C-VTON: Context-Driven Image-Based Virtual Try-On Network

Benjamin Fele[1,2] Ajda Lampe[1,2] Peter Peer[2] Vitomir Štruc[1]
[1]Faculty of Electrical Engineering, [2]Faculty of Computer and Information Science
University of Ljubljana, SI-1000 Ljubljana, Slovenia
{benjamin.fele, vitomir.struc}@fe.uni-lj.si, {ajda.lampe, peter.peer}@fri.uni-lj.si

Figure 1. Example results generated with the proposed Context-Driven Virtual Try-On Network (C-VTON). The original input image is shown in the upper left corner of each example and the target clothing in the lower left. Note that C-VTON generates convincing results even with subjects in difficult poses and realistically reconstructs on-shirt graphics.

Abstract

Image-based virtual try-on techniques have shown great promise for enhancing the user-experience and improving customer satisfaction on fashion-oriented e-commerce platforms. However, existing techniques are currently still limited in the quality of the try-on results they are able to produce from input images of diverse characteristics. In this work, we propose a Context-Driven Virtual Try-On Network (C-VTON) that addresses these limitations and convincingly transfers selected clothing items to the target subjects even under challenging pose configurations and in the presence of self-occlusions. At the core of the C-VTON pipeline are: (i) a geometric matching procedure that efficiently aligns the target clothing with the pose of the person in the input images, and (ii) a powerful image generator that utilizes various types of contextual information when synthesizing the final try-on result. C-VTON is evaluated in rigorous experiments on the VITON and MPV datasets and in comparison to state-of-the-art techniques from the literature. Experimental results show that the proposed approach is able to produce photo-realistic and visually convincing results and significantly improves on the existing state-of-the-art.

1. Introduction

In the age of online shopping, virtual try-on technology is becoming increasingly important and a considerable amount of research effort is being directed towards this area as a result [2]. Especially appealing here are *image-based* virtual try-on solutions that do not require specialized hardware and dedicated imaging equipment, but are applicable with standard intensity images, e.g., [3, 4, 7, 8, 12]. As illustrated in Figure 1, the goal of such solutions is to replace a piece of clothing in an input image with a target garment as realistically as possible. This allows for the design of virtual fitting rooms that let consumers try-on clothes without visiting (brick and mortar) stores, and also benefits retailers by reducing product returns and shipping costs [1, 2, 20].

Existing solutions to image-based virtual try-on are dominated by two-stage approaches (and their extensions) that typically include: (i) *a geometric matching* stage that aligns the target clothing to the pose of the person in the input image, and estimates an approximate position and shape of the target garment in the final try-on result, and (ii) *an image synthesis* stage that uses dedicated generative models (e.g., Generative Adversarial Networks (GANs) [9]), together with various refinement strategies to synthesize the final try-on image based on the aligned clothing and different auxiliary sources of information, e.g., pose keypoints, parsed body-parts, or clothing annotations among others [3, 12, 35]. To further improve on this overall framework, various enhancements have also been proposed, including: (i) refinements of the data fed to the geometric matching stage [23, 35, 36], (ii) integration of clothing segmentations into the synthesis procedure [35, 36], and (iii) the use of knowledge distillation schemes to minimize the impact of parser-related errors [8, 16].

While the outlined advances greatly improved the quality of the generated try-on results, the loss of details on the transferred garments that is often a consequence of difficulties with the geometric matching stage still represents a ma-

jor challenge to image-based virtual try-on solutions [4, 12, 34]. Additionally, poor quality (human/clothing) parsing results typically still lead to unconvincing try-on images with garment textures being placed over incorrect body parts. Although recent (destilled) parser-free models, e.g., [16, 8], address this issue to some degree, they still inherit the main characteristics of the teacher models (including parsing issues) and often struggle with the generation of realistic body-parts, such as hands or arms.

In this paper, we propose a Context-Driven Virtual Try-On Network (C-VTON) that aims to address these issues. To improve the quality of the generated on-garment textures and logotypes, we design a novel geometric matching module that conditions the pose-matching procedure on body-segmentations only, and, therefore, minimizes the dependence on multiple (potentially error-prone) pre-processing steps. Additionally, we formulate learning objectives for the module's training procedure that penalize the appearance of the aligned clothing solely within the body area (while ignoring other body parts) to ensure that challenging pose configuration and self-occlusions (e.g. from hands) do not adversely affect performance. This design leads to realistic virtual try-on results with convincing details, as also shown in Figure 1. Finally, we develop a powerful context-aware image generator (CAG) that utilizes contextual cues in addition to the warped clothing to steer the synthesis process. The generator is designed as a standard residual network, but relies on conditional (context-dependent) normalization operations (akin to SPADE layers [24]) to ensure the contextual information is considered to a sufficient degree when generating virtual try-on result. In summary, we make the following main contributions in this paper:

- We propose a novel image-based approach to virtual try-on, named Context-Driven Virutal Try-On Network (C-VTON), that produces state-of-the-art results with input images of diverse characteristics.
- We design a simplified geometric matching module, termed Body-Part Geometric Matcher (BPGM), capable of producing accurate garment transformations even with subjects in challenging poses and arm configurations.
- We introduce a Context-Aware Generator (CAG) that allows for the synthesis of high-quality try-on results by making use of various sources of contextual information.

