Cloth Interactive Transformer for Virtual Try-On

Bin Ren[1], Hao Tang[1], Fanyang Meng[2], Runwei Ding[3], Ling Shao[4], Philip H.S. Torr[5], Nicu Sebe[16]

[1]University of Trento [2]Peng Cheng Laboratory [3]Peking University Shenzhen Graduate School
[4]Inception Institute of AI [5]University of Oxford [6]Huawei Research Ireland

Abstract

2D image-based virtual try-on has attracted increased attention from the multimedia and computer vision communities. However, most of the existing image-based virtual try-on methods directly put both person and the in-shop clothing representations together, without considering the mutual correlation between them. What is more, the long-range information, which is crucial for generating globally consistent results, is also hard to be established via the regular convolution operation. To alleviate these two problems, in this paper we propose a novel two-stage Cloth Interactive Transformer (CIT) for virtual try-on. In the first stage, we design a CIT matching block, aiming to perform a learnable thin-plate spline transformation that can capture more reasonable long-range relation. As a result, the warped in-shop clothing looks more natural. In the second stage, we propose a novel CIT reasoning block for establishing the global mutual interactive dependence. Based on this mutual dependence, the significant region within the input data can be highlighted, and consequently, the try-on results can become more realistic. Extensive experiments on a public fashion dataset demonstrate that our CIT can achieve the new state-of-the-art virtual try-on performance both qualitatively and quantitatively. The source code and trained models are available at https://github.com/Amazingren/CIT.

1. Introduction

Virtual try-on (VTON), derived from fashion editing [43, 19], aims to transfer a desired in-shop clothing onto a customer's body image. There is no doubt that successfully achieving this will bring both time and energy-saving shopping experience in our daily life. In practice, VTON has been deployed in some big brand clothing stores or e-commerce shopping applications for its convenience. However, most of the existing methods are based on 3D model pipelines [1, 17, 11, 10, 28] and follow the traditional computer graphics convention. Despite the detailed results they require many labor resources, a huge time investment, and a complex data acquisition such as multi-view videos or 3D scans [23] that hinder its widespread application. As an alternative, conditional GAN based methods such as image-to-image translation or other conditional image generation approaches [33, 34, 7], made some positive progress. However, there are still some obvious artifacts that exist in the generated results. To further make the final results of 2D image-based methods more realistic, the classic two-stage pipeline VTON [12] was proposed, utilizing the first stage to warp the in-shop clothing to the desired deformation style and in the second stage the warped clothing is aligned to the consumer's image. Though the results are improved, the performance is still far from the plausible generation. Many approaches following this pipeline, e.g., CP-VTON [37], ACGPN [40], and CP-VTON+ [21], have been proposed

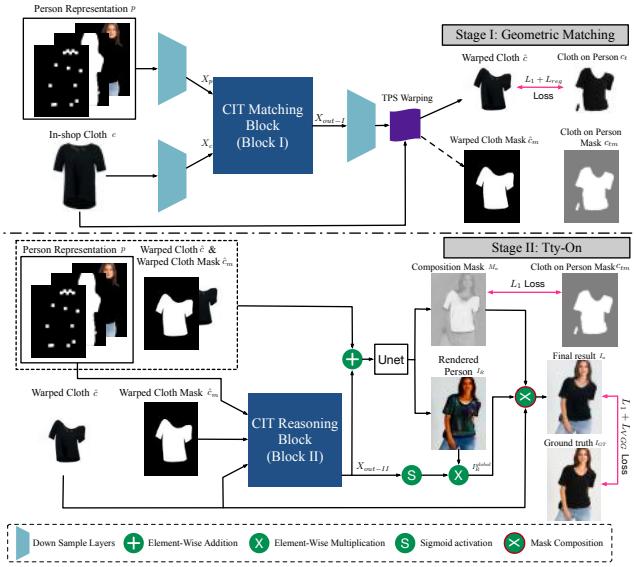

Figure 1: The overall architecture of the proposed Cloth Interactive Transformer (CIT) for virtual try-on.

with more competitive performance. However, their generated try-on results are only good for plain clothes. When there exist rich textures or complex patterns, the performance is considerably lower. This is mostly due to the fact that they did not pay sufficient attention to the input data leading to some serious mismatch phenomena in the warped in-shop clothing and consequently obtaining unsatisfying final try-on results. Moreover, most of the aforementioned methods directly adopt convolution neural networks (CNN) for modeling the relation within input data and the limited receptive fields of the convolution kernels are not able to capture the global long-range relations.

Based on these observations, we also decided to adopt the two-stage pipeline like VTON's [12], but in order to address the aforementioned limitations we propose a novel Cloth Interactive Transformer (CIT) for virtual try-on. The overall architecture of the proposed CIT is shown in Fig. 1.

