IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2023, accepted 11 August 2023, date of publication 18 August 2023, date of current version 23 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3306449

RESEARCH ARTICLE

LC-VTON: Length Controllable Virtual Try-On Network

JINLIANG YAO[1,2] **AND HAONAN ZHENG**[1]

[1]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[2]Zhejiang Key Laboratory of Brain–Machine Collaborative Intelligence, Hangzhou 310018, China

Corresponding author: Haonan Zheng (212050264@hdu.edu.cn)

This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2019C03127, and in part by the Zhejiang Provincial Basic Public Welfare Research Project under Grant LGG20F020012.

ABSTRACT Image-based virtual try-on provides customers with convenient online clothes selections by transferring garments onto a reference person. Despite the emergence of several solutions to generate photo-realistic images and adapt to complex poses, controlling clothing length remains a challenge. We argue that the clothing reconstruction did not consider clothing length information, which results in clothing length being uncontrollable in most virtual try-on methods. To overcome this limitation, a novel clothing-agnostic person representation is proposed, which eliminates clothing information and quantifies clothing length as a numerical value. A new segmentation generator is designed to predict try-on segmentation maps of any length conditioned on this representation. Moreover, we correct two inaccurate labels, which enables our model to utilize clothing length control to generate a wider range of garment interactions in images, such as the top tucked into or worn over the bottom, as well as the top and bottom worn separately without intersecting. Extensive experiments demonstrate that our method achieves the goal of continuous clothing length control and generates photo-realistic images with fine details that outperform most baseline methods in terms of quantitative and qualitative metrics.

INDEX TERMS Virtual try-on, clothing length controllable, conditional semantic generation, generative adversarial network.

I. INTRODUCTION

In recent years, customers' growing reliance on online shopping has increased their need to try clothes on virtually. A virtual try-on task has been proposed to facilitate clothing selection and augment the convenience of the online shopping experience. Still, it remains a significant challenge for virtual try-on to control clothing length while ensuring realistic results.

Existing virtual try-on techniques can be classified into 2D image-based and 3D model-based methods. 3D model-based methods [1], [2], [3] rely on 3D measurement data to reconstruct a 3D model and render multiple output images onto the model body with precise geometric transformations. However, modeling for characters and clothing requires extensive 3D data collection, lengthy rendering times, and expensive computing devices.

Han et al. [4] first proposed a two-stage virtual try-on network based on 2D images, which employed clothing-agnostic person representation to generate a coarse try-on result and refined it with warped in-shop clothes. Since then, various solutions have been proposed to improve performance in different aspects. Most works [5], [6], [7], [8] focused on improving deformation methods to reduce misalignment, which can retain more textures and reduce artifacts. Several works [6], [7], [8], [9] have improved clothing-agnostic person representations to maintain parts irrelevant to try-on, leading to better adaptation to complex poses and retention of realistic details. However, current research can generate photo-realistic results but overlooks the possibility of combining garment modification tasks, such as clothing length control, to enhance user functionality. Some methods

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

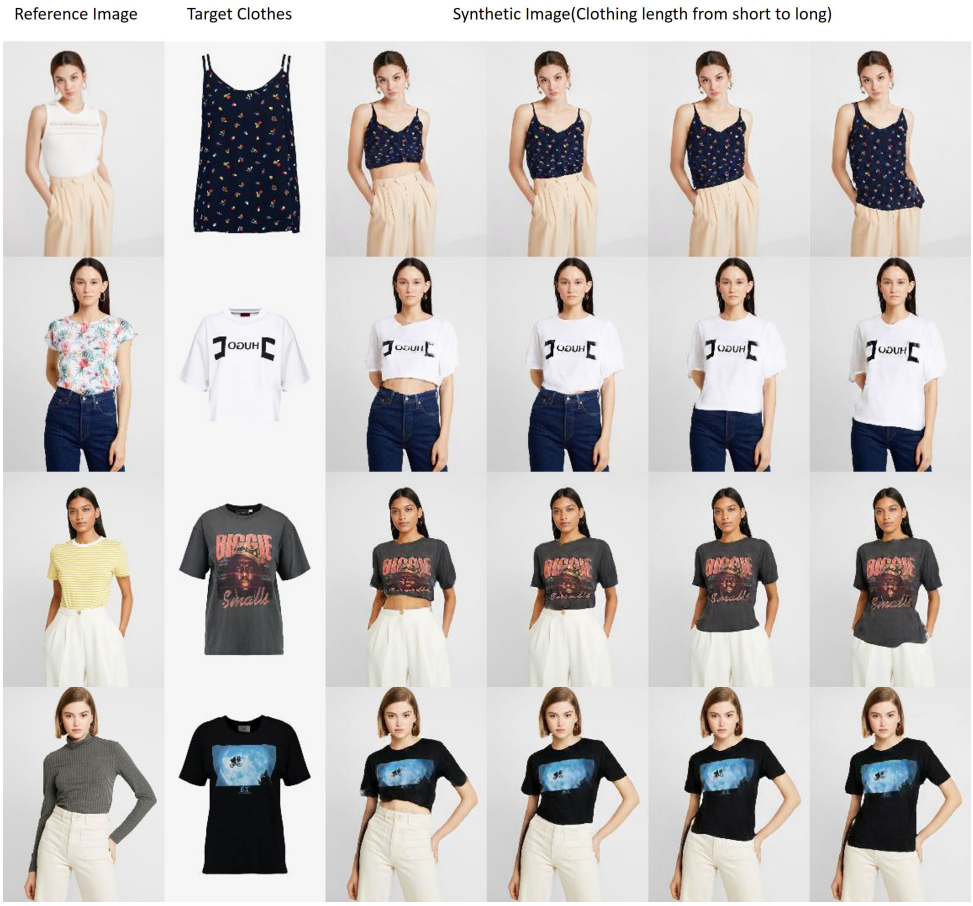

FIGURE 1. Clothing length controllable visual results generated by LC-VTON.

that have incorporated this functionality are based on texture transfer [10] and underwear model reconstruction [11]. In their research, the authors have all mentioned that these methods still have deficiencies in terms of photorealism.

