters.

We adopt an adaptive method to determine the convolution kernel k, where the size of the kernel k is proportional to the channel dimension C (i.e., there may be a mapping between k and C). Due to the limitations of linear mapping, we employ non-linear mapping instead. Additionally, we know that the channel dimension is typically a power of 2; thus, we use an exponential function with base 2 to represent the non-linear mapping relationship, as shown below:

$$C = \phi(k) = 2^{(\gamma * k - b)}. \tag{10}$$

Once the mapping relationship between the kernel size and channel dimension is determined, we can determine the range of local cross-channel interactions based on the channel dimensions of different feature maps, using different kernel sizes for different channels. The kernel size k can be calculated using the following equation:

$$k = \psi(C) = \left| \frac{log_{2(C)}}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{11}$$

where k is the size of the convolution kernel, C is the number of channels, and $\mid t \mid odd$ represents the nearest odd number to t. We set γ and b as 2 and 1, respectively.

4.5. Parser-Free Framework Based on Knowledge Distillation

4.5.1. Parser-Free Method

To achieve a parser-free method, we employ a framework based on knowledge distillation. The entire framework consists of two parts: a teacher network and a student network, as illustrated in the structure in Figure 4. The input of the person feature extraction module of the teacher network includes various human parsing results. The garment warping module infers the appearance flow between the person and the garment based on previously extracted multi-level features and utilizes this flow to generate a warped garment. Finally, the warped garment, along with the preserved regions on the image, is used as the input to train the try-on generation module, supervised by ground truth. After training the teacher network, the student network takes the output of the teacher network as the input for person images and randomly selects different garment images to generate try-on results. The output of the student network is supervised by the human images input to the teacher network, effectively solving the problem of insufficient paired images under supervision. Additionally, since the input information of the teacher network contains human parsing, resulting in richer learned features, we use channel-wise knowledge distillation to guide the student network in fully learning the rich channel information from the teacher network.

4.5.2. Channel-Wise Distillation (CWD)

Due to the teacher model's input containing various parser-based prior knowledge, we introduce channel-wise knowledge distillation [42]. This method focuses on features from different channels, softly adjusting the activations of corresponding channels between the teacher and student networks. To perform this, we first convert the channel activations into probability distributions and measure their differences using probability distribution metrics, adopting KL divergence in our text. Subsequently, we use a trained teacher model to obtain activation maps for predicting channel-specific masks, enabling the student network to learn useful knowledge from the teacher network.

We denote the feature maps of the teacher network and the student network as y_c^T and y_c^S, respectively. Then, the channel-wise distillation loss can be expressed in general form as follows:

$$L_{cwd} = \varphi\left(\phi\left(y_c^T\right), \phi\left(y_c^S\right)\right), \tag{12}$$

where $\phi(\cdot)$ is used to convert the feature values into probability distributions as follows:

$$\phi(y_c) = \frac{exp\left(\frac{y_{c,i}}{\tau}\right)}{\sum_{i=1}^{W\cdot H} exp\left(\frac{y_{c,i}}{\tau}\right)}, \tag{13}$$

where $c = 1, 2, \ldots, C$ represents the channel; i represents the pixel position within the channel; and τ represents the temperature parameter for distillation, where a higher τ leads to softer output probability distributions, meaning each channel focuses on a larger spatial area. If the number of channels between the teacher and student networks does not match, a 1×1 convolutional layer is used to upsample the number of channels in the student network. This paper uses KL divergence to evaluate the difference between the output channel distributions of the teacher and student. The formula is as follows:

$$\varphi\left(y^T, y^S\right) = \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W\cdot H} \phi\left(y_{c,i}^T\right) \cdot \log\left[\frac{\phi\left(y_{c,i}^T\right)}{\phi\left(y_{c,i}^S\right)}\right]. \tag{14}$$

KL divergence is an asymmetric measure. In the above equation, it can be seen that when $\phi\left(y_{c,i}^T\right)$ is larger, $\phi\left(y_{c,i}^S\right)$ should be as large as $\phi\left(y_{c,i}^T\right)$ to minimize KL divergence. When $\phi\left(y_{c,i}^T\right)$ is smaller, KL divergence pays less attention to minimizing $\phi\left(y_{c,i}^S\right)$. Therefore, the student network can learn significant features from the teacher network, especially when student networks do not have human-agnostic images as inputs.