2. Related work

Image-based virtual try-on techniques have recently appeared as an appealing alternative to traditional try-on solutions that rely on 3D modeling and dedicated computer-graphics pipelines [10, 25, 26, 30]. The pioneering work from [17, 27], for example, approached the virtual try-on task as an image analogy problem with promising results. However, due to the lack of explicit (clothing) deformation modelling, the generated images exhibited only limited photo realism. To address this shortcoming, Han *et al*. [12] proposed a two-stage approach, named VITON, that used a coarse-to-fine image generation strategy and utilized a Thin-Plate Spline (TPS) transformation [6] to align the image of the desired clothing with the pose of the target subject. Wang *et al*. [34] improved on this approach with CP-VTON, which introduced a Geometric Matching Module (GMM) that allowed to learn the TPS clothing transformations in an end-to-end manner (similarly to [28]) and led to impressive try-on results. Follow-up work further refined the geometric matching stage using various mechanisms. CP-VTON+ [23], for instance, improved the human mask fed to the GMM, VTNFP [36] designed an elaborate person representation as the input to the GMM, whereas LA-VITON [21] and ACGPN [35] introduced additional transformation constraints when training their warping/matching modules. With C-VTON we follow the outlined body of work and design a novel matching module based on simplified inputs that can be estimated reliably even in the presence of considerable appearance variability. We achieve this by conditioning the module on body-parts only and leveraging the power of recent human-parsing models.

Several solutions have also been presented in the literature to improve the quality of the generated try-on results during the image synthesis stage. MG-VTON [4], ACGPN [35] and VITON-HD [3], for example, proposed using secondary neural networks that generate clothing segmentations matching the target garment and utilizing these as additional sources of information for the generator. S-WUTON [16] and PF-AFN [8] employed a teacher-student knowledge distillation scheme to alleviate the need for error-prone (intermediate) processing steps that often contribute to difficulties with existing try-on approaches. FE-GAN [5] and VITON-HD [3] followed recent developments in image synthesis [24] and introduced generators with conditional normalization layers to help with the quality and realism of the synthesized try-on results. Similarly to these techniques, C-VTON also uses an advanced image generator with conditional normalization layers in the synthesis step, but capitalizes on the use of contextual information to steer the generation process. Furthermore, three powerful discriminators are employed in an adversarial training procedure to make full use of the available contextual information and improve the realism of the generated results.

3. Context-Driven Virtual Try-On Network

We propose a Context-Driven Virtual Try-On Network (C-VTON) that relies on robust pose matching and contextualized synthesis to generate visually convincing virtual try-on results. Formally, given an input image of a subject, $I \in \mathbb{R}^{w \times h \times 3}$, and a reference image of some target clothing, $C \in \mathbb{R}^{w \times h \times 3}$, the goal of the model is to synthesize a

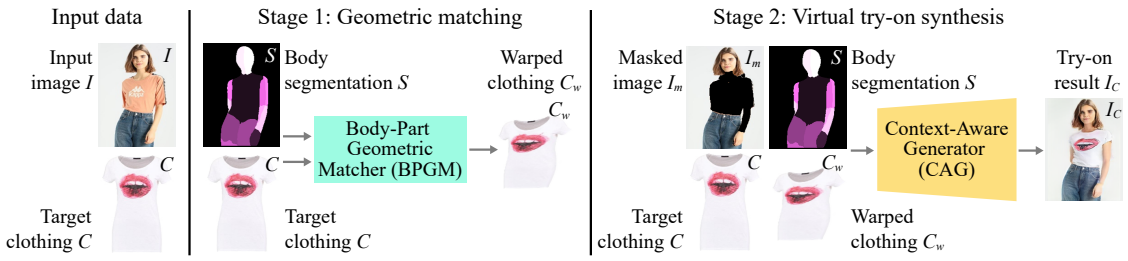


Figure 2. Overview of the proposed Context-Driven Virtual Try-On Network (C-VTON) that given an input image of a subject I and some target clothing C generates a visually convincing virtual try-on image I_C. C-VTON is designed as a two-stage pipeline comprising a Body-Part Geometric Matcher (BPGM) that pre-aligns the target clothing C with the pose of the subject in I and a Context-Aware Generator (CAG) that generates the final try-on image I_C based on the warped clothing and other sources of contextual information.

photo-realistic output image, $I_C \in \mathbb{R}^{w \times h \times 3}$, of the subject from I wearing the target clothing C (see Figure 1).

As illustrated in Figure 2, C-VTON is designed as a two-stage pipeline with two main components: (i) a *Body-Part Geometric Matcher (BPGM)* that warps the reference image of the target clothing C to match the pose of the subject in the input image I, and (ii) a *Context-Aware Generator (CAG)* that uses the output of the BPGM together with various sources of contextual information to generate the final (virtual) try-on result I_C. Details on the two components are given in the following sections.