However, controlling clothing length is essential in virtual try-on as it produces different garment interactions (e.g., wearing a top tucked into or over the bottom or wearing a top and bottom separately without intersecting, resulting in the belly being naked) [10]. We argue that previous methods have been limited for two reasons. Firstly, incomplete elimination of clothing information leads to the reproduction of clothing length. With the introduction of the segmentation map, most previous approaches train a try-on segmentation map generator to reconstruct clothing semantics based on the clothing-agnostic person segmentation map. The clothing-agnostic person segmentation map is intended to eliminate clothing information. However, these approaches only remove the clothing shape and omit clothing length information, leading to its reproduction instead of reconstruction. Second, the clothing length information is implicitly encoded in the image, making it difficult to change as a separate condition. As illustrated in Fig. 2, the reference displays a common garment interaction in datasets where the top and bottom intersect. These clothing-agnostic person

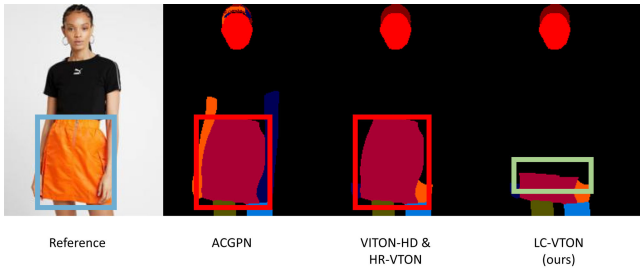

FIGURE 2. We compare clothing-agnostic person segmentation maps generated by different methods. The blue box indicates the bottom part of the reference image. In the red box, the corresponding clothing-agnostic person segmentation map retains the full bottom semantics. In contrast, the green box shows our clothing-agnostic segmentation map with clothing length information removed.

segmentation maps try various methods to remove clothing shapes, but all keep the entire bottom. After training, the prediction of the generator reproduces the entire bottom semantics, which makes the bottom fixed. As a result, generators lose the ability to control clothing length, leading to only one look being achieved.

To overcome this limitation, we propose a novel Length Controllable Virtual Try-On Network (LC-VTON). Our

approach introduces a novel clothing-agnostic person representation that combines graphic and numerical elements and eliminates clothing information. We quantify clothing length as a numerical value that allows for continuous length control by users. We then predict a target segmentation map based on the desired clothing item and length. Considering the case wearing a top and bottom separately without intersecting, we employ the Context Incompatibility Handling module to ensure the target segmentation map is compatible with the reference. Next, we deform the clothes to align with the target segmentation map. Finally, we generate the clothing-agnostic person image corresponding to the target segmentation map and synthesize the try-on image by fusing all inputs.

We summarize our contributions as follows:

- We propose LC-VTON, a novel image-based virtual try-on network that enables the generation of multiple try-on effects by controlling clothing length.
- We introduce a novel numerical-graphical clothing-agnostic person representation that eliminates clothing information and quantifies clothing length as a numerical value, providing continuous control of clothing length.
- We correct two inaccurate labels in VITON-HD: We complete the bottom label by incorporating belt content, and we add the 'belly' label into the human presentation.
- Extensive experimental results show that the method outperforms most existing methods in qualitative and quantitative terms.

II. RELATED WORK

A. CONDITIONAL IMAGE GENERATION

Conditional Generative Adversarial Networks (CGANs) are a variation of Generative Adversarial Networks (GANs), which generate particular data based on additional condition data. The condition data can be class labels [12], [13], text [14], [15], images, and attributes [16]. Previous virtual try-on networks generated try-on images using clothing as the sole variable condition. In this paper, we introduce a new variable condition, the clothing length value, which enables users to control clothing length while trying on various clothing items, thus enhancing the try-on experience.

B. TOWARDS PHOTO-REALISTIC IMAGE-BASED VIRTUAL TRY-ON

Since the proposal of 2D image-based virtual try-on methods, generating photo-realistic images has been a primary objective. Han et al. [4] first proposed a two-stage framework named VITON to solve the 2D image-based virtual try-on task. VITON built a sub-task to produce a coarse result and warped clothes and then generated a composition mask to refine the coarse result. Following VITON, CP-VTON [5] refined the framework by introducing a neural network that could learn TPS parameters. Benefiting from this deformation network, CP-VTON could generate try-on results directly by fusing warped clothes and clothing-agnostic person representations. VTNFP [9] employed a segmentation map prediction to improve performance in complex person pose situations where body parts and clothing intersect. To further enhance fine details and increase perceptual quality, ACGPN [6] composited a segmentation map preserving non-target body parts, retaining the details irrelevant to clothing. Moreover, they used a constrained TPS transformation to prevent clothing from being over-distorted. As the demand for high-resolution virtual try-on grows, VITON-HD [7] proposed a novel clothing-agnostic person representation and employed alignment-aware segment normalization to address the issue of misalignment, achieving high resolutions virtual try-on results. ClothFlow [17] was proposed to tackle the problem of pose-guided virtual try-on. It predicted an optical flow map to warp the source clothes, which can also be applied to the 2D image-based virtual try-on. Based on this deformation, HR-VTON [8] constructed a try-on condition generator that generated the segmentation map and deformed clothes simultaneously. HR-VTON prevented images from pixel-squeezing in the try-on task of complex poses by obtaining a more suitable clothes transformation through conditional alignment. However, these methods focus only on generating photo-realistic images, ignoring the significant need for users to control clothing length. The length-controllable virtual try-on technology offers users a broader range of dressing options for clothing, along with increased stylistic potential resulting from variations in clothing length.

C. GARMENT CONTROLLABLE VIRTUAL TRY-ON

Garment-controllable virtual try-on is a sub-field of virtual try-on that incorporates garment modification tasks to enhance functionality. As presented in [18], a virtual try-on task should produce multiple results with a single initial input by allowing users to control garment attributes such as clothing length. Although some methods have already provided such functionality, they are mainly based on clothing texture transfer or model reconstruction, leading to their respective limitations in photorealism. DIOR [10] produced various dress effects, such as wearing a top tucked into the bottom or over it, by simulating the order in which people put on clothes. However, this approach was based on texture transfer and could not reproduce realistic clothing patterns. The most closely related work is SC-VTON [11], which proposed a new task to render deformed clothing to the reconstructed underwear model, firstly realizing shape-controllable virtual try-on. Although SC-VTON achieved clothing shape control, the underwear model reconstruction network was trained on pseudo-labeled pairs, which limits its performance on real data. Furthermore, users could not use their own bottoms as references since the reference bottom was discarded in the reconstruction. Compared with underwear model reconstruction, the methods based on target segmentation map prediction can produce more photo-realistic results.