4.5.3. Loss Function

Due to the fact that the input of parser-based models contains more semantic information compared to parser-less models, we also adopt a scheme based on knowledge distillation learning with a distillation loss to guide the feature extractors of the student network as follows:

$$L_{dis} = \varphi \sum_l \|p_l^t - p_l^s\|_2, \tag{15}$$

$$\varphi = \begin{cases} 1, \; if \|p^t - p^G\|_1 < \|p^s - p^G\|_1 \\ 0, otherwise \end{cases}, \tag{16}$$

where p_l^t and p_l^s represent the feature maps extracted by the teacher network and the student network at the l-th scale, p^t and p^s are the parsing results of the teacher network and the student network, respectively, and p^G is the ground truth of the person image.

The loss function used in the extracting and warping stages is defined as follows:

$$L_w = \lambda_{per} L_{per} + \lambda_1 L_1 + \lambda_{sec} L_{sec} + \lambda_{dis} L_{dis} + \lambda_{cwd} L_{cwd}, \tag{17}$$

where L_{per} is the perceptual loss [43] of the warped garment, L_1 represents the pixel-wise L1 loss of the warped garment, L_{sec} is the second-order smoothness loss of the appearance flow from PFAFN [6], L_{dis} is the distillation loss, and L_{cwd} is the channel-wise distillation loss.

During the training of the generator, the model parameters are optimized by minimizing L_g as follows:

$$L_g = \alpha_{euc} L_{euc} + \alpha_{per} L_{per}, \tag{18}$$

where L_{euc} is the pixel-wise L1 loss and L_{per} is the perceptual loss [43], and the formulas are as follows:

$$L_{euc} = \|I^G - I\|_1, \tag{19}$$

$$L_{per} = \sum_m \|\phi_m(I^G) - \phi_m(I)\|_1, \tag{20}$$

where I^G and I are the generated image and ground truth, respectively, and $\phi_m(\cdot)$ represents the m-th layer feature map in the VGG-19 [44] network pre-trained on ImageNet.

5. Experiments

5.1. Experimental Settings

We conduct our experiments on the VITON [5] and VITON-HD [45] datasets. VITON is the most popular dataset used in image-based virtual try-on tasks. The VITON dataset consists of a training set containing 14,221 pairs of images and a test set containing 2032 pairs of images. Each image pair includes a target garment image and a person image, with both having a resolution of 256 × 192. VITON-HD is the same as VITON, except that the image resolution is 512 × 384.

We train our model on the VITON dataset using the Adam optimizer. We first train the teacher network using the clothes image and the image of the person wearing the clothes. The parsing results are also utilized in this phase. Then, we train the student network. Each model is trained for 100 epochs with an initial learning rate of 5×10^{-4} and is linearly decayed after 50 epochs. The hyperparameters are set as follows: λ_{per} = 1.0, λ_1 = 1.0, λ_{dis} = 0.2, λ_{sec} = 6.0, and λ_{cwd} = 0.8. During testing, the reference person image and the target clothes image are provided as inputs to the student network to generate the output image. Unlike the training phase, additional inputs, such as human parsing results, are not used.

We compare FA-VTON with several methods, including ACGPN [10], PF-AFN [6], SDAFN [30], FS-VTON [7], and DM-VTON [28]. All models are trained on the same VITON dataset by using the official codes provided by the authors to ensure the fairness of the experiments.