3.1. The Body-Part Geometric Matcher (BPGM)

The first stage of the C-VTON pipeline consists of the proposed BPGM, and is responsible for estimating the parameters of a Thin-Plate Spline (TPS) [6] transformation that is used to align the target clothing with the pose of the person in the input image I. This allows for approximate positional matching of the target garment and helps to make the task of the generator in the next stage easier. As shown in Figure 3, the BPGM takes a reference image of the target clothing C and body segmentations S as input, and then produces a warped version C_w of the target clothing at the output. The body segmentations $S \in \{0,1\}^{w \times h \times d}$ are generated using the DensePose model from [11] and contain $d = 25$ channels (classes), each corresponding to a different body part. Compared to other virtual try-on architectures that utilize complex clothing-agnostic person representations, e.g., [3, 16, 34], to obtain geometric clothing transformations, BPGM relies on body-part segmentations only, which are sufficient for reliably matching target garments to person images, as we show in our experiments.

BPGM Architecture. The proposed BPGM follows the design of the Geometric Matching Module (GMM) from [34] and consists of two distinct encoders, E_1 and E_2. The first takes the target clothing C as input and generates a corresponding feature representation $\psi_{E_1} \in \mathbb{R}^{w_f \times h_f \times d_f}$. Similarly, the second encoder accepts the body segmentations S at the input and produces a feature representation $\psi_{E_2} \in \mathbb{R}^{w_f \times h_f \times d_f}$ at the output. Here, w_f and h_f represent spatial dimensions that depend on the depth of the encoders and d_f denotes the number of output channels. Next, the feature representations are normalized channel-wise to unit L_2 norm, spatially flattened and organized into a matrix $\Psi_{\{E_1,E_2\}} \in \mathbb{R}^{d_f \times w_f h_f}$, which serves as the basis for computing the correlation matrix $Corr$ [28], i.e.:

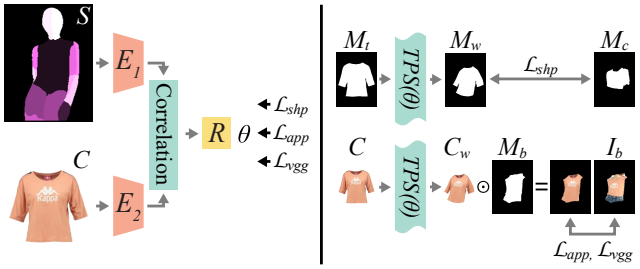


Figure 3. Overview of the Body-Part Geometric Matcher (BPGM). The BPGM architecture is shown on the left and the training losses designed to preserve on-garment textures and overall garment shape are on the right. Unlike competing solutions, BPGM estimates the warping function based on body-part locations only, leading to robust performance even in the presence of challenging poses, e.g., with crossed arms, arms occluding the body, etc.

$$Corr = \Psi_{E_1}^{\top} \Psi_{E_2} \in \mathbb{R}^{(w_f h_f) \times (w_f h_f)}. \quad (1)$$

The correlation matrix $Corr$ is then fed into a regressor, R, which predicts a parameter vector θ (with $2n^2$ dimensions) that corresponds to x and y offsets on an $n \times n$ grid, according to which the target clothing C is warped.

Training Objectives. Three loss functions are used to learn the parameters of the BPGM,, i.e.:

- A **target shape loss** ($\mathcal{L}_{shp}$) that encourages the warping procedure to render the target clothing in a shape that matches the pose of the subject in I, i.e.:

$$\mathcal{L}_{shp} = \|M_w - M_c\|_1 = \|T_\theta(M_t) - M_c\|_1, \quad (2)$$

where M_t and M_w are binary masks corresponding to the original (C) and warped target clothing (C_w), respectively, M_c is a binary mask corresponding to the clothing area in the input image (generated with the segmentation

model of Li *et al*. from [22]), and T_θ denotes the TPS transformation parameterized by θ.

- An **appearance loss** ($\mathcal{L}_{app}$) that forces the visual appearance of the warped clothing C_w within the body area M_b to be as similar as possible to the input image I, i.e.:

$$\mathcal{L}_{app} = \|C_w \odot M_b - I_b\|_1, \quad (3)$$

where $\odot$ is the Hadamard product and M_b a binary masks of the body area (a channel in S), and $I_b = I \odot M_b$.

- A **perceptual loss** ($\mathcal{L}_{vgg}$) that ensures that the target clothing and its warped version contain the same semantic content within the body area, i.e. [19]:

$$\mathcal{L}_{vgg} = \sum_{i}^{n} \lambda_i \|\phi_i(C_w \odot M_b) - \phi_i(I \odot M_b)\|_1, \quad (4)$$

where $\phi_i(\cdot)$ is a feature map generated before each (of the $n = 5$) max-pooling layer of a VGG19 [32] model (pre-trained on ImageNet), and λ_i is the corresponding weight.

Among the above losses, $\mathcal{L}_{shp}$ aims to match the general garment area, while $\mathcal{L}_{app}$ and $\mathcal{L}_{vgg}$ are designed to specifically match the on-garment graphics, without forcing the BPGM matcher to align sleeves, which are often a source of unrealistic transformations used later by the generator.