Within the first geometric matching stage, we design the CIT matching block. It can model long-range relations between the person and clothing representations in an interactive manner. As such a valuable correlation map will be produced to boost the performance of the thin-plate spline (TPS) transformation [4]. Unlike the traditional hand-crafted shape context matching strategies [20, 18, 27] which are only suitable for one certain feature type, the proposed CIT matching block has learnable features and can model global relationships via the cross-modal transformer encoder. Consequently, the warped clothing becomes more natural and can fit a wearer's pose and shape more appropriately. In the second stage, unlike previous methods [37, 21] that treat the warped in-shop clothing and and its corresponding mask as only one modality, we proposed a novel CIT reasoning block that takes the person representation, the warped clothing, and the warped clothing mask separately as inputs. After that the activated global consistency correlation can be established and can strengthen the significant region in input data, which will lead to more natural intermediate result from the UNet. Besides, it serves as an attention map to activate the rendered person image, making the final results sharper and more realistic.

In summary, the contributions of our paper are as follows:

- We propose a novel two-stage Cloth Interactive Transformer (CIT) for the challenging virtual try-on task, which can model long-range interactive relations between the clothing and the person representations. To the best of our knowledge, the proposed CIT is the first transformer-based framework for virtual try-on.
- We propose a new two-modality CIT matching block in the geometric matching stage, making the in-shop clothing to be warped to the desired direction more reasonable.
- We propose a new three-modality CIT reasoning block; based on this block, the latent global long-range correlation can be strengthened to make the final try-on results more realistic.
- Extensive experiments show that the proposed CIT achieves new state-of-the-art results on the dataset collected by Han et al. [12] in the 2D image-based VTON task both qualitatively and quantitatively.

2. Related Work

Virtual Try-On (VTON), as one of the most popular tasks within the fashion area, has been widely studied by the research community due to its great commercial profits and practical potential. Traditionally, this task was realized by computer graphic techniques, which build 3D models and render the output images via the precise control of geometric transformations or physical constraints [9, 5, 6, 10, 28]. By using these 3D measurements or representations, these methods can generate great results for VTON, but the additional need for 3D scanning equipment, computation resource, and heavy labor cannot be ignored.

Compared to 3D-based methods, 2D image-based methods are more applicable to online shopping scenarios. Jetchev and Bergmann [15] propose a conditional analogy GAN to swap fashion articles with only 2D images. However, they did not consider pose variations and also require the paired images of in-shop clothes and the wearer during inference, limiting the applicability in practical scenarios. VITON [12] tackles this problem by a coarse-to-fine architecture, which firstly computes a shape context [3] thin-plate spline (TPS) transformation [4] for warping an in-shop cloth on the target person and then blends the warped clothing to a given person. What is more, CP-VTON [37] and CP-VTON+ [21] adopt similar two-stage frameworks as VITON but make the original TPS transformation learnable based on [24] via a convolutional geometric matcher. While the generated try-on results seem more natural, there are still serious artifacts when there are heavy occlusions, rich texture, or large transformations. Recently, ACGPN [40] was proposed to tackle this issues. Compared to CP-VTON, ACGPN adds an additional semantic generation module to generate a semantic alignment of spatial layout. Though the performance is improved, it is still similar to the one of the previous methods [12, 37, 21]. The problem is that they do not consider the latent global long-range interactive correlation between the person representation and the in-shop clothing.

To alleviate these problems, based on the natural property of the transformer [36], we propose a novel two-stage Cloth Interactive Transformer (CIT) for virtual try-on to model the latent global relation in both stages. As a result, our method can generate sharper and more realistic try-on person images.

Long-Range Dependence Modeling. Although CNN-based methods have shown an excellent representation abil-

ity in various tasks such as classification, segmentation, and so on, it is not easy to build global dependencies due to limited receptive fields of the convolution kernels. This limitation raises great challenges to many application where the long-range relations are needed.

To overcome this limitation, attention mechanism [2, 31, 32, 30] are widely used with CNN models though they were initially designed for natural language processing. Moreover, non-local neural networks [38] were designed based on the self-attention mechanism, allowing to capture long distance dependencies in the feature maps. However, this approach suffers from high memory and computation costs. Also, [26] proposed an attention gate model to increase the sensitivity of a base model. Additionally, Transformers were firstly introduced for neural machine translation tasks [36] because they can model long-range dependencies in sequence-to-sequence tasks and can capture the relations between arbitrary positions in a certain sequence. Unlike previous CNN-based methods, Transformers are built solely on self-attention operations, which is strong in modeling the global context being hence successfully applied to other tasks such as language modeling [22], semantic role labeling [29] and so on. Recently, Transformer-based frameworks have also demonstrated their effectiveness on various vision tasks. Vision Transformer (ViT) [8] splits the image into patches and models the correlation between these patches as sequences with Transformer. The cross-modal transformer [35] was proposed to learn representations directly from unaligned multi-modal streams.