We evaluate the similarity between the generated images and ground truth using three widely used metrics: the Structural Similarity Index (SSIM) [46], Learned Perceptual Image Patch Similarity (LPIPS) [47], and Fréchet Inception Distance (FID) [48]. The SSIM and LPIPS are used for paired images, while FID is used for unpaired images.

5.2. Qualitative Results

Figure 10 illustrates the qualitative comparison results of FA-VTON with the latest baselines on the VITON-HD dataset.

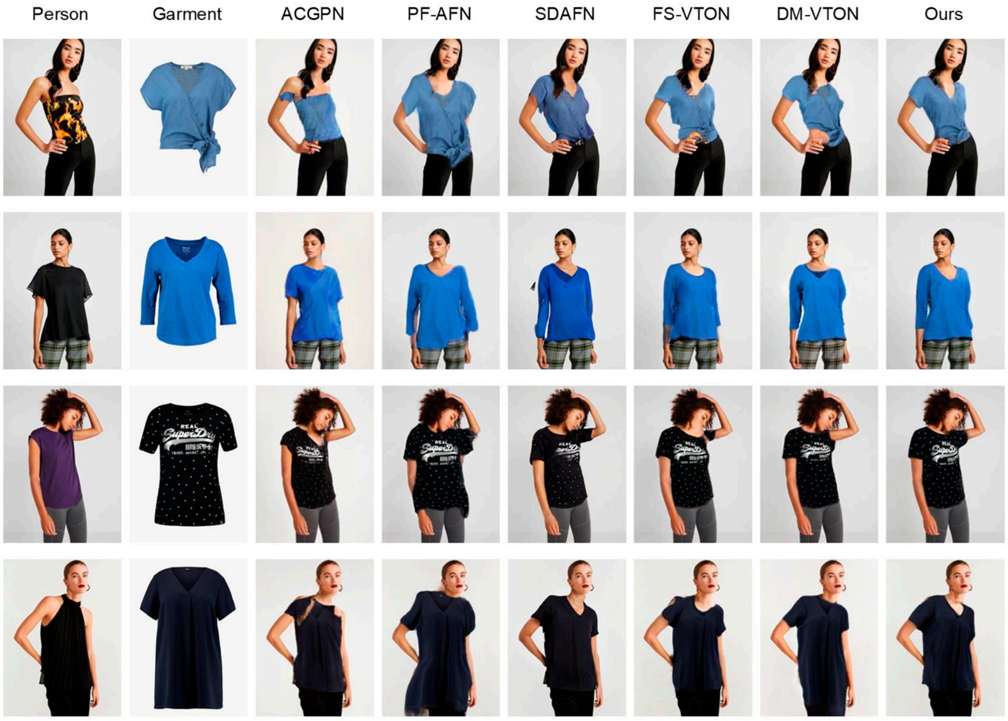

Figure 10. Qualitative results from different models (ACGPN, PF-AFN, SDAFN, DM-VTON, and ours) on the VITON-HD testing dataset.

In the first row, our goal is to correctly distinguish between the front and back parts of the garment to avoid retaining excess fabric at the back. Previous models tended to retain excess fabric, especially around the neckline and hem. By aligning features in the feature extraction stage, we can accurately identify the boundaries of the garment, providing correct guidance for the subsequent garment warping module to remove the excess parts.

In the second row, our objective is to achieve smooth deformation at the edges of the garment, such as at the cuffs, and generate arms wearing clothing with a realistic appearance. Blurry regions for garment replacement often exacerbate errors in deformation refinement, leading to excessive deformation at the edges, particularly at the cuffs and shoulders. Artifacts often appear at the junction between the garment and the skin.

In the third row, we aim to show that FA-VTON can adapt to complex poses and deform the garment to reasonable positions. Previous methods were sensitive to misinterpreting the human body due to the influence of the original image, resulting in unreasonable garment deformation and body part generation during synthesis. However, our model, provided with a robust feature extractor, can accurately identify the regions for garment replacement, guiding the garment to warp correctly. Additionally, because of the attention mechanism added to our generation module for weighted feature processing, we can effectively suppress unimportant features, thus preserving complex patterns, like clothing textures, and generating more natural-looking images.

