VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations

Matteo Fincato[1], Federico Landi[1], Marcella Cornia[1], Fabio Cesari[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia, [2]YOOX NET-A-PORTER GROUP
Email: [1]{name.surname}@unimore.it, [2]{name.surname}@ynap.com

***Abstract*—The large spread of online shopping has led computer vision researchers to develop different solutions for the fashion domain to potentially increase the online user experience and improve the efficiency of preparing fashion catalogs. Among them, image-based virtual try-on has recently attracted a lot of attention resulting in several architectures that can generate a new image of a person wearing an input try-on garment in a plausible and realistic way. In this paper, we present VITON-GT, a new model for virtual try-on that generates high-quality and photo-realistic images thanks to multiple geometric transformations. In particular, our model is composed of a two-stage geometric transformation module that performs two different projections on the input garment, and a transformation-guided try-on module that synthesizes the new image. We experimentally validate the proposed solution on the most common dataset for this task, containing mainly t-shirts, and we demonstrate its effectiveness compared to different baselines and previous methods. Additionally, we assess the generalization capabilities of our model on a new set of fashion items composed of upper-body clothes from different categories. To the best of our knowledge, we are the first to test virtual try-on architectures in this challenging experimental setting.**

I. Introduction

Online fashion shopping has seen a significant increase in recent years, ranking first for development and growth compared to any other e-commerce sector. Today, all brands present their catalog on the web using either their own online shop, third-party online sales, or both. Computer vision and pattern recognition communities have devoted large research efforts to make e-commerce customer experience more efficient and enjoyable, resulting in different solutions for the fashion domain that range from clothing retrieval [1], [2] and segmentation [3] to compatibility prediction [4], [5], [6] and virtual try-on [7], [8], [9]. While the retrieval of similar garments and the compatibility prediction of different clothing items can improve the visual search of online products, virtual try-on architectures can offer the opportunity to try on clothes in a virtual way. Although in some specific cases it can be considered as a rather mature technique (*e.g.* the virtual try-on of eyeglasses on a customer face), the virtual try-on of a generic garment is still an open problem.

In spite of its intrinsic complexity, virtual try-on is an important challenge and a priority for e-commerce companies. Indeed, it could help the users decision on what to buy and potentially reduce the returns of clothes already purchased. It could also make the supply chain more efficient and less expensive, limiting the number of fashion models required to wear a specific garment and reducing the time of manual photo-composition. Both points, and especially the second, become an issue in time of pandemic lock-down, when it is imperative to maximize the distance among workers. Using virtual try-on solutions, fashion companies would require only a few shots of a particular fashion model to present many different clothes in their catalogs.

Fig. 1: Given an in-shop garment and a front-view target model, virtual try-on architectures generate a new image where the input garment is virtually worn while preserving the body pose and identity of the target model. The proposed architecture, VITON-GT, can generate high-quality images thanks to multiple geometric transformations.

The virtual try-on task can be approached both in 2D and 3D space. In 3D space, the approach asks for advanced computer graphic techniques to build human 3D models and render the try-on results [10], [11]. These methods can achieve high quality details and can maintain the physical properties of the clothes, but their large-scale deployment is limited by the excessive manual labor and expensive devices to collect necessary 3D data. Conversely, a fully 2D pipeline on simple RGB data is inexpensive for data preparation and potentially automatic. With this premise, recent work has proposed to reformulate virtual try-on as a conditional image generation problem, thus not requiring 3D information [7], [8]. Given a try-on garment and a target person, a virtual try-on architecture can generate a new image of the target person wearing the input piece of clothing while maintaining the original body pose and identity of the wearer (Fig. 1).