Inspired by [35], we propose the novel Cloth Interactive Transformer (CIT) for virtual try-on to model long-range interactive dependence between the in-shop clothing and person representations.

3. Cloth Interactive Transformer

In this section, we firstly give an overall introduction of the proposed Cloth Interactive Transformer (CIT) for virtual try-on (see Fig. 1) then we provide details on the CIT matching ($Block$-I) and the CIT reasoning ($Block$-II) blocks, respectively.

3.1. Overview and Notation

In the 2D image-based VTON task, consider as given a person image $I \in R^{3\times h\times w}$ and an in-shop clothing image $c \in R^{3\times \times h\times w}$. Our goal is to generate the image $I^o \in R^{3\times h\times w}$ where a person I wears the cloth c. The basic structure of our proposed method is similar to CP-VTON [37] and CP-VTON+ [21]. There are two stages in our proposed CIT, i.e., geometric matching and try-on stages as shown in Fig. 1. The former produces a warped clothing representation and a warped mask according to a given person's pose and shape. The latter utilizes the warped items together with the person appearance to generate the final person image with the worn in-shop clothing.

In the first stage, we design the CIT matching block ($Block$-I, the upper area in Fig. 2) which takes the person feature X_p and the in-shop cloth feature X_c as inputs. Then it produces a correlation feature X_{out-I} followed by a down-sample layer for regressing the parameter θ. θ is used for warping the in-shop clothing c using the on-body style $\hat{c}$ via the thin-plate spline (TPS) warping module. Meanwhile, the relevant mask $\hat{c_m}$ of $\hat{c}$ is also produced based on θ. In the second stage, we utilize the warped cloth and warped mask together with the person representation p as inputs for the CIT reasoning block ($Block$-II, the bottom area in Fig. 2), and the output X_{out-II} is used for two purposes: 1) activate the important region of the overall input $I_{(p,\hat{c},\hat{cm})}$ and 2) guide the final mask composition operation in order to generate more realistic try-on results.

Note that the first stage of CIT is trained with sample triplets $I_{(p,c,cm)}$, while the second stage is trained with $I_{(p,\hat{c},\hat{cm})}$. What is more, in the first matching stage, we adopt the same loss function as CP-VTON+ [21]:

$$L_{Matching} = \lambda_1 \cdot L_1(\hat{c}, c_t) + \lambda_{ref} \cdot L_{reg}, \tag{1}$$

where L_1 indicates the pixel-wise L1 loss between the warped result $\hat{c}$ and the ground truth c_t. L_{reg} indicates the grid regularization loss. In the second stage, the loss item is as follows:

$$\begin{aligned} L_{Try-on} = \lambda_1 \cdot ||I_0 - I_{GT}||_1 + \lambda_{vgg} \cdot L_{VGG} \\ +\lambda_{mask} \cdot ||M_o - c_{tm}||_1, \end{aligned} \tag{2}$$

The first item aims to minimize the discrepancy between the output I_o and the ground truth I_{GT}. The second item, the VGG perceptual loss [16], is a widely used strategy in image generation tasks while the third item is used to encourage the composition mask M_o to select the warped clothing mask c_{tm} as much as possible.

3.2. Interactive Transformer

Based solely on the self-attention mechanism instead of convolutional operations, Transformer is powerful for modeling global dependence. Consider the Transformer's outstanding global relation modeling ability, we propose a novel Interactive-Transformer for exploring the relationship between the person and the clothing in the VTON domain. We design two kinds of Interactive Transformers in this paper, the first version i.e., Interactive Transformer I is utilized in the geometric matching stage, and the second version, i.e., Interactive Transformer II is used in the try-on stage. They are both built on the basic Transformer encoders and on the cross-modal Transformer encoders that are showed in Fig. 2.

Before we describe the proposed Interactive Transformer, let's consider one basic network of total N layers