In the fourth row, the garment patterns in the try-on images generated by our model do not retain excessive deformation, and they naturally blend with the lower garment. Furthermore, the edges are smooth, and the wrinkles are realistic.

Additionally, we also present some try-on results, especially those with complex neckline structures, as shown in Figure 11.

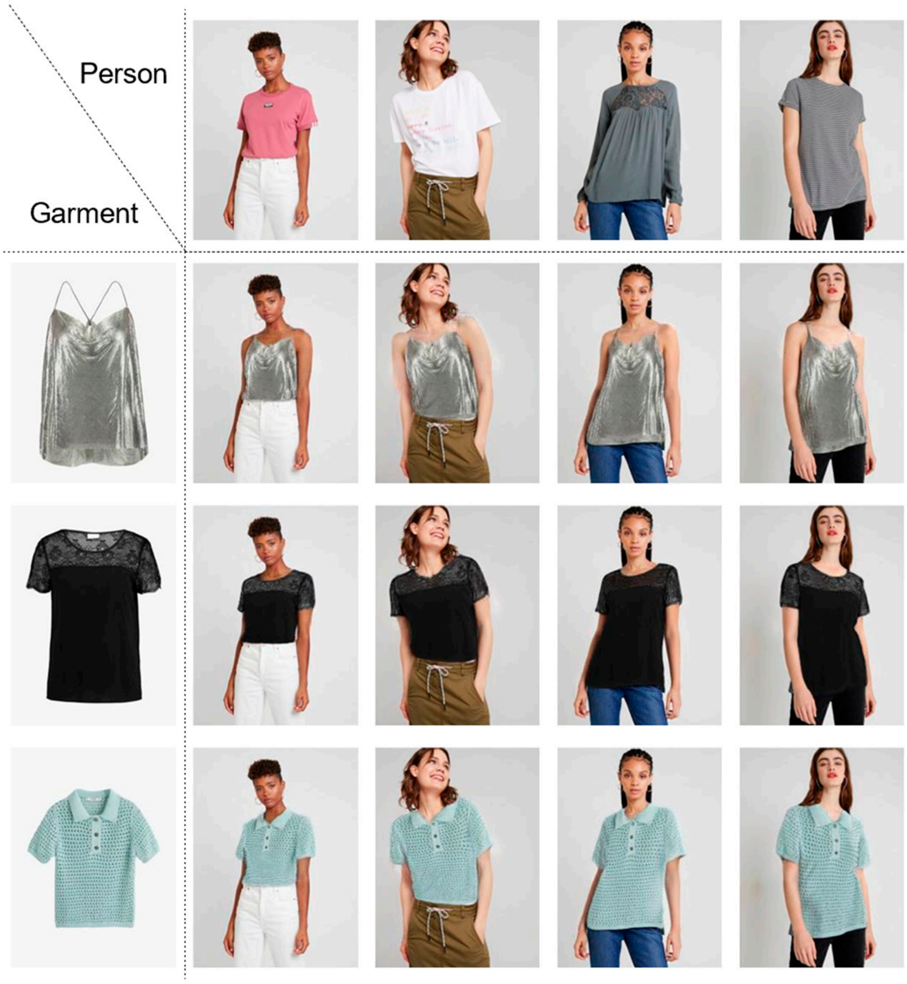

Figure 11. Try-on results of complex clothes.

5.3. *Quantitative Results*

We evaluated the performance of our proposed method on the VITON and VITON-HD datasets. Table 1 compares our method with several baseline methods, including ACGPN, PF-AFN, SDAFN, FS-VTON, and DM-VTON. Our method achieved excellent scores of 0.86, 8.72, and 0.184 for the SSIM, FID, and LPIPS on the VITON dataset and achieved excellent scores of 0.85, 9.53, and 0.096 on the VITON-HD dataset. These scores outperform those of other methods, showing the effectiveness of our proposed virtual try-on model. In summary, our model achieves outstanding quantitative results on the VITON dataset.