Finally, the **joint learning objective** for the BPGM is:

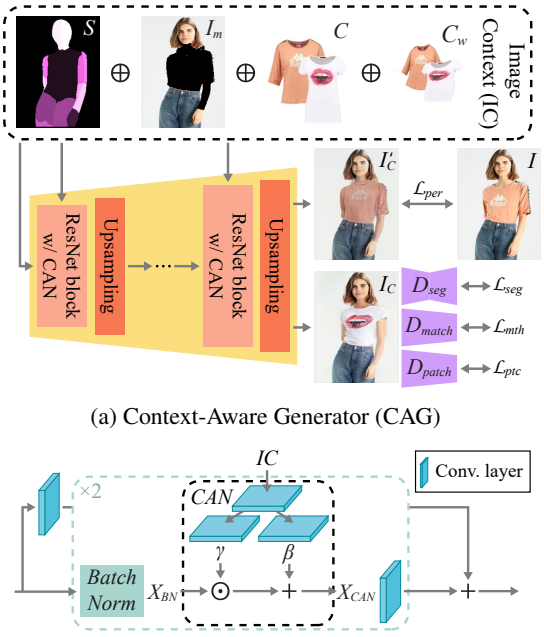
$$\mathcal{L}_{BPGM} = \lambda_{shp}\mathcal{L}_{shp} + \lambda_{app}\mathcal{L}_{app} + \lambda_{vgg}\mathcal{L}_{vgg}, \quad (5)$$

where λ_{shp}, λ_{app} and λ_{vgg} are balancing weights. The parameters of the BPGM are learned over a dataset of input images I with matched images of target clothing C.

3.2. The Context-Aware Generator (CAG)

The second stage of the C-VTON pipeline consists of the Context-Aware Generator (CAG) and is responsible for synthesizing the final virtual try-on image I_C given the warped target clothing and other contextual cues as input. To simplify the discussion in the remainder of the section, we jointly refer to all inputs of the generator as *Image Context (IC)* hereafter, and define it as a channel-wise concatenation of the body-part segmentations S, the input image with masked clothing area I_m (computed as $I_m = I \odot M_c$), and the target and warped clothing images, C and C_w, respectively, i.e., $IC = S \oplus I_m \oplus C \oplus C_w$. A visual illustration of IC is presented at the top of Figure 4(a).

CAG Design. The context-aware generator consists of a sequence of ResNet blocks [13] and ($2\times$) upsampling layers augmented with what we refer to as *Context-Aware Normalization (CAN)* operations. Similarly to recent spatially-adaptive normalization mechanisms used in the field of conditional image synthesis [24, 31], the proposed CAN layers are designed to efficiently utilize the information from the



(a) Context-Aware Generator (CAG)

(b) ResNet block w/ Context-Aware Normalization (CAN)

Figure 4. Overview of the Context-Aware Generator (CAG): (a) the proposed generator with associated losses used during training, (b) a schematic representations of a ResNet block with Context-Aware Normalization (CAN). The proposed CAG is designed to exploit contextual information for the synthesis step - at the input, for activation normalization and through a series of discriminators.

image context IC and feed the generator with critical contextual information. As illustrated Figure 4(a), this is done at different resolutions to ensure (i) that the activations of the generator are spatially normalized at different levels of granularity, and (ii) that the information on the targeted semantic layout and desired appearance of the synthesized output is propagated efficiently throughout the generator.

Each ResNet block of the generator has two inputs: the image context IC, and the activation map from the previous model layer. The only exception here is the first ResNet block of the generator that uses image context IC at the smallest resolution (8×6 pixels) for both inputs, as shown in Figure 4(a). The utilized ResNet blocks consist of a sequence of batch-normalization and convolutional layers repeated twice with CAN operations preceding the convolutional layers. If the output of the batch normalization is denoted as X_{BN}, then the context-aware normalization can formally be defined as: $X_{CAN} = X_{BN} \odot \gamma + \beta$, where $\odot$ denotes the Hadamard product and X_{CAN} is the normalized output. γ and β stand for (spatial) scale and bias parameters with the same dimensionality as X_{BN}. The parameters are learned during the training procedure and computed using three convolutional layers, one of which is shared to first project IC onto a joint embedding space before estimating the values of γ and β with distinct convolutional operations.

Training Objectives. Two types of losses are designed to learn the parameters of C-VTON's generator, i.e.:

- A **perceptual loss** ($\mathcal{L}_{per}$) that encourages the generator to produce a virtual try-on result as close as possible to the reference input image I in terms of semantics. Because the loss assumes the desired target appearance I_C is known, the image context IC is constructed with target clothing C that matches the one in the input image I, i.e.:

$$\mathcal{L}_{per} = \sum_{i}^{n} \tau_i \|\phi_i(I'_C) - \phi_i(I)\|_1, \tag{6}$$

  where $\phi_i(\cdot)$ are feature maps produced by a pretrained VGG19 [32] model before each of the $n = 5$ max-pooling layer, τ_i is the i-th balancing weight, and I'_C is the try-on result generated with the matching target clothing.