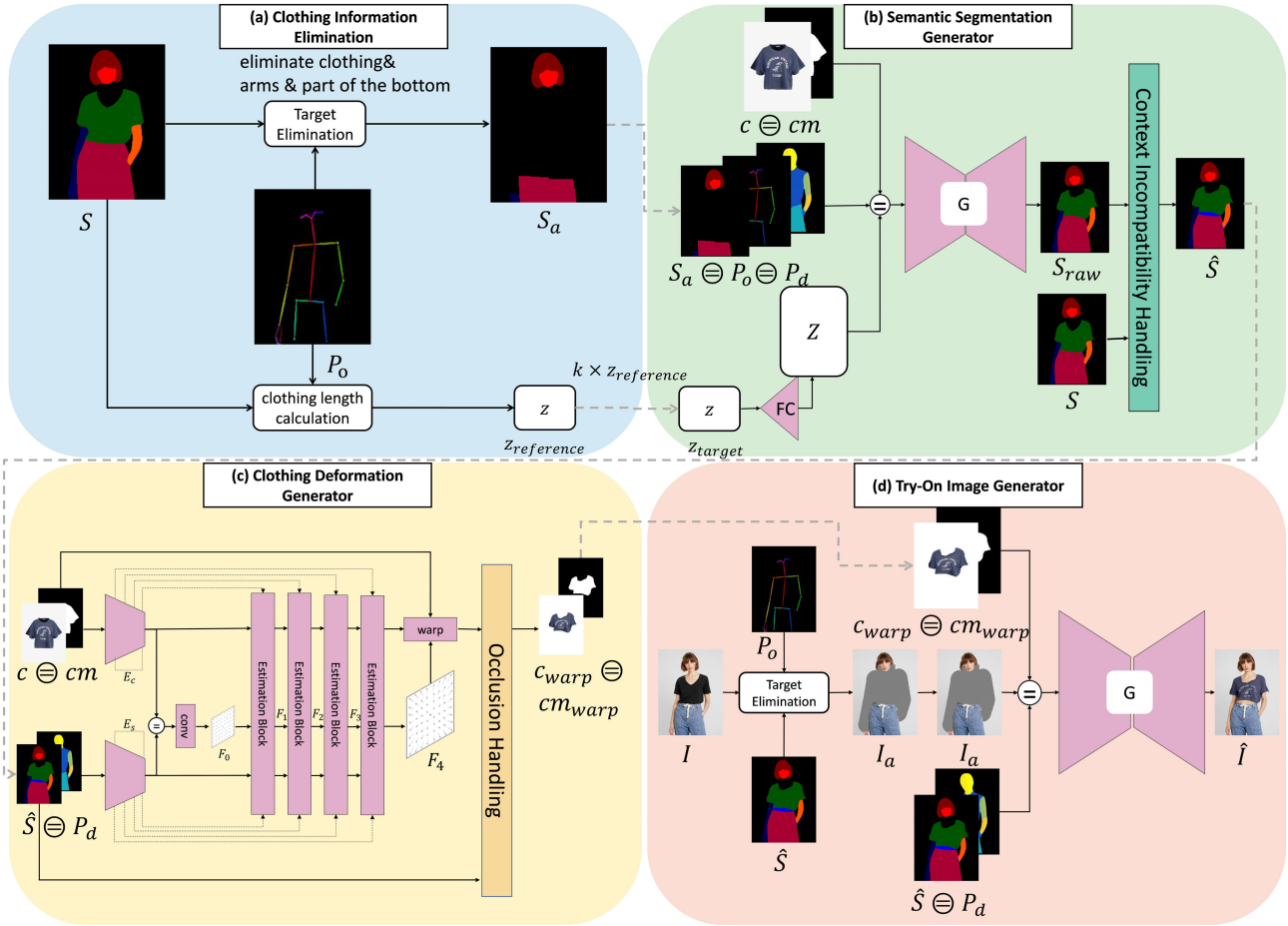

FIGURE 3. The LC-VTON architecture consists of four stages. We illustrate the model process in the case where the top and bottom are worn separately without intersecting. (a) Firstly, given a reference segmentation map S, we utilize pose map P_o to remove clothing information from S, resulting in a clothing-agnostic person segmentation S_a. Simultaneously, we use S and P_o to calculate clothing length value $z_{reference}$ and multiply a small coefficient k by $z_{reference}$ to derive shorter clothing length value z_{target}. (b) Next, the Semantic Segmentation Generator (SSG) predicts a raw layout S_{raw} using (c, c_m, S_a, P_o, P_d, z_{target}). The Context Incompatibility Handling module uses S to process S_{raw} and obtain the target layout $\hat{S}$. (c) Then, the Clothing Deformation Generator (CDG) aligns c with $\hat{S}$. (d) Lastly, the Try-on Image Generator (TOIG) synthesizes the final image $\hat{I}$ by fusing the previous outputs and clothing-agnostic person image I_a.

III. PROPOSED MODEL

In Fig. 3, we illustrate that the objective of LC-VTON is to generate the image of a person $\hat{I} \in \mathbb{R}^{3\times H\times W}$ wearing a target clothing item $c \in \mathbb{R}^{3\times H\times W}$ with clothing length varying from the given reference image $I \in \mathbb{R}^{3\times H\times W}$ while preserving the pose and body shape. However, training directly with triplets $(I, c, \hat{I})$ is challenging since collecting such triplets in practice is difficult. To address this issue, we employ a clothing-agnostic person representation that removes clothing information and allows us to reconstruct I where the original c is worn on the person already.