Table 1. Quantitative results of different models on the VITON and VITON-HD datasets. For SSIM, the higher is the better. For FID and LPIPS, the lower is the better.

Dataset	VITON			VITON-HD		
Method	**SSIM ↑**	**FID ↓**	**LPIPS ↓**	**SSIM ↑**	**FID ↓**	**LPIPS ↓**
ACGPN	0.69	16.64	0.226	0.78	14.99	0.170
PF-AFN	0.79	10.09	0.213	0.69	25.44	0.229
SDAFN	0.78	9.42	0.228	0.82	9.97	0.113
FS-VTON	0.85	8.89	0.200	0.83	10.00	0.102
DM-VTON	0.81	10.57	0.213	0.82	11.81	0.125
Ours	**0.86**	**8.72**	**0.184**	**0.85**	**9.53**	**0.096**

5.4. Ablation Study

We conducted ablation experiments to validate the effectiveness of our designed feature extraction module, the try-on image generation module, and channel-wise knowledge distillation.

We compared the performance with and without the FASM in both the clothing feature extraction module and the human body feature extraction module, demonstrating how the features from the first stage gradually influence the final try-on results. We visualized the feature maps of the last layer and obtained grayscale images by compressing the color channels. As shown in Figure 12, the first two columns are the model inputs, and columns 3 to 8 show the comparisons with and without the FASM. In the first row, the impact of the human body feature extraction map on garment warping and try-on image generation is explained. It can be seen that without adding the FASM, the model fails to learn the contour information of the human body, especially around the right shoulder, resulting in excessive deformations at the edges by the subsequent garment warping module. The coarse-to-fine warping module aggravates the erroneous effects, resulting in excessively deformed garments in the final try-on image. In contrast, the model using the FASM can extract boundary information well, providing more reasonable information for subsequent warping and generating more realistic try-on images. The second row in Figure 12 illustrates the impact of the garment feature extraction map on garment warping and try-on image generation. It is evident that with the FASM, the model can effectively learn the boundaries of the garment. Due to the irregularities in garment design, their layout images may reveal fabric areas that are unnecessary for the try-on images, posing a challenge to whether the model can correctly retain the corresponding garment areas. The feature maps generated by the traditional feature extractor cannot distinguish between the front and back of the garment, while our model can effectively identify the garment shapes and guide subsequent garment warping.

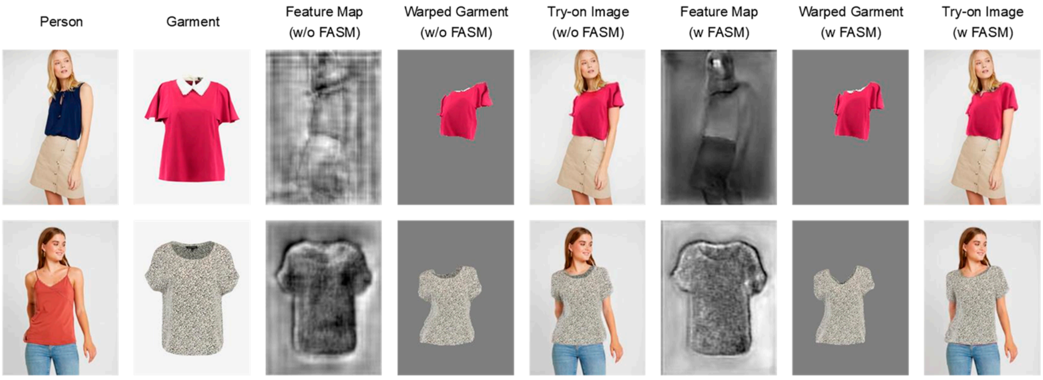

Figure 12. The ablation study of the FASM.