- Three **adversarial losses** defined through three discriminators, each aimed at realistic generation of different aspects of the final try-on image, i.e.:

  – The **segmentation discriminator** (D_{seg}), inspired by [31], aims to ensure realistic body-part generation by predicting (per-pixel) segmentation maps S and their origin (real or fake). Given an input image (I or I_C), D_{seg} outputs a $w \times h \times (d+1)$ dimensional tensor, where the first d channels contain segmented body parts and the $(d+1)$-st channel encodes whether a pixel is from a real or generated data distribution. D_{seg} is trained by minimizing a $(d+1)$-class cross-entropy loss, i.e.:

$$\begin{aligned}\mathcal{L}_{D_{seg}} = &- \mathbb{E}_{(I,S)} \left[\sum_{k=1}^{d} \alpha_k \left(S_k \odot \log D_{seg}(I)_k \right) \right] \\ &- \mathbb{E}_{I_C} \left[\log D_{seg}(I_C)_{d+1} \right],\end{aligned} \tag{7}$$

    where the first (segmentation-related) term applies to the (real) input images I and penalizes the first d output channels, and the second term penalizes the last remaining channel generated from the synthesized image $I_C = CAG(IC)$. α_k is a balancing weight calculated as the inverse frequency of the body-part in the given channels of the segmentation map S, i.e., $\alpha_k = hw/\lfloor S_k \rfloor$, where $\lfloor \cdot \rfloor$ is a cardinality operator. The corresponding adversarial loss ($\mathcal{L}_{seg}$) for the generator is finally defined as:

$$\mathcal{L}_{seg} = -\mathbb{E}_{(I_C,S)} \left[\sum_{k=1}^{d} \alpha_k \left(S_k \odot \log D_{seg}(I_C)_k \right) \right]. \tag{8}$$

  – The **matching discriminator** (D_{mth}) aims to encourage the generator to synthesize output images with the desired target clothing by predicting whether the target garment C corresponds to the clothing being worn in either I or I_C. Formally, we train D_{mth} by minimizing the following learning objective:

$$\begin{aligned}\mathcal{L}_{D_{mth}} = &- \mathbb{E}_{(I,C)} \left[\log D_{mth}(I, C) \right] \\ &- \mathbb{E}_{(I_C,C)} \left[\log(1 - D_{mth}(I_C, C)) \right],\end{aligned} \tag{9}$$

  leading to the following generator loss ($\mathcal{L}_{mth}$):

$$\mathcal{L}_{mth} = -\mathbb{E}_{(I_C,C)} \left[\log D_{mth}(I_C, C) \right]. \tag{10}$$

  – The **patch discriminator** (D_{ptc}) contributes towards realistic body-part generation by focusing on the appearances of local patches, $P = \{p_0, ..., p_m\}, p_i \in \mathbb{R}^{w_p \times h_p \times 3}$, centered at $m = 5$ characteristic body-parts, i.e., the neck, and both upper arms and forearms. Different from PatchGAN [15], where patches are extracted implicitly through convolutional operations in the discriminator, we sample the patches from fixed locations based on the segmentation map S. The discriminator is trained to distinguish between real and generated body areas based on the following objective:

$$\begin{aligned}\mathcal{L}_{D_{ptc}} = &- \mathbb{E}_{P_{real}} \left[\log D_{ptc}(P_{real}) \right] \\ &- \mathbb{E}_{P_{fake}} \left[\log(1 - D_{ptc}(P_{fake})) \right],\end{aligned} \tag{11}$$

    where P_{real} and P_{fake} correspond to patches extracted from the real and generated images I and I_C, respectively. The generator loss then takes the following form:

$$\mathcal{L}_{ptc} = -\mathbb{E}_{P_{fake}} \left[\log D_{ptc}(P_{fake}) \right]. \tag{12}$$

The **joint objective** for training C-VTON's context-aware generator is a weighted sum of all loss terms, i.e.:

$$\mathcal{L}_G = \lambda_{per}\mathcal{L}_{per} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{mth}\mathcal{L}_{mth} + \lambda_{ptc}\mathcal{L}_{ptc}, \tag{13}$$

where λ_{per}, λ_{seg}, λ_{mth} and λ_{ptc} denote hyperparameters that determine the relative importance of each loss term.

4. Experiments

In this section we now present experiments that: (i) compare C-VTON to competing models, (ii) demonstrate the impact of key components on performance, and (iii) explore the characteristics of the proposed model.

4.1. Datasets

Following prior work [8, 16, 23, 35], two datasets are selected for the experiments, i.e., VITON [12] and MPV [4].