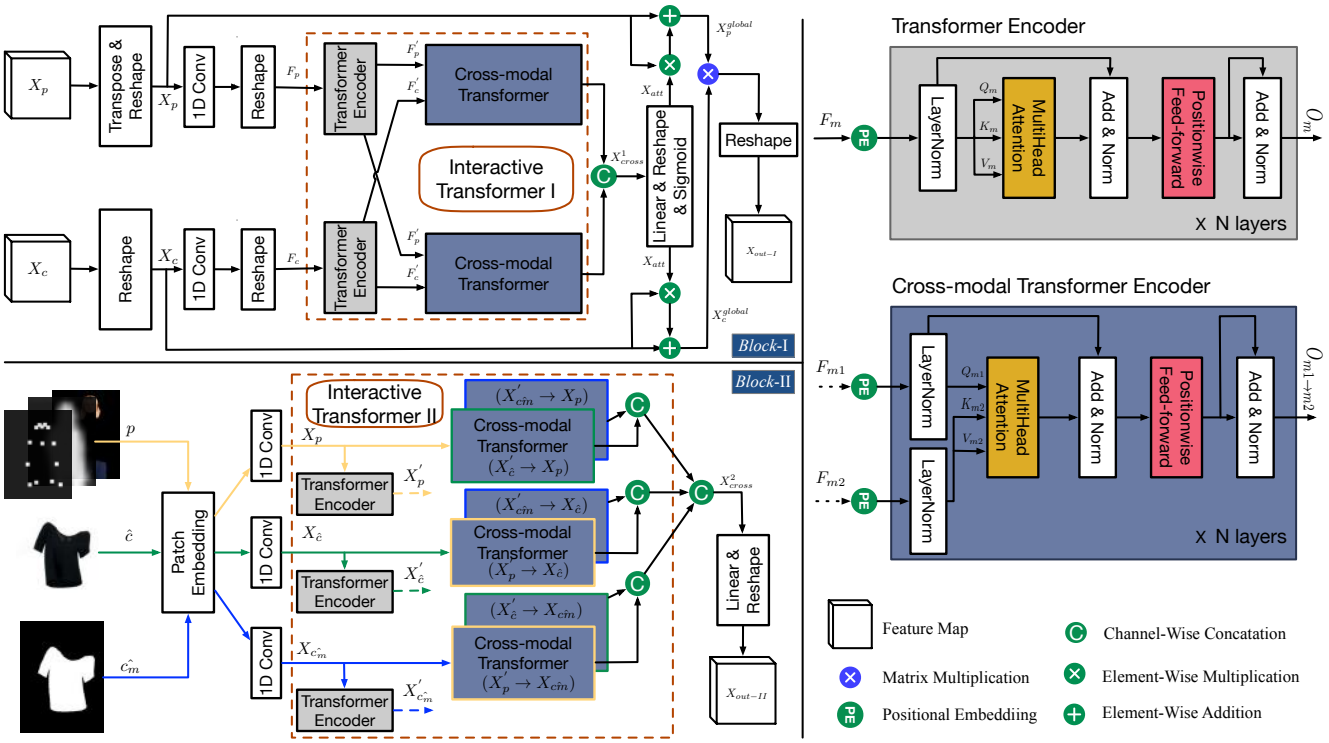

Figure 2: The key components of the proposed Cloth Interactive Transformer (CIT) for virtual try-on. The upper area on the left is the CIT Matching Block (*Block*-I) while the bottom area of the left indicate the CIT Reasoning Block (*Block*-II). On the right, normal Transformer encoder and the proposed cross-modal Transformer encoder are shown in detail.

as an example. For a regular Transformer encoder, given an input feature F_m, a positional embedding is firstlu added to the input feature as discussed in [36]. This is useful for extracting context from the original input feature based on orders. The input feature after the positional embedding will be transformed into queries Q_m, keys K_m, and values V_m by a linear projection with relevant weight. Then the output of the attention layer A_m is computed as:

$$A_m = \text{softmax}(\frac{Q_m K_m^T}{\sqrt{d}})V_m, \tag{3}$$

In practice, we also use the multi-head mechanism. The outputs from every single head will be combined for making the model jointly attend to information from different representation sub-spaces at different positions. Similarly to one modal input, cross-modal Transformer encoder's attention layer $A_{m2\to m1}$ can be computed as:

$$A_{m2\to m1} = \text{softmax}(\frac{Q_{m1} K_{m2}^T}{\sqrt{d}})V_{m2}, \tag{4}$$

Finally, after skip connection and position-wise feed-forward, we will obtain the output of each Transformer encoder unit O_m or $O_{m2\to m1}$. Note that in this cross-modal Transformer encoder, each modality keeps updating its sequence via the low-level external information from the multi-head cross-modal attention module. As a result, one modality will be transformed into a different set of key/value pairs to interact with another modality.

Interactive Transformer I is highlighted in the upper area of Fig. 2. Each Interactive Transformer encoder consists of two regular Transformer encoders and two cross-modal Transformer encoders. $selfTrans()$ indicates the regular Transformer operator while $crossTrans()$ indicated the cross-modal Transformer operator. Given two inputs X_p and X_c, each of them will go through a N-layer regular Transformer encoder to produce $X_p^{'}$ and $X_c^{'}$. This procedure can be described as follows:

$$\begin{aligned} X_p^{'} &= selfTrans(X_p), \\ X_c^{'} &= selfTrans(X_c), \end{aligned} \tag{5}$$

We then use the cross-modal Transformer operator for generating the final output of the Interactive Transformer I:

$$\begin{aligned} X_{cross}^1 = \text{cat}(&crossTrans(X_p^{'}, X_c^{'}), \\ &crossTrans(X_c^{'}, X_p^{'})), \end{aligned} \tag{6}$$

where $crossTrans(X_p^{'}, X_c^{'})$ indicates that we utilize feature $X_c^{'}$ to form keys and values while we use $X_p^{'}$ to form queries within an self-attention module. On the other hand, $crossTrans(X_c^{'}, X_p^{'})$ indicates that keys and values come from $X_p^{'}$ and queries come from $X_c^{'}$. After concatenat-

ing the outputs from the two cross-modal Transformer encoders, we finally get the output X^1_{cross} from the Interactive Transformer I. It can strengthen the correlation matching ability.