Furthermore, we specifically decomposed the FASM for qualitative comparisons. As shown in Table 2, we evaluated three metrics of the final generated try-on images when only adding the FSM, only adding the FAM, and adding both parts. It can be observed that the image quality is higher when both modules are added compared to when only one module is added.

Table 2. Comparing the results with only the FAM, FSM, and FAM + FSM.

	SSIM ↑	FID ↓	LPIPS ↓
FAM	0.857	8.86	0.193
FSM	0.855	8.99	0.196
FAM + FSM	**0.860**	**8.82**	**0.189**

We also conducted ablation experiments on the ECA module. As shown in Figure 13, we can observe that adding the ECA module can make the synthesized images more realistic, while also mitigating, to some extent, the unreasonable results caused by deformation errors. As shown in Table 3, after adding the ECA module, the two metrics—FID and LPIPS—are all superior to those of the model without the ECA module.

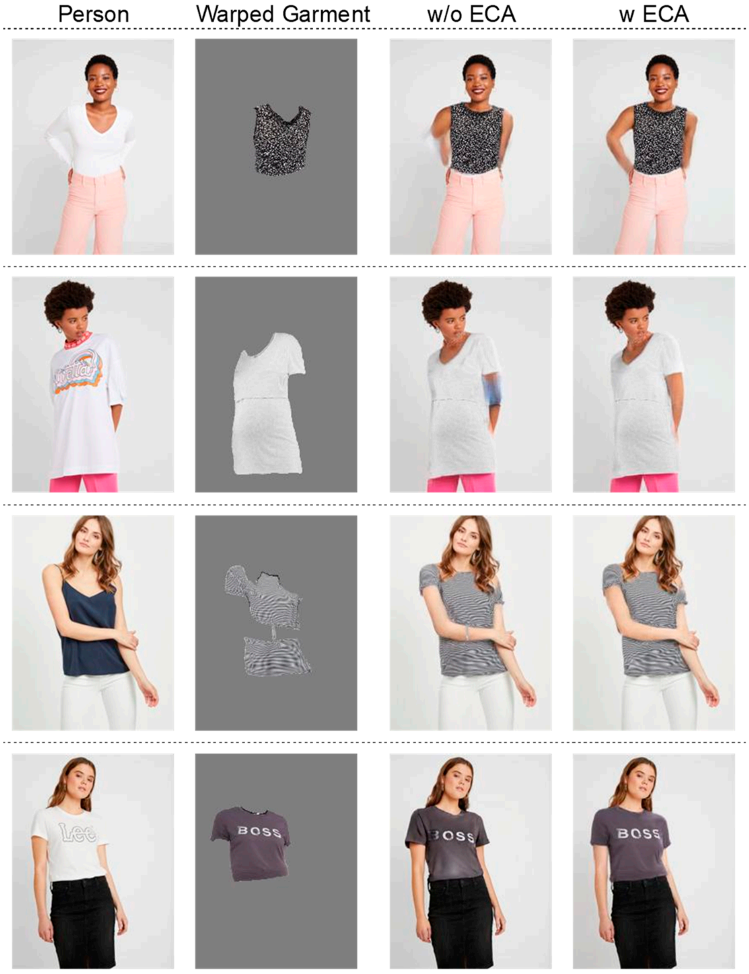

Figure 13. The ablation study of ECA.

Table 3. The ablation quantitative results of ECA.

	SSIM ↑	FID ↓	LPIPS ↓
w/o ECA	0.859	8.79	0.191
w ECA	0.856	**8.72**	**0.187**

Finally, we conduct ablation experiments on the CWD loss function. The CWD loss function helps the student network effectively utilize the teacher network's knowledge based on parsers. Since the input of the student model retains the original garment, under the guidance of CWD, the student network focuses on learning parsing representations unrelated to the original garment, such as the teacher network's prior knowledge of human pose key points. As shown in Figure 14, the student network is able to focus on learning representation knowledge unrelated to the original garment, reducing interference from the original image.