VITON [12] is a popular dataset for evaluating virtual try-on solutions and consists of $14,221$ training and 2032 testing pairs of images (i.e., subjects and target clothing) with a resolution of 256×192 pixels. For the experiments, duplicate images in the training and test sets are filtered out, leaving 8586 image pairs in the training set and 416 image pairs in the test set. After duplicate removal, the test set

Data	Model	Published	FID↓	LPIPS↓ ($\mu \pm \sigma$)
VITON	CP-VTON [34]	ECCV 2018	47.36	0.303 ± 0.043
	CP-VTON+ [23]	CVPRW 2020	41.37	0.278 ± 0.047
	ACGPN [35]	CVPR 2020	37.94	0.233 ± 0.047
	PF-AFN [8]	CVPR 2021	27.23	0.237 ± 0.049
	C-VTON	This work	**19.54**	$\mathbf{0.108 \pm 0.033}$
MPV	S-WUTON [16]	ECCV 2020	8.188	0.161 ± 0.070
	PF-AFN† [8]	CVPR 2021	6.429	n/a
	C-VTON	This work	**4.846**	$\mathbf{0.073 \pm 0.039}$

† As reported in the original publication

Table 1. Quantitative comparison of C-VTON and competing state-of-the-art models in terms of FID and LPIPS scores - lower is better, as also indicated by the corresponding arrows.

contains unique images not seen during training and allows for a fair comparison between different approaches.

MPV [4] represents another virtual try-on dataset with $35,687$ person images (256×192) wearing $13,524$ unique garments. Different from VITON, MPV exhibits a higher degree of appearance variability with larger differences in zoom level and view point. For the experiments, the images in MPV are prefiltered to feature only (close to) frontal views in accordance with standard methodology, e.g., [8, 16]. The final train and test sets contain $17,400$ paired and 3662 unpaired person and clothing images.

4.2. Implementation Details

C-VTON is implemented in Python using PyTorch. Most modules utilized in the processing pipeline build on ResNet-like blocks [13] that consist of two conv+ReLU layers and a trainable shortcut connection. Architectural details for the main C-VTON components are given below.

The Body-Part Geometric Matcher (BPGM) consists of 2 encoders, E_1 and E_2, with 5 stacked convolutional layers, followed by a downsampling operation, a ReLU activation function and batch normalization. The feature regressor R is implemented with 4 convolutional layers, each followed by a ReLU activation and batch normalization layers. An 18-dimensional linear output layer is used to obtain the parameters (θ) for thin-plate spline transformation.

The Context-Aware Generator (CAG) consists of ResNet blocks with context-aware normalization added before every convolutional layer. We use 6 such blocks each followed by an ($2\times$) upsampling layer. Contextual inputs are resized to match each block's input resolution. An exponential moving average (EMA) is applied over generator weights with a decay value of 0.9999, similarly to [33].

Discriminators. The matching discriminator D_{mth} is implemented with two encoders (one for C and one for I_C) consisting of 6 ResNet blocks each. The output of the encoders is concatenated and fed to a linear layer that produces the final output. The patch discriminator D_{ptc} comprises 4 ResNet blocks arranged in an encoder architecture, with a fully-connected layer on top. The segmentation discriminator D_{seg} has an UNet [29] encoder-decoder architecture and consists of a total of 12 ResNet blocks.

Dataset	Model	Published	vs. C-VTON
VITON	CP-VTON [34]	ECCV 2018	0.766
	CP-VTON+ [23]	CVPRW 2020	0.756
	ACGPN [35]	CVPR 2020	0.674
	PF-AFN [8]	CVPR 2021	0.527
MPV	S-WUTON [16]	ECCV 2020	0.607

Table 2. Results of the human perceptual study reported in terms of the frequency C-VTON generated results were preferred over others. The study was conducted with 100 randomly selected images for each dataset and 70 human participants.

Training Details. The ADAM optimizer [18] is used for the training procedure with a learning rate of $lr_{BPGM} = 0.0001$ for the BPGM, $lr_G = 0.0001$ for the generator and $lr_D = 0.0004$ for the discriminators. All weights in the learning objectives from Eqs. (4), (5), (6) and (13) are set to 1, except for $\lambda_{vgg} = 0.1$ and $\lambda_{per} = 10$. The geometric matcher is trained for 30 epochs and the generator for 100 in all configurations. Source code is available at `https://github.com/benquick123/C-VTON`.

4.3. Quantitative Results

To demonstrate the performance of C-VTON, we first analyze Fréchet Inception Distances (FID [14]) and Learned Perceptual Image Patch Similarities (LPIPS) [37] over processed VITON and MPV test images and conduct a human perceptual study (similarly to [8, 12, 34]) on the MTurk platform. For comparison purposes, we also report result for multiple state-of-the-art models, i.e., CP-VTON [34], CP-VTON+ [23], ACGPN [35], PF-AFN [8] and S-WUTON [16]. Pretrained (publicly released) models are used for the experiments to ensure a fair comparison, except for S-WUTON, where synthesized test images were made available for scoring by the authors of the model.

FID and LPIPS Scores. A quantitative comparison of C-VTON and the selected competitors is presented in Table 1. We note that the results for PF-AFN on MPV are borrowed from [8], since no pretrained model is publicly available for this dataset. As can be seen, C-VTON significantly outperforms all competing models on both datasets. On VITON it reduces the FID score by 28.2% compared to the runner-up and the LPIPS measure by 53.6%. Similar (relative) performances are also observed on MPV, where C-VTON again leads to comparable reductions in FID and LPIPS scores when compared to the runner-ups. We attribute these results to the simplified geometric matching procedure used in C-VTON and the inclusion of diverse contextual information in the final image synthesis step.