To enable the model to learn the ability of clothing length control during the reconstruction, we use a clothing-agnostic person representation that eliminates clothing length and quantifies it as a numerical value (Section III-A). The Semantic Segmentation Generator (SSG) first generates the raw segmentation map S_{raw} and then processes it to obtain the target segmentation map $\hat{S}$ using a Context Incompatibility Handling module that leverages the semantic context information from S (Section III-B). After being fed with $\hat{S}$, the Clothing Deformation Generator (CDG), which cascades elimination blocks and an occlusion handling module, predicts an optical flow map for deforming c. Finally, the Try-On Image Generator (TOIG) generates the clothing-agnostic person image I_a based on S and P_o and then synthesizes the final try-on result by fusing all previous outputs (Section III-D).

A. NUMERICAL-GRAPHICAL CLOTHING-AGNOSTIC PERSON REPRESENTATION

VITON-HD used a processing method of target elimination to obtain clothing-agnostic person representation, aiming to remove the clothing shape and preserve the body parts that need to be reproduced. HR-VTON extended this representation to the virtual try-on task with complex poses. This representation cleanly leaves out the clothing information and is competent for most virtual try-on tasks. Still, they

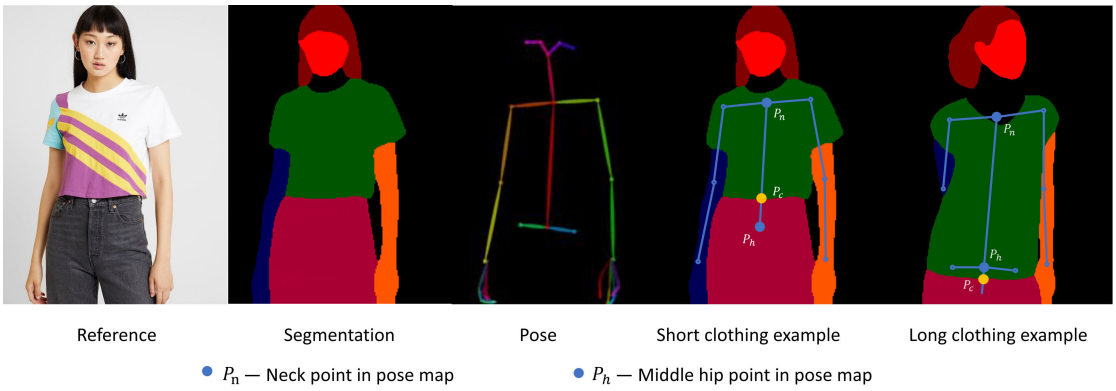

FIGURE 4. We provide two examples of calculating clothing length. The neck point P_n and the middle hip point P_h are marked with solid blue dots, and line $L_{P_nP_h}$ intersects the upper boundary of the bottom at point P_c, which is marked with a solid yellow dot. In the short clothing example, P_c is located between P_n and P_h, while in the long clothing example, P_c lies on the extension line of P_n and P_h.

kept the full bottom semantics, which implicitly specifies clothing length, rendering it useless for our task. To address the issue, we propose a novel person representation consisting of several person representation images and a numerical value that represents clothing length.

1) CLOTHING-AGNOSTIC PERSON SEGMENTATION MAP

The clothing-agnostic person segmentation map $S_a \in \mathbb{L}^{H\times W}$ illustrates the layout of an unclothed person and is used to predict the try-on segmentation map. Inspired by the reconstruction, we remove the upper part of the bottom semantics and reconstruct it with the clothing length value to achieve the purpose of controllable clothing length. We remove the clothing shape in a similar way to VITON-HD. Differently, we use P_o to further remove the upper part of the bottom in S and retain the lower part of the bottom in S_a to embody the bottom type. By removing the upper part of the bottom semantics, the clothing length information is eliminated, as the generator loses the basis for predicting clothing length.

2) CLOTHING LENGTH VALUE

To reconstruct the bottom semantics, we propose a quantized clothing length value $z \in \mathbb{R}$, which can represent clothing length and achieve continuous control of clothing length. As shown in Fig. 4, we map the neck key point $P_n(x_n, y_n)$ and hip key point $P_h(x_h, y_h)$ from the pose map $P_o \in \mathbb{R}^{3\times H\times W}$ onto the segmentation map $S \in \mathbb{L}^{H\times W}$ and connect them with a line $L_{P_nP_h}$ that intersects the upper boundary of the bottom at a point we refer to as $P_c(x_c, y_c)$. We take P_n and P_h as reference points and use the position of P_c in this reference system to reflect clothing length information. It works because a person who tries on various lengths of clothes uses the same pose map. The coordinates of P_c vary with the change of clothing length, while P_n and P_h remain fixed. We conclude two qualitative patterns from the example images: P_c is between P_n and P_h in short clothing, while in long clothing, P_c is on the extension of the line $L_{P_nP_h}$.

To further achieve continuous control of clothing length, we use a signed line segment ratio function to quantize clothing length as z, which can be formulated in coordinates as:

$$z = \frac{y_c - y_h}{y_h - y_n} \tag{1}$$

In the case of short clothing, z takes on a negative value, while in the case of long clothing, z is positive. To encode more information, we substitute P_n and P_h with the entire pose map P_o. As a result, the combination of z and P_o is used to capture the clothing length information.

B. SEMANTIC SEGMENTATION GENERATOR (SSG)

Given (S_a, P_o, P_d, z) and (c, cm), the Semantic Segmentation Generator(SSG) predicts the try-on segmentation map $\hat{S}$ that separates the different regions of the try-on image. By incorporating clothing length as an additional variable condition, SSG can predict try-on segmentation map of desired clothing length. As clothing length is considered as low-level information in the image, we begin by mapping z_{target} to $Z \in \mathbb{R}^{H\times W}$ using a fully connected layer. This value is then concatenated with the other inputs. To enhance the prediction ability to complex pose tasks, the dense pose map P_d is also employed to provide additional spatial information. The concatenated inputs are then fed into the generator to produce the raw segmentation map S_{raw}. However, as mentioned, the segmentation generator tends to generate the garment interaction where the top and bottom intersect. As shown in Fig. 3 (b), when generating images where the top and bottom are worn separately without intersecting (e.g. the target top is shorter than the reference), the predicted bottom semantics will go beyond the reference bottom semantics, leading to an inconsistency between S_{raw} and S. This is because I cannot provide sufficient content to guide the image generation for the exceeded bottom semantics. Therefore, to address this issue, we propose a Context Incompatibility Handling module to correct S_{raw}.