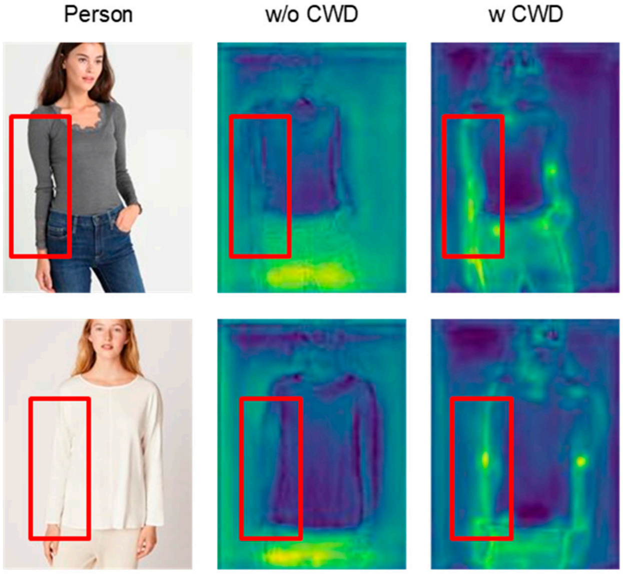

Figure 14. The ablation study of CWD.

6. Conclusions

We first analyzed the impact of feature alignment on try-on image generation. Subsequently, we proposed a novel parser-free model, FA-VTON, which achieves more proper garment warping by aligning upsampled features and filtering important features. Next, we introduced an ECA attention mechanism into the last module to generate more realistic try-on images. Finally, we achieved parser-free try-on modeling based on channel-wise distillation to help the student network learn rich semantic knowledge from the teacher network. Experimental results show that our model performs better in both qualitative and quantitative aspects.

Author Contributions: Conceptualization, Y.W., N.D. and L.Y.; methodology, N.D.; validation, N.D.; formal analysis, N.D.; resources, Y.W. and L.Y.; data curation, N.D.; writing—original draft preparation, N.D.; writing—review and editing, Y.W. and L.Y.; visualization, N.D.; supervision, Y.W. and L.Y.; project administration, Y.W. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available from https://github.com/shadow2496/VITON-HD, accessed on 15 March 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bhatnagar, B.L.; Tiwari, G.; Theobalt, C.; Pons-Moll, G. Multi-garment net: Learning to dress 3d people from images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5420–5430.
2. Mir, A.; Alldieck, T.; Pons-Moll, G. Learning to transfer texture from clothing images to 3d humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7023–7034.
3. Saito, S.; Simon, T.; Saragih, J.; Joo, H. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 84–93.
4. Han, X.; Hu, X.; Huang, W.; Scott, M.R. Clothflow: A flow-based model for clothed person generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10471–10480.
5. Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7543–7552.
6. Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; Luo, P. Parser-free virtual try-on via distilling appearance flows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8485–8493.
7. He, S.; Song, Y.-Z.; Xiang, T. Style-based global appearance flow for virtual try-on. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3470–3479.
8. Issenhuth, T.; Mary, J.; Calauzenes, C. Do not mask what you do not need to mask: A parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 619–635.
9. Lee, S.; Gu, G.; Park, S.; Choi, S.; Choo, J. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 204–219.
10. Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L.; Yang, M. Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 589–604.
11. Ge, C.; Song, Y.; Ge, Y.; Yang, H.; Liu, W.; Luo, P. Disentangled cycle consistency for highly-realistic virtual try-on. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16928–16937.
12. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* **2015**, *34*, 248. [CrossRef]
13. Jetchev, N.; Bergmann, U. The conditional analogy gan: Swapping fashion articles on people images. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2287–2292.
14. Duchon, J. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Proceedings of the Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*; Springer: Berlin/Heidelberg, Germany, 1977; pp. 85–100.
15. Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; Efros, A.A. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part IV 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 286–301.
16. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
17. Lin, C.; Li, Z.; Zhou, S.; Hu, S.; Zhang, J.; Luo, L.; Zhang, J.; Huang, L.; He, Y. Rmgn: A regional mask guided network for parser-free virtual try-on. *arXiv* **2022**, arXiv:2204.11258.
18. Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; Wen, F. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18381–18391.
19. El Ogri, O.; Daoui, A.; Yamni, M.; Karmouni, H.; Sayyouri, M.; Qjidaa, H. New set of fractional-order generalized Laguerre moment invariants for pattern recognition. *Multimedia Tools Appl.* **2020**, *79*, 23261–23294. [CrossRef]
20. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
21. Yamni, M.; Karmouni, H.; Sayyouri, M.; Qjidaa, H. Image watermarking using separable fractional moments of Charlier–Meixner. *J. Franklin Inst.* **2021**, *358*, 2535–2560. [CrossRef]
22. Karmouni, H.; Jahid, T.; El Affar, I.; Sayyouri, M.; Hmimid, A.; Qjidaa, H.; Rezzouk, A. Image analysis using separable Krawtchouk-Tchebichef's moments. In Proceedings of the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fes, Morocco, 22–24 May 2017; pp. 1–5.
23. Karmouni, H.; Jahid, T.; Hmimid, A.; Sayyouri, M.; Qjidaa, H. Fast computation of inverse Meixner moments transform using Clenshaw's formula. *Multimedia Tools Appl.* **2019**, *78*, 31245–31265. [CrossRef]
24. Yang, X.; Ding, C.; Hong, Z.; Huang, J.; Tao, J.; Xu, X. Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-on. *arXiv* **2024**, arXiv:2404.01089.
25. Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; Cucchiara, R. LaDI-VTON: Latent diffusion textual-inversion enhanced virtual try-on. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 8580–8589.