Interactive Transformer II is highlighted in the bottom area of Fig. 2. Similarly to Interactive-Transformer I, the Interactive Transformer II is also constructed by using normal single-modal and cross-modal Transformer encoders. However, with Interactive-Transformer II, the number of normal single-modal transformer encoders is increased to three and consequently, the number of cross-modal Transformer encoders is increased to six. Given three kinds of input feature X_p, $X_{\hat{c}}$, and $X_{c\hat{m}}$, after feeding them to their corresponding normal single-modal Transformer encoder, we obtain $X^{'}_p$, $X^{'}_{\hat{c}}$, and $X^{'}_{c\hat{m}}$ respectively.

To simplify the description, we use the input X_p for explanation. Firstly, as queries, X_p will receive $X^{'}_{\hat{c}}$ to form the cross-modal input $(X^{'}_{\hat{c}} \rightarrow X_p)$ to one of the cross-modal Transformer encoder, then it will form another cross-modal input $(X^{'}_{c\hat{m}} \rightarrow X_p)$ for another cross-modal Transformer encoder.

$$X^{cross}_p = \text{cat}(crossTrans(X^{'}_p, X^{'}_{\hat{c}}), \\ crossTrans(X^{'}_p, X^{'}_{c\hat{m}})), \tag{7}$$

Similarly, when we take $X_{\hat{c}}$ and $X_{c\hat{m}}$ as queries respectively, we willget two other intermediate results X^{cross}_c and X^{cross}_{cm}. Finally, the overall output of Interactive Transformer II becomes:

$$X^2_{cross} = \text{cat}(X^{cross}_p, X^{cross}_c, X^{cross}_{cm}), \tag{8}$$

3.3. CIT Matching Block

For the thin-plate spline (TPS) transformation for warping the in-shop clothing c, VTON [12] has utilized handcraft shape-context features. This procedure is not only time-consuming but is also not robust when facing a new hard sample. As an improvement, CP-VTON and CP-VTON+ directly adopt the approach of [24] by making this procedure learnable based on convolutional neural networks (CNN). Though the TPS transformation correlation can be established in this way, we argue that the global long-range relation modeling is still unsatisfactory for two reasons. Firstly, due to the limited receptive fields of the convolution kernels, the convolution operation cannot capture global dependencies. Secondly, the main idea behind [24] is only to build a correlation map directly via matrix multiplication without considering the interaction between two input data, so correlation itself does not benefit from their mutual properties.

Based on these considerations, we propose a novel Cloth Interactive Transformer Matching (CIT matching) block, aiming to add a global long-range modeling ability to the original geometric matching stage. As illustrated in the upper area of Fig. 1, there are also four parts in our geometric matching stage similar to CP-VTON and CP-VTON+. Besides the proposed $Block$-I, the other parts like feature extractor, regression network, and TPS warping module share the same structure as CP-VTON+. Our goal is to utilize $Block$-I to capture a more robust global long-range correlation between two kinds of input feature, i.e., $X_p \in R^{B\times C\times H\times W}$ and $X_c \in R^{B\times C\times H\times W}$ based on our designed Interactive-Transformer I encoder. B, C, H, and W indicate batch size, channel number, image height, and image width, respectively.

To utilize the Transformer encoder for modeling long-range dependencies we utilize two reshape operations for each input data to adjust the size to a sequence-likewise data form. Besides, a 1D temporal convolution layer is also adopted to ensure that each element of the input sequences has sufficient awareness of its neighborhood elements and the convolved sequences are expected to contain the local structure of the sequence. These are depicted in Fig. 2 with detailed annotations. After preprocessing the person and the in-shop clothing representations X_p and X_c, we get F_p and F_c. Then based on the aforementioned N layers Interactive Transformer I, we get X^1_{cross} which models the global long-range correlation between the person data and the in-shop clothing data.

Instead of directly adding this long-range relation to features X_p or X_c, we strengthen both by a global strengthened attention X_{att} as follows:

$$X^{global}_{(.)} = X_{(.)} + X_{(.)} \times X_{att}, \tag{9}$$

Here $\times$ means an element-wise multiplication and $(.)$ indicates that both features X_p and X_c follow the same strengthen form. Note that X_{att} is produced from X^1_{cross} by a linear projection and a sigmoid activation. Then a matrix multiplication is conducted to these two global strengthened features. The output X_{out-I} of the proposed CIT matching block is finally obtained after a reshaping operation.