Human Perceptual Study. We also evaluate C-VTON through a human perceptual study to analyze the (subjectively) perceived quality of the generated try-on images. In

Figure 5. Comparison of C-VTON (ours) and several recent state-of-the-art models on the VITON (left) and MPV (right) datasets. Areas of interest in the synthesized images are marked with a red bounding box. C-VTON performs considerably better than competing models when synthesizing arms and hands and also better preserves on-shirt graphics. Best viewed electronically and zoomed-in for details.

Model	VITON		MPV	
	FID↓	LPIPS↓	FID↓	LPIPS↓
C-VTON	**19.535**	$\mathbf{0.108 \pm 0.033}$	**4.846**	$\mathbf{0.073 \pm 0.039}$
A1: w/o CAN	24.521	0.162 ± 0.037	12.096	0.159 ± 0.049
A2: w/o BPGM	24.422	0.140 ± 0.036	6.728	0.096 ± 0.046
A3: w/o $\mathcal{D}^\dagger$	21.359	0.109 ± 0.033	5.898	0.076 ± 0.040
A4: w/o EMA	24.571	0.150 ± 0.035	5.304	0.102 ± 0.043

† $\mathcal{D}$ stands for the set of discriminators $\mathcal{D} = \{D_{seg}, D_{mth}, D_{ptc}\}$

Table 3. Ablation study results. For each C-VTON variant (A1-A4) one key component is ablated to demonstrate its contribution.

Figure 6. Qualitative ablation-study results. The aggregation of all component results in noticeable improvements in sleeve and arm generation, on-garment graphics and realistic garment shapes.

the scope of the study, participants were shown the original input image, the target garment and two distinct try-on results, where one was always the result of C-VTON and the other was generated by one of the competing solutions. The participants had to choose the more convincing of the two images based on multiple factors, i.e.: texture transfer quality, arm generation capabilities, pose preservation, and overall quality of results. 100 randomly selected images from each dataset were used for the study, which featured 70 participants in total. The results in Table 2, reported in terms of frequency C-VTON generated results were preferred over others, show that the proposed approach was clearly favored among the human raters.

4.4. Qualitative Results

Next, we explore the performance of C-VTON through visual comparisons with competing models in Figure 5. Due to the unavailability of a pretrained PF-AFN model for MPV, C-VTON is only compared to S-WUTON on this dataset. As can be seen from the presented examples, the proposed approach generates the most convincing virtual try-on results and performs particularly well with hand and on-shirt graphics synthesis. The results clearly show that visually convincing virtual try-on results can be produced with C-VTON even with subjects imaged in difficult poses and with challenging arm/hand configurations.

Among the evaluated competitors, PF-AFN produces the most convincing results on the VITON dataset. However, as illustrated by the presented examples, the method sometimes does not preserve arms, the initial body shape and/or the pose of the subjects, whereas C-VTON fares much better in this regard. The remaining approaches, i.e., CP-VTON, CP-VTON+ and ACGPN, produce less convincing results and often fail to preserve certain (non-transferable) image parts (e.g. trousers and skirts) and textures from the target garment. On MPV, S-WUTON similarly struggles to preserve arms and body shape, while our model synthesizes both well. The excellent performance of C-VTON in this regard is the results of the body-part segmentation procedure used and the set of carefully designed discriminators that ensure realism of the generated images.

4.5. Ablation Study

C-VTON relies on several key components to facilitate image-based virtual try-on. To demonstrate the impact of

Figure 7. Sample results with multiple target garments and different subjects. Challenging examples with long to short-sleeved garment transfer are presented. Best viewed zoomed-in for details.

Figure 8. Comparison of the geometric matching module (GMM) from [34] and our Body-Part Geometric matcher (BPGM). S-w-GMM: Synthesis with GMM, S-w-BPS: Synthesis with BPGM.

these components on performance, an ablation study is conducted. Specifically, four C-VTON variants and implemented, i.e.: (i) C-VTON without CAN operations (A1), (ii) C-VTON without the BPGM (A2), (iii) C-VTON without the discriminators (A3), and (iv) C-VTON without the exponential moving average - EMA (A4).

The results in Table 3 show that FID and LPIPS scores increase when any of the key components is ablated. Interestingly, the absence of CAN operations seems to affect results the most, while the discriminators appear to contribute the least. However, when looking at the visual examples in Figure 6, we see that the discriminators critically affect the final image quality despite the somewhat smaller change in quantitative scores. The difference is especially noticeable when comparing sharpness and artefacts when training C-VTON with or without the discriminators. Furthermore, CAN operations contribute to sleeve and arm generation, the BPGM to a higher quality of on-garment graphics, and EMA to more realistic garment shapes. C-VTON combines these contributions without creating new artefacts, as illustrated by the results for the complete model.