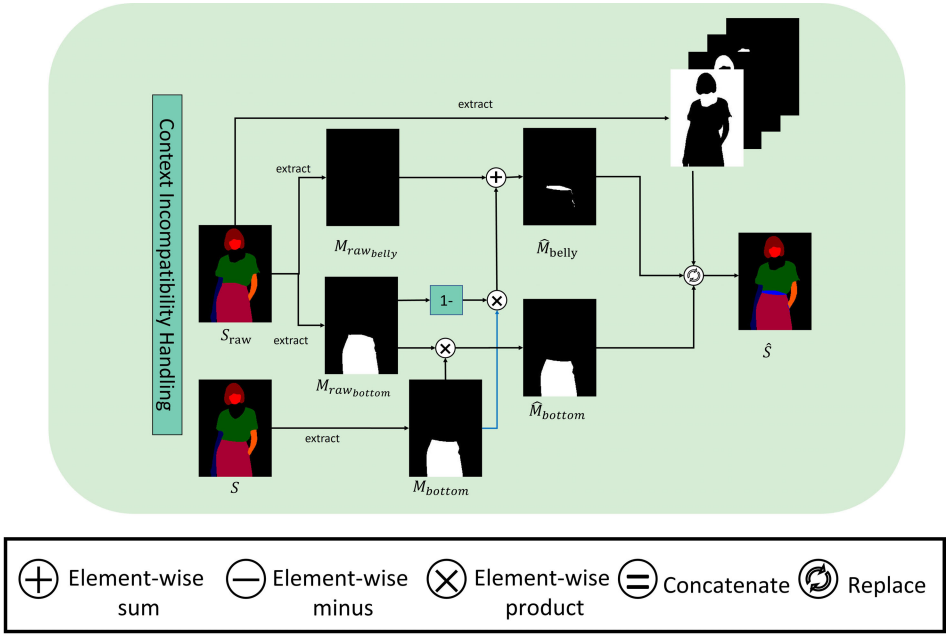

FIGURE 5. The architecture of context incompatibility handling module.

Context Incompatibility Handling: As described in Fig. 5, we extract the top channel $M_{raw_{top}}$ and bottom channel $M_{raw_{bottom}}$ from S_{raw}, as well as the bottom channel M_{bottom} from S. M_{bottom} is used as a shape mask to remove all bottom semantics outside the mask, ensuring that the entire bottom semantics are contained within the mask region. The removed semantics are then filled in the belly channel to protect the semantic integrity of the segmentation map. The following formula expresses the process of Context Incompatibility Handling:

$$\hat{S}^{k,i,j} = \begin{cases} S^{k,i,j}_{raw}, & if \quad k \neq k_{bottom}, k_{belly} \\ S^{k,i,j}_{raw} \cdot S^{k,i,j}, & if \quad k = k_{bottom} \\ S^{k_{bottom},i,j}_{raw} \cdot (1 - S^{k_{bottom},i,j}) + S^{k,i,j}_{raw}, & if \quad k = k_{belly} \end{cases} \tag{2}$$

where S^{kij} and S^{kij}_{raw} indicate the pixel values of the segmentation map S and S_{raw} corresponding to the coordinates (i, j) in channel k. k_{bottom} and k_{belly} denote the index of the bottom and belly. The SSG is trained without Context Incompatibility Handling Module during training. We train SSG to establish a mapping between S and $(S_a, P_o, P_d, z_{reference}, c, cm)$, where $z_{reference}$ is calculated based on S. We adopt U-net [19] as the generator architecture and employ pixel-wise cross-entropy loss $\mathcal{L}_{CE}$ between S and $Sraw$ during training. In addition, we introduce a conditional adversarial loss $\mathcal{L}_{cGAN}$ to encourage SSG to generate segmentation maps with various lengths. The complete objective function used for training SSG can be expressed as:

$$\mathcal{L}_{SSG} = \lambda_1 \mathcal{L}_{cGAN} + \lambda_2 \mathcal{L}_{CE} \tag{3}$$

$$\mathcal{L}_{CE} = -\frac{1}{HW} \sum_{k \in C, i \in H, j \in W} S^{k,i,j} log(S^{k,i,j}_{raw}) \tag{4}$$

$$\begin{aligned} \mathcal{L}_{cGAN} = & E_{X,S,Z}[log(D(X, S, Z))] \\ & + E_{X,S,Z}[log(1 - D(X, S, -Z))] \\ & + E_{X,S,Z}[log(1 - D(X, G(X, Z), Z))] \end{aligned} \tag{5}$$

where λ_1 and λ_2 are hyper-parameters controlling relative importance between two losses, respectively set to 1 and 10. In Eq. (4), the symbols H, W, and C indicate the height, width, and channel number of S. In Eq. (5), the symbols X, S, and Z respectively denote the image inputs of SSG, reference segmentation map, and z.

C. CLOTHING DEFORMATION GENERATOR (CDG)

In this stage, we aim to deform the target clothing item c to achieve alignment with $\hat{S}$. While the flexible deformation of clothing based on TPS transformation conforms to the inherent flexibility of clothing, it falls short in achieving pixel-level alignment between c and $\hat{S}$. Failure to fulfill such alignment could make the next task of the TOIG challenging and potentially lead to artifacts. Inspired by HR, we cascade the clothing flow estimation blocks with the occlusion handling module. The module enables more reasonable cloth flow deformation by reducing pixel-squeezing of patterns and textures when dealing with complex poses where the arms overlap the body.

We adopt the clothing flow estimation from ClothFlow [17] as the structure of our estimation block, and the occlusion handling module adopts the same design proposed in HR. Formally, given (c, cm) and $(\hat{S}, P_d)$ as input, two encoders extract the feature pyramid $\{E_{c_k}\}^4_{k=0}$ and $\{E_{s_l}\}^4_{l=0}$

respectively. Then, we feed the concatenated E_{c_4} and E_{s_4} into a convolution layer to estimate the initial clothing flow F_0. The estimation blocks subsequently upsample and refine F_0 level-by-level to obtain the final flow F_4. Lastly, we utilize F_4 to warp c and use the Occlusion Handling module to remove the self-occlusion, obtaining the well-aligned warped clothing c_{warp}.