26. Kim, J.; Gu, G.; Park, M.; Park, S.; Choo, J. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-on. *arXiv* **2023**, arXiv:2312.01725.
27. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
28. Nguyen-Ngoc, K.-N.; Phan-Nguyen, T.-T.; Le, K.-D.; Nguyen, T.V.; Tran, M.-T.; Le, T.-N. DM-VTON: Distilled Mobile Real-time Virtual Try-on. In Proceedings of the 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Sydney, Australia, 16–20 October 2023; pp. 695–700.
29. Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; Liang, X. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23550–23559.
30. Bai, S.; Zhou, H.; Li, Z.; Zhou, C.; Yang, H. Single stage virtual try-on via deformable attention flows. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 409–425.
31. Huang, S.; Lu, Z.; Cheng, R.; He, C. Fapn: Feature-aligned pyramid network for dense image prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 864–873.
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
33. Güler, R.A.; Neverova, N.; Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7297–7306.
34. Yan, K.; Gao, T.; Zhang, H.; Xie, C. Linking garment with person via semantically associated landmarks for virtual try-on. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17194–17204.
35. Feng, R.; Ma, C.; Shen, C.; Gao, X.; Liu, Z.; Li, X.; Ou, K.; Zhao, D.; Zha, Z.-J. Weakly supervised high-fidelity clothing model generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3440–3449.
36. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
37. Sun, D.; Roth, S.; Black, M.J. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *Int. J. Comput. Vis.* **2014**, *106*, 115–137. [CrossRef]
38. Janai, J.; Guney, F.; Ranjan, A.; Black, M.; Geiger, A. Unsupervised learning of multi-frame optical flow with occlusions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 690–706.
39. Jin, X.; Wu, L.; Shen, G.; Chen, Y.; Chen, J.; Koo, J.; Hahm, C.-h. Enhanced bi-directional motion estimation for video frame interpolation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5049–5057.
40. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
42. Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; Shen, C. Channel-wise knowledge distillation for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5311–5320.
43. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Choi, S.; Park, S.; Lee, M.; Choo, J. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
46. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
47. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
48. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 25–34.