3.4. CIT Reasoning Block

Previous works like CP-VTON and CP-VTON+ directly concatenate the person image p, the warped clothing image $\hat{c}$, and the warped clothing mask image $c\hat{m}$. Then the concatenated input is sent to a UNet to generate a composition mask M_o and a rendered person image I_R. For the same reason described in the last subsection we design the Cloth Interactive Transformer Reasoning (CIT reasoning) block in short $Block$-II depicted in Fig. 2, aiming to model the global dependence among these three input images.

Inspired by [8], we adopt the patch embedding to the three inputs, for making the image data compatible. Then each goes through a 1D temporal convolution to ensure the

relation modeling of each element with its neighbor elements. Then the Interactive-Transformer II is utilized for modeling the global long-range correlation. The output X_{out-II} of Interactive Transformer II is obtained after a linear projection and a reshape operation as shown in the bottom area of Fig. 2. Then X_{out-II} is utilized for two proposes; one is to activate the important region of the overall input by adding X_{out-II} to $I_{(p,\hat{c},\hat{cm})}$, another is to guide the final mask composition as follows:

$$\begin{aligned} I_R^{global} &= \text{sigmoid}(X_{out-II}) \times I_R, \\ I_o &= M_o \times \hat{c} + (1 - M_o) \times I_R^{global}, \end{aligned} \tag{10}$$

where $\times$ represents the element-wise multiplication and $sigmod$ indicates the Sigmod activation function.

4. Experiments

Datasets. We conduct all our experiments on the same dataset collected by Han et al. [12] and used in VITON, CP-VTON, and CP-VTON+. Because of copyright issues, we could only utilize the reorganized version. It contains around 19,000 front-view women and top clothing image pairs. There are 16,253 cleaned pairs, which are split into a training set and a validation set with 14,221 and 2,032 pairs, respectively. In the training set, the target clothing and the clothing worn by the wearer are the same. However, in the test set, the two are different.

Evaluation Metrics. We follow [21, 40, 13] and adopt Jaccard Score (JS) [14], Structure Similarity (SSIM) [39] and Learned Perceptual Image Patch Similarity (LPIPS) [42] metrics for the same clothing try-on cases (with ground truth cases) in the first geometric matching stage and the second try-on stage, respectively. The original target human image is used as the reference image for SSIM and LPIPS (lower score means better), and the parsed segmentation area for the current upper clothing is used as the JS reference. For different clothing try-on (where no ground truth is available), we used the Inception Score (IS) [25].

Implementation Details. In our experiments, the settings are similar to CP-VTON and CP-VTON+. We set λ_1=λ_{VGG}=λ_{mask}=1 and λ_{ref}=0.5 during training. Both stages are trained for 200K steps with batch size 4. Moreover, for Adam optimizer, β_1 and β_2 are set to 0.5 and 0.999, respectively. The learning rate was firstly fixed at 0.0001 for the first 100K steps and then linearly decayed to zero for the rest of the steps. All input images are resized to 256×192 and the output images have the same resolution.

4.1. Quantitative Evaluation

To evaluate our method, we adopt four metrics, i.e. JS, SSIM, LPIPS, and IS. JS is to evaluate the quality of the warped mask in the first geometric matching stage with same-pair test samples, which is similar to the IoU metric used in CP-VTON+ but is more convenient for implementation. Note that we take cloth masks on person as the reference images. Other metrics are designed to evaluate the performance of the second try-on stage. Note that SSIM and LPIPS are also used for same-pair images (retry-on) while only IS is utilized for evaluating the unpaired generated try-on results. Besides, numerical results of CP-VTON were computed by us while for CP-VTON+, the results were obtained based on the officially provided checkpoints.

Table 1: Quantitative results on paired setting (JS, SSIM, LPIPS, IS, Q1, Q2) and on unpaired setting. For LPIPS, the lower is better. For JS, IS, SSIM, Q1, and Q2, the higher is better. Q1 denotes 'Which image is the most photo-realistic?', and Q2 denotes 'Which image preserves the details of the target clothing the most?' in the user study. Note that ACGPN⋆ indicates special cases, more explanation are given in Section 4.1.

Method	JS	SSIM	LPIPS	IS	Q1 (%)	Q2 (%)
CP-VTON [37]	0.759	0.800	0.1256	2.832	19.3	14.6
CP-VTON+ [21]	**0.812**	0.817	0.117	**3.074**	26.9	25.2
VTNFP [41]	-	0.803	-	2.784	-	-
CIT (Ours)	0.800	**0.827**	**0.115**	3.060	**32.6**	**35.4**
ACGPN⋆ [40]	-	0.845	-	2.829	21.2	24.8

From Table 1 we can see that the proposed CIT achieves the best numerical evaluation results for two metrics, the highest SSIM score while the lowest LPIPS score among other state-of-the-art methods such as CP-VTON [37], CP-VTON+ [21], and VTNFP [41]. Though we do not obtain not the best quantitative scores in JS and IS metrics, our proposed CIT can generate sharper and more realistic try-on images compared to CP-VTON+. For ACGPN, we also test the performance based on the official checkpoints. However, when we utilize the same test pair samples, the numerical scores are extremely unsatisfactory, so we put the reported evaluation results in Table 1 and give a visualization comparison in the next part.