Following HR [8], the CDG is optimized using a loss function $\mathcal{L}_{CDG}$ that comprises three terms, which are $\mathcal{L}_{L1}$, $\mathcal{L}_{VGG}$, and $\mathcal{L}_{TV}$. Both $\mathcal{L}_{L1}$ and $\mathcal{L}_{VGG}$ are partitioned into two components, one for direct comparison with the ground truth and another that incorporates intermediate flow estimations to improve performance. The $\mathcal{L}_{L1}$ and $\mathcal{L}_{VGG}$ are expressed as follows:

$$\mathcal{L}_{L1} = \|cm_{warp} - S_c\|_1 + \sum_{i=0}^{3} w_i \cdot \|W(cm, F_i) - S_c\|_1 \tag{6}$$

$$\mathcal{L}_{VGG} = \phi(c_{warp}, I_c) + \sum_{i=0}^{3} w_i \cdot \phi(W(c, F_i), I_c) \tag{7}$$

where w_i specifies the relative importance between each term. $\mathcal{L}_{TV}$ is a total-variation loss to enforce the smoothness of the appearance flow, which is written as:

$$\mathcal{L}_{TV} = \|\nabla F_4\|_1 \tag{8}$$

The total loss of the CDG is as follows:

$$\mathcal{L}_{CDG} = \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV} \tag{9}$$

where λ_{L1} and λ_{TV} are the hyper-parameters showing importance of λ_{L1} and λ_{TV} in L_{CDG}, respectively set to 10 and 2.

D. TRY-ON IMAGE GENERATOR (TOIG)

In our task, the goal of TOIG is slightly different from previous work in that it needs to synthesize corresponding try-on results based on the $\hat{S}$ for different clothing lengths. As the length of the top increases and gradually covers the bottom, we need to decrease the bottom content that will be reproduced. Therefore, I_a must be dynamic obtained in TOIG rather than fixed. We need to regenerate I_a in every inference by removing the target content from I and keeping the bottom content according to $\hat{S}$. Eventually, given $(I_a, \hat{S}, P_d, c_{warp}, cm_{warp})$, TOIG fuses all the inputs as the final try-on image $\hat{I}$.

We attempted to train TOIG with the same human representation as VITON-HD. However, we observed severe artifacts in the belly area of the output images when inferring the results where the top was shorter than the reference (see Fig. 11). With deep research, we found that the human parse in VITON-HD lacks the belt label, leaving the bottom semantics in the human representation incomplete. Additionally, the omission of belly semantics in the human representation made it difficult to generate belly content in the semantic region designated as background. Hence, as shown in Fig. 6,

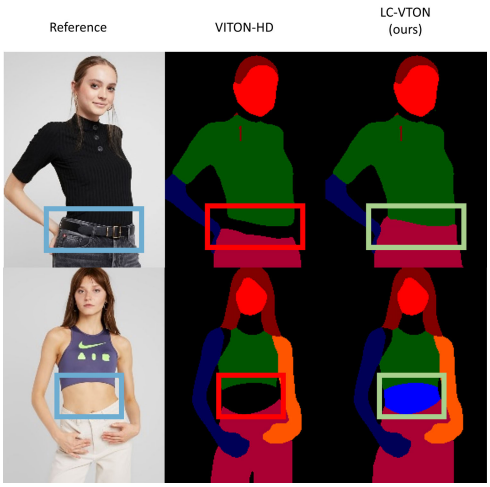

FIGURE 6. The specific repaired semantics. We mark the belt and belly contents in the reference image with blue boxes. The red box is used to reveal the mislabeled semantics. The Green box is used to indicate the correct semantics after repair.

a semantic parser [20] is trained on the ATR dataset [21] and utilized to complete the missing belt label into the bottom label. Moreover, we employ the pose map to preserve the semantics of the belly and introduce the 'belly' into the label set to encourage the generation of accurate belly content.

In this stage, the U-net [19] architecture is utilized as the backbone of TOIG. We train TOIG using conditional adversarial loss, feature matching loss, and perceptual loss, following the same approach as pix2pixHD [22].

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

1) DATASET

All experiments are conducted using the dataset provided by VITON-HD [7], which comprises 13,679 pairs of frontal-view images of women wearing tops. The dataset was divided into two parts, consisting of 11,647 pairs for training and 2,032 pairs for inference. Throughout training and inference, images are downsampled to a resolution of 256×192. The comparison experiments against CP-VTON, CP-VTON+, ACGPN, VITON-HD, and HR-VTON are also conducted using this dataset.

2) TRAINING

We train three separate generators and combine them into one for the try-on task. During training, the target clothing item is identical to the clothing worn in the reference image, and the clothing length $z_{reference}$is calculated from the reference segmentation map. CDG and TOIG are trained for 20 epochs using a batch size of 4, while SSG is trained for 30 epochs with the same batch size. We use the Adam optimizer with hyper-parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All code is implemented using the PyTorch deep learning toolkit, and experiments are conducted on an NVIDIA 1080Ti GPU.

FIGURE 7. Qualitative comparison with baselines.

3) INFERENCE

In inference, LC-VTON incorporates the Context Incompatibility Handling module to infer results. In addition, the input clothes and clothing length differ from those used during the training phase to produce results for various clothing items and lengths. Further results will be presented in subsequent sections.

B. QUALITATIVE RESULTS

1) COMPARISON WITH BASELINES

We evaluate the performance of LC-VTON by comparing it with several state-of-the-art baselines at a resolution of 256 × 192, using publicly available codes. Figure 7 illustrates that LC-VTON produces more realistic images than CP-VTON, CP-VTON+, and ACGPN. Compared with VITON-HD and HR-VTON, LC-VTON produces competitive results in photorealism, exhibiting clear clothing patterns and textures, and the body shape and details are more realistic and natural. These results demonstrate that LC-VTON is capable of generating convincing and photo-realistic outputs.

2) EFFECTS OF THE CLOTHING LENGTH VALUE

In Fig. 8, we display many results to demonstrate our approach's ability to generate try-on images of different clothing lengths while preserving the patterns and textures of the garments. The results presented in this study show that LC-VTON successfully achieves the goal of clothing length control.