4.6. Strengths and Weaknesses

Finaly, we demonstrate some of C-VTON's strengths and weaknesses by: (i) presenting virtual try-on results for multiple target garments and different subjects, (ii) highlighting the benefits of the proposed geometric matcher, and (iii) illustrating some of the model's limitations.

Figure 9. Examples of less convincing try-on results. With certain image characteristics C-VTON generates blurry clothing edges and only partially transfers target garments. Zoom in for details.

Multiple Targets and Subjects. Figure 7 shows visual try-on results for two distinct subjects and 6 different target garments with varying sleeve lengths. Note that despite the fact that the arms are completely covered in the input images, C-VTON is able to generate realistic try-on results with convincing arm appearances. Varying sleeve lengths are also transferred well onto the synthesized images.

GMM vs. BPGM. Figure 8 shows a comparison between the original geometric matching module (GMM) from [34] and the proposed body-part geometric matcher (BPGM). Note that our BPGM generates more realistic warps that better preserve the shape and texture of the target clothing in the final try-on result. As illustrated by the example in the top row, the better alignment ensured by the BPGM leads to a correctly rendered V-neck. Similarly, on-shirt graphics are better preserved when the proposed BPGM is used instead of the original GMM, as seen by the example in the bottom row of Figure 8.

Limitations. Issues with the masking procedure (of I_m) when generating the image context IC, loose clothing in the input images, and the inability of the model to differentiate between the front and backside of the target garment C are among the main causes for some of the less convincing virtual try-on results produced with C-VTON. These causes lead to unrealistic and soft garment edges, incorrectly synthesized clothing types and improperly rendered neck areas. Similar limitations are also observed with the competing models, as seen from the presented examples in Figure 9.

5. Conclusion

In this paper, we proposed C-VTON, a novel approach to image-based virtual try-on capable of synthesizing high-quality try-on results across a wide range of input-image characteristics. The model was evaluated in extensive experiments on the VITON and MPV datasets and was shown to clearly outperform the state-of-the-art. Additional results that further highlight the merits of the proposed approach are available in the Supplementary Material.

References

[1] Rose Francoise Bertram and Ting Chi. A Study of Companies' Business Responses to Fashion E-Commerce's Environmental Impact. *International Journal of Fashion Design, Technology and Education*, 11(2):254–264, 2018.

[2] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion Meets Computer Vision: A Survey. *ACM Computing Surveys*, 54(4):1–41, 2021.

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14131–14140, June 2021.

[4] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards Multi-Pose Guided Virtual Try-on Network. In *International Conference on Computer Vision (ICCV)*, pages 9026–9035, 2019.

[5] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion Editing With Adversarial Parsing Learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8120–8128, 2020.

[6] Jean Duchon. Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, 1977.

[7] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled Cycle Consistency for Highly-realistic Virtual Try-On. *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-Free Virtual Try-on via Distilling Appearance Flows. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8485–8493, 2021.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

[11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An Image-based Virtual Try-on Network. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6626–6637, 2017.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[16] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do Not Mask What You Do Not Need to Mask: A Parser-Free Virtual Try-On. In *European Conference on Computer Vision (ECCV)*, pages 619–635, 2020.

[17] Nikolay Jetchev and Urs Bergmann. The Conditional Analogy GAN: Swapping Fashion Articles on People Images. In *International Conference on Computer Vision Workshops (ICCV-W)*, pages 2287–2292, 2017.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution using a Generative Adversarial Network. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[20] Hanna Lee and Yingjiao Xu. Classification of Virtual Fitting Room Technologies in the Fashion Industry: From the Perspective of Consumer Experience. *International Journal of Fashion Design, Technology and Education*, 13(1):1–10, 2020.

[21] Hyug Jae Lee, Rokkyu Lee, Minseok Kang, Myounghoon Cho, and Gunhan Park. LA-VITON: A Network for Looking-Attractive Virtual Try-On. In *International Conference on Computer Vision Workshops (ICCV-W)*, 2019.

[22] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1 – 12, 2021.

[23] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. CP-VTON+: Clothing Shape and Texture Preserving Image-based Virtual Try-on. In *Computer Vision and Pattern Recognition Workshops (CVPR-W)*, page 11, 2020.

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

[25] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7365–7375, 2020.

[26] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Transactions on Graphics*, 36(4):1–15, 2017.

[27] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. SwapNet: Image Based Garment Transfer. In *European Conference on Computer Vision (ECCV)*, pages 679–695, 2018.

[28] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional Neural Network Architecture for Geometric Match-

ing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6148–6157, 2017.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[30] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based Animation of Clothing for Virtual Try-on. *Computer Graphics Forum*, 38(2):355–366, 2019.

[31] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only Need Adversarial Supervision for Semantic Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[32] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[33] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You Only Need Adversarial Supervision for Semantic Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2020.

[34] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward Characteristic-Preserving Image-based Virtual Try-on Network. In *European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.

[35] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 7850–7859, 2020.

[36] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation. In *International Conference on Computer Vision (ICCV)*, pages 10510–10519, 2019.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.