4.2. Qualitative Evaluation

To further validate the performance of the proposed CIT for virtual try-on, we present the visualization comparison results from both stages, including warped clothing and final try-on person images.

Comparison of Warping Results. We visualize the warped cloths results for both retry-on (same-pair) and try-on (different-pair) cases in Fig. 3. We can see that the proposed CIT can generate sharper and more realistic warped clothing than the other state-of-the-art methods like CP-VTON, CP-VTON+, and ACGPN. We want to argue that the JS or IoU scores are not always able to capture the actual image generation quality. For instance, though Table 1 shows that CP-VTON+ has the best JS score of 0.812, which is higher than ours 0.800, the qualitative results show that our method is superior to CP-VTON+. This typically happens

Figure 3: Qualitative comparisons of the warped cloths by the proposed CIT based geometric matching stage. To the left of the line of dashes are same-pair (retry-on) cases while on the right are the different-pair cases for try-on cases.

Table 2: Ablation studies of the proposed CIT for virtual try-on.

Method	JS	SSIM	LPIPS	IS
Baseline [21]	0.812	0.817	0.117	**3.074**
B1 (CIT Matching only)	0.800	0.808	0.123	3.020
B2 (CIT Reasoning only)	0.812	0.821	0.125	**3.105**
B3 (Full: B1+B2)	0.800	0.827	0.115	3.060
B4 (Full + L_1 mask loss)	**0.813**	**0.829**	**0.110**	3.005

for texture-rich cases such as the cases with line-stripes in the last row of the same-pair cases and the second row in the different-pair cases, or in the presence of logos in the second row of the same-pair cases and third row of the different-pair cases, and so on. That ids to say, that the higher JS or IoU scores do not always mean a better result, because the competing results over-warp the in-shop clothing toward an uncontrollable direction.

Comparison of Try-on Results. Fig. 4 shows the final try-on results (different-pair cases). There is no doubt that our CIT outperforms the other state-of-the-art methods. The proposed CIT can keep the original clothing texture and its pattern as much as possible, and the final resulting images are more realistic and natural. Compared to our method, the other approaches display many artifacts, for example, the irregular logo pattern (the first row), the over-warped cloth texture (the third row), the ridiculous results for unique cloth (the last row), etc. All these try-on results further strengthen the evidence that our proposed CIT is superior to others.

User Study. We further evaluate the proposed CIT and other baselines via a user study. We randomly select 120 sets of reference and a target clothing images from the test dataset. Given the reference images and the target clothes, 30 users are asked to choose the best outputs of our model and baselines (i.e., CP-VTON, CP-VTON+, and ACGPN) according to the two questions: (Q1) Which image is the most photo-realistic? (Q2) Which image preserves better the details of the target clothing? As shown in Table 1, we can see that the proposed CIT achieves significantly better results than the other baselines, which further demonstrates that our model generates more realistic images, and preserves the details of the clothing items compared to the baselines.

4.3. Ablation Study and Discussion

To validate the effectiveness of CIT, we conduct four ablation experiments (i.e., B1, B2, B3, B4 in Table 2) and take CP-VTON+ [21] as our baseline: B1 means we only use the CIT Matching block in the first geometric matching stage; B2 denotes we only use the CIT Reasoning block in the second try-on stage; B3 is the final version adopted in this paper; B4 is based on B3, while we additionally add a L_1 loss between generated warped clothing mask $\hat{c_m}$ and the cloth on person mask c_{tm}, its overall matching loss can be

Figure 4: Qualitative comparisons of different state-of-the-art methods.

summarized as follows:

$$L_{Matching} = \lambda 1 \cdot L_1(\hat{c}, c_t) + \lambda 2 \cdot L_1(\hat{c_m}, c_{tm}) + \lambda_{ref} \cdot L_{reg}, \quad (11)$$

where $\lambda 1$, $\lambda 2$, and λ_{ref} are all set to 1. The other settings are the ones of B3.