3) ABLATION STUDY ON THE EFFECT OF CONTEXT INCOMPATIBILITY HANDLING

We conduct an ablation study to evaluate the efficacy of the Context Incompatibility Handling module in the Semantic Segmentation Generator. The results in Fig. 9 demonstrate that without the Context Incompatibility Handling module, the model erroneously retains the gray content of the clothing-agnostic person image, synthesizing a flawed image. After applying the correction to the bottom semantics, the model successfully synthesizes the belly content. It should be noted that when trying to generate a longer target top than the reference, the module will hardly work because the bottom does not exceed the reference bottom.

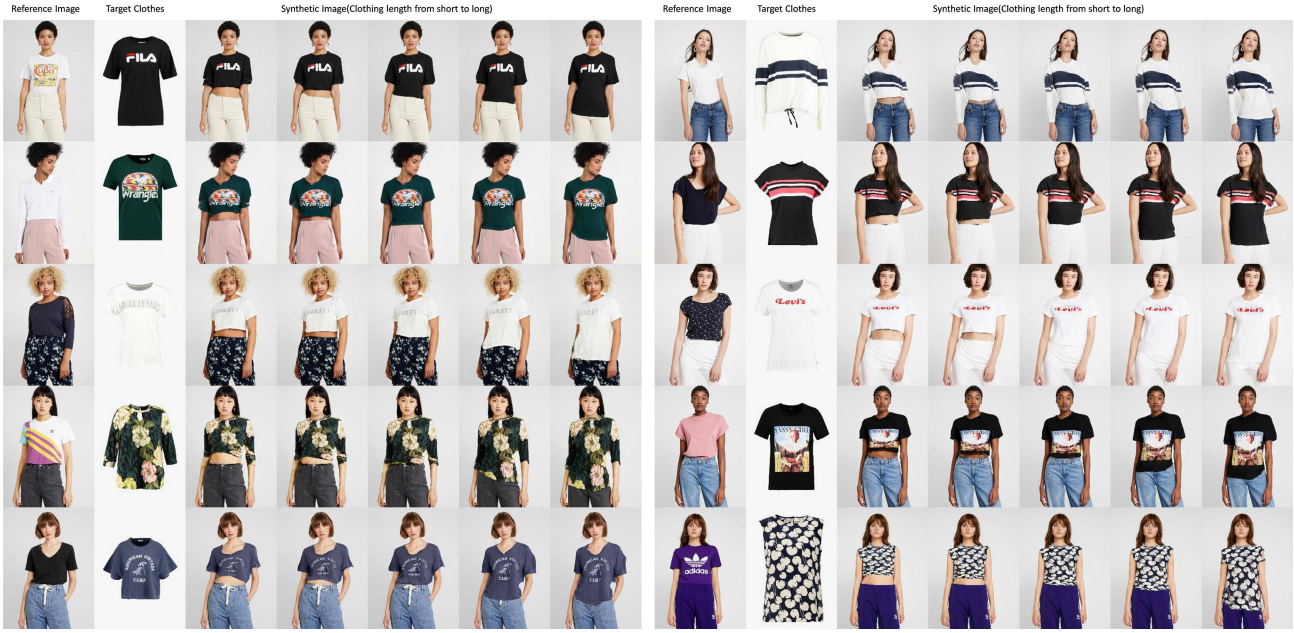

FIGURE 8. Effects of the clothing length value.

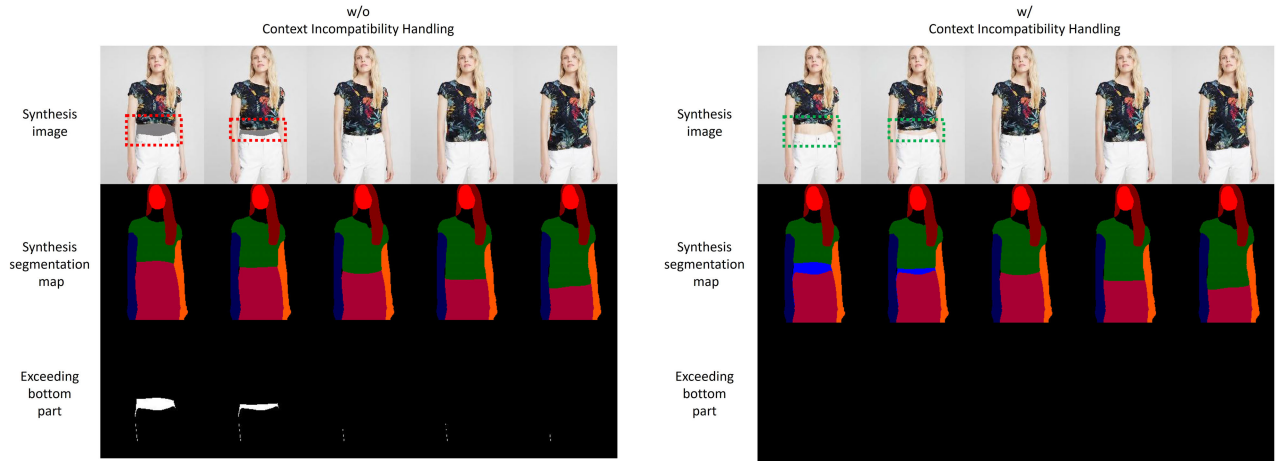

FIGURE 9. Ablation study on the effect of Context Incompatibility Handling. The red box highlights belly content incorrectly synthesized without the Context Incompatibility Handling module. In contrast, the green box indicates the correct belly content generated using the module.

4) EFFECTS OF TARGET ELIMINATION IN TOIG

Figure 10 illustrates the clothing-agnostic person image under various clothing lengths. With increasing clothing length, a greater portion of the upper part of the bottom needs to be removed to ensure the accurate synthesis of the target content.

5) ABLATION STUDY ON THE EFFECT OF LABEL CORRECTION

To demonstrate the adverse effects of incorrect labels on LC-VTON, we perform an ablation study of the label correction. Figure 11 shows the model trained on the incorrect labels produces increasingly blurry belt-like artifacts as the clothing length decreases. However, the model can generate accurate belly content at any clothing length after training on the corrected data.