The results are displayed in Table 2. From the comparison of baseline [21] and B1, though CP-VTON+ has a better JS score, the qualitative results presented in Fig. 3 indicate that B1 can generate more natural warped cloths. The only difference is that B1 adopts the CIT Matching block in the first stage. In other words, the warped cloth generated by our CIT Matching block seems more reasonable because it can model latent global cues within two input features. Another interesting comparison is between the baseline and B2. We only adopt the CIT Reasoning block in the second try-on stage and keep the other settings the same as [21]. Numerical results such as SSIM and IS further demonstrate that the proposed reasoning block can boost the try-on results via its global long-range relationship modeling ability. B3 is a combination of B1 and B2, and is also the final strategy we adopt in this paper. B3 produces not only the more natural warped clothing but is also able to obtain the more realistic try-on result. Consequently, we argue that for the

Figure 5: Qualitative comparisons of ablation studies between B3 and B4.

two-stage 2D image-based VTON task, it is not always true that the final try-on results are better when JS or IoU scores are higher. To validate this, we conduct B4 for further exploration. In Table 2 we can see B4 obtains all the best numerical results but the IS score. Does this mean that B4 can generate the most realistic try-on results? We give several visualization samples in Fig. 5 for this question.

The try-on results of B3 are better than B4. The only thing different is that B4 adopts another L_1 loss to make a further constraint between the warped cloth mask and the cloth on person mask. Though it boosts the numerical results such as JS, SSIM, and LPIPS scores, the overall results are worsen (see the artifacts marked out by a red line of dashes of B4 Try-on results in the fourth column of Fig. 5). We argue that this worsening mainly comes from the over-warped cloth shown in the third column; we also mark out those ambiguous parts in the third column of Fig. 5. From B3 we can see that when combining the CIT Matching (B1) and CIT Reasoning (B2) bloks, both SSIM and LPIPS are better than in the case when only one of them is taken into consideration. Though IS seems to decrease from 3.105 from 3.060, the qualitative results become more realistic.

4.4. Failure Cases and Analysis

Though impressive person try-on images can be generated by our CIT, there are still several failure cases that are inevitable. We found three such cases and we visualize them in Fig. 6. In the first case (first row), the difference between the clothing in the reference image and the in-shop cloth image is too big. As such the person's mask cannot match the new in-shop cloth anyway. Moreover,

Figure 6: Several failure cases of our proposed CIT for virtual try-on.

self-occlusion is another big problem, leading to blurry of ambiguity-prone generated images (second row). In the third case (last row), very large poses and shape transformations may also lead to ambiguous results.

For the first two cases, the main reason is that the input data itself lack information on whether a body region should be covered with the cloth or not. We can give a further organization to input data to remedy this, such as using more accurate segmentation maps or adopting more fine-grained human annotations. For the last case, the 2D image-based method cannot completely capture the large pose and shape transformations. We think that utilizing 3D input data such as body mesh and 3D models for the clothing may alleviate this problem.

5. Conclusion

In this paper, we proposed a novel two-stage Cloth Interactive Transformer (CIT) for image-based virtual try-on tasks. To the best of our knowledge, our work is the first to utilize Transformer for this task. In the first stage, we proposed a transformer-based matching block, which can model global long-range relations when warping a cloth via learnable thin-plate spline transformation and as a result, the warped cloth can be more natural. We also designed a new transformer-based reasoning block in the second stage. On one hand, the reasoning block can strengthen the important regions within the input data, on the other hand, the mutual interactive relations established via the reasoning block further improve the rendering process to make the final try-on results more realistic. Lastly, we conducted extensive experiments in terms of quantitative and qualitative comparisons and provided in-depth discussions, validating that our proposed CIT achieves new state-of-the-art performance.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*. CVPR Spotlight Paper. 1

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[3] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002. 2

[4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 2

[5] Remi Brouet, Alla Sheffer, Laurence Boissieux, and Marie-Paule Cani. Design preserving garment transfer. *ACM Transactions on Graphics*, 31(4):Article–No, 2012. 2

[6] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5

[9] Jun Ehara and Hideo Saito. Texture overlay for virtual clothing based on pca of silhouettes. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 139–142. Citeseer, 2006. 2

[10] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 1, 2

[11] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 1

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 1, 2, 5, 6

[13] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. *arXiv preprint arXiv:2007.02721*, 2020. 6

[14] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912. 6

[15] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2287–2292, 2017. 2

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[17] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018. 1

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[19] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 1

[20] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *European conference on computer vision*, pages 128–142. Springer, 2002. 2

[21] MR Minar, TT Tuan, H Ahn, P Rosin, and YK Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, volume 2, page 11, 2020. 1, 2, 3, 6, 7, 8

[22] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019. 3

[23] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 1

[24] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2, 5

[25] Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 6

[26] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019. 3

[27] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 19(5):530–535, 1997. 2

[28] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014. 1, 2

[29] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*, 2018. 3

[30] Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *ACM MM*, 2020. 3

[31] Hao Tang, Song Bai, Philip HS Torr, and Nicu Sebe. Bipartite graph reasoning gans for person image generation. In *BMVC*, 2020. 3

[32] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 3

[33] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019. 1

[34] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020. 1

[35] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 3, 4

[37] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 1, 2, 3, 6

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[40] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020. 1, 2, 6

[41] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10511–10520, 2019. 6

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[43] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pages 1680–1688, 2017. 1