C. QUANTITATIVE RESULTS

We perform quantitative experiments in paired and unpaired settings separately. The paired setting is used to reconstruct the person wearing the original clothing, while the unpaired setting is used to infer the try-on result. We evaluate our method using four widely adopted metrics in virtual experiments. For paired setting, the Structural Similarity (SSIM) [23] and the Learned Perceptual Image Patch Similarity (LPIPS) [24] are employed to evaluate the similarity between the reconstructed images and reference

FIGURE 10. The clothing-agnostic person images correspond to different lengths examples.

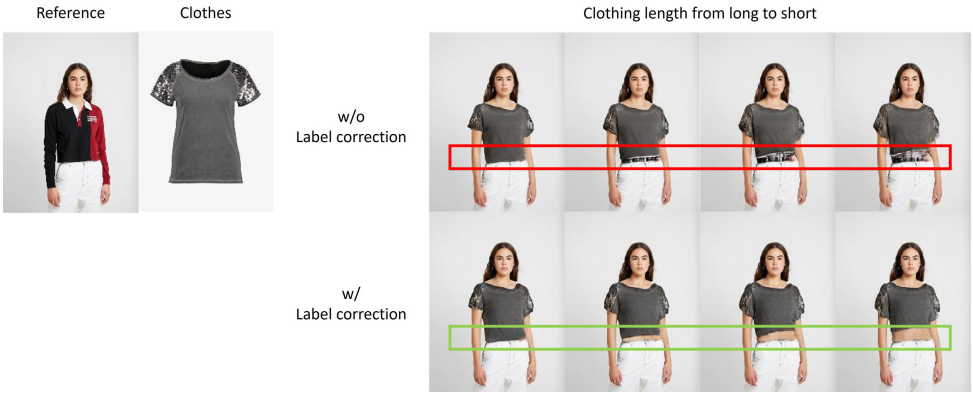

FIGURE 11. Ablation study on the effect of label correction. The red box indicates the artifacts, while the green box shows the correct content.

TABLE 1. Quantitative comparison with baselines in 256 × 192 resolution. The KID value was multiplied by 100. For the SSIM, higher is better. For the LPIPS, FID, and KID, lower is better.

	LPIPS↓	SSIM↑	FID↓	KID↓
CP-VTON	0.159	0.739	30.11	2.034
ACGPN	0.124	0.857	15.75	1.599
VITON-HD	0.084	0.811	16.36	0.871
HR-VTON	0.096	0.864	8.45	0.654
LC-VTON	**0.073**	0.858	**8.21**	0.681

images. For unpaired settings, the Frechet Inception Distance (FID) [25] and the Kernel Inception Distance (KID) are adopted to measure the visual quality of the generated images.

We quantitatively compare our methods with several state-of-the-art baselines at a resolution of 256 × 192. Table 1 shows that LC-VTON outperforms CP-VTON, ACGPN, and VITON-HD on every metric. While LC-VTON does not achieve the same performance as HR-VTON regarding KID and SSIM scores, it demonstrates comparable quantitative levels. Additionally, LC-VTON performs better than HR-VTON in terms of LPIPS and FID scores, indicating its ability to produce photo-realistic images. We consider that, when generating images with short clothing, HR-VTON tends to overlook the person's belly content, whereas LC-VTON is capable of generating belly content. This difference contributes to LC-VTON outperforming HR-VTON in terms of LPIPS and FID scores. The lower SSIM and KID scores may arise from the independent nature of the processes of semantic prediction and clothing deformation, resulting in slight imperfections during the alignment.

V. CONCLUSION

In this paper, we introduce a novel Length Controllable Virtual Try-On Network (LC-VTON), which allows users to control the length to achieve various garment interactions while trying on clothes. We use the newly proposed clothing length value to control the generation of the try-on segmentation map, guiding the generation of length-controllable try-on results. We correct mislabeled semantics in human parse and add a 'belly' label to human representation, which enables LC-VTON to produce images of top and bottom intersecting or belly-naked while continuously controlling the length. The clothing length editing function allows users to personalize

their clothing based on their fashion style, which is significant for virtual try-on applications. Extensive qualitative and quantitative experiments demonstrate that LC-VTON outperforms most existing models.

REFERENCES

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1175–1186.

[2] B. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5419–5429.

[3] J. Li, J. Ye, Y. Wang, L. Bai, and G. Lu, "Fitting 3D garment models onto individual human models," *Comput. Graph.*, vol. 34, no. 6, pp. 742–755, Dec. 2010.

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

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

[6] 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 Recognit. (CVPR)*, Jun. 2020, pp. 7847–7856.

[7] 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 Recognit. (CVPR)*, Jun. 2021, pp. 14126–14135.

[8] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, Oct. 2022, pp. 204–219.

[9] 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. (ICCV)*, Oct. 2019, pp. 10510–10519.

[10] A. Cui, D. McKee, and S. Lazebnik, "Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14618–14627.

[11] X. Gao, Z. Liu, Z. Feng, C. Shen, K. Ou, H. Tang, and M. Song, "Shape controllable virtual try-on for underwear models," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 563–572.

[12] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[13] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[15] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[16] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1225–1233.

[17] X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10470–10479.

[18] J. Son, T. C. Pedroso, C. Siga, and J. Lee, "Controllable garment transfer," 2022, *arXiv:2204.01965*.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[20] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3260–3271, Jun. 2022.

[21] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[23] 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, Apr. 2004.

[24] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[25] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

JINLIANG YAO received the Ph.D. degree in computer application technology from the Institute of Automation, Chinese Academy of Sciences, in 2009. In 2009, he went to Hangzhou Dianzi University to conduct teaching and research work. During the study period, he participated in a number of national and provincial-level projects, published 20 papers at domestic and international journal conferences, and obtained two national patents granted by the first inventor. His current research interests include pattern recognition and image processing. He was a recipient of the Third Prize of the Zhejiang Science and Technology Progress Award.

HAONAN ZHENG was born in Zhejiang, China, in 1998. He received the bachelor's degree from Zhejiang Agricultural and Forestry University, in 2020. He is currently pursuing the master's degree with Hangzhou Dianzi University.