In this paper, we propose a novel solution for virtual try-on and we introduce a new image-based architecture. To preserve the original characteristics of the input clothes (*e.g.* colors,

textures, and logos), we devise a multi-stage geometric transformation technique that can reduce distortions and artifacts in the generated images. We call this new architecture VITON-GT, which stands for VIrtual Try-ON with Geometric Transformations. Specifically, the proposed model includes two different components: a two-stage geometric transformation module and a transformation-guided try-on module. In the first module, our model learns an affine transformation to move the input in-shop clothes closer to the target body pose, and a warping transformation to generate a warped version of the input garment by taking advantage from the previous transformation result. Then, the second component exploits the previously learned geometric transformations and generates the try-on result. To increase the realism of generated images, we integrate adversarial training in the second stage of our architecture and we devise a finetuning strategy that further improves the visual quality of output images. We validate our solution on the standard dataset for image-based virtual try-on [7] in comparison to different baselines and previously proposed methods. Additionally, we assess the ability of VITON-GT to generalize to different try-on clothing categories. To that end, we collect a proprietary set of upper-body clothes from YOOX, an online fashion retailer, that we use as additional analysis of our model performance.

To summarize, our main contributions are as follows:

- We propose a new image-based virtual try-on model that can generate high-quality images thanks to a two-stage geometric transformation of the input garment and a generative network that effectively exploits the learned geometric transformations.
- We demonstrate the effectiveness of the proposed solution both in terms of visual similarity with ground-truth images and realism of the generated try-on results.
- We show the ability of our model to generalize to a collected set of upper-body clothes, consisting of $5,000$ images of five different clothing categories. To the best of our knowledge, we are the first to propose this challenging experimental setting for virtual try-on.

II. Related Work

In this section, we provide an overview of the most important image-based virtual try-on architectures focusing on single- and multi-pose guided models.

Image-based Virtual Try-On. Han *et al.* [7] were the first to present a virtual try-on network relying only on 2D information to transfer the clothes to the corresponding region of the target person. They proposed an encoder-decoder architecture to synthesize a coarse clothed person wearing the target clothing item, and a refinement network to composite the warped garment with the previous coarse result and generate the final output. After this work, many different image-based virtual try-on networks have been introduced [8], [12], [13], [14], [15]. Among them, Wang *et al.* [8] proposed a learnable transformation module to align in-shop clothes to the target person thus improving the generation of virtual try-on images.

On a similar line, Jandial *et al.* [16] presented a two-stage training pipeline consisting of a coarse-to-fine warping network and a texture transfer network conditioned on a learned segmentation mask and trained with a triplet loss strategy to further improve the quality of try-on results. More recently, Yang *et al.* [17] proposed to generate the semantic layout of the target person and predict whether the corresponding image content needs to be generated or preserved, thus leading to more photo-realistic try-on results.

Although all these solutions have focused on the try-on of a specific garment category (*i.e.* typically t-shirts), a few models have tried to generate an image of a model wearing a complete outfit of different try-on clothes [18], [9]. Specifically, in [18] the problem is addressed in a paired setting with the generation of high-resolution images. On the contrary, the model presented in [9] does not exploit any paired training data and transfers clothing items selected from various reference images to a target model.

Multi-Pose Guided Virtual Try-On. While all previous methods can generate try-on images with the same pose of the input model, other approaches can deal with multiple poses to guide the generation [19], [20], [21], [22]. One of the first solutions in this direction has been presented in [19]. In particular, the proposed architecture is composed of a human parsing network to estimate a plausible human parsing with the target clothes and pose, and a warping network followed by a refinement stage to generate the final result. Similarly, Hsieh *et al.* [21] introduced a new model for this task consisting of three stages: pose-guided parsing translation, segmentation region coloring, and salient region refinement. Han *et al.* [22] presented instead a model for pose-guided image generation and virtual try-on that estimates a dense flow between source and target clothing region. Differently, Dong *et al.* [23] went beyond image-based virtual try-on and proposed a video-based solution that learns to synthesize a video of try-on results based on a person image, a try-on garment, and a series of target body poses.

III. Proposed Method

We tackle the problem of 2D single-pose virtual try-on: given a target image I of a clothed person in a given pose and a try-on garment c, our goal is to generate a new image $\tilde{I}$ in which the same person is wearing the item c maintaining the original body pose. We propose a new architecture consisting of two main components: a two-stage geometric transformation module and a transformation-guided try-on module. Our method is depicted in Figure 2 and detailed below.

A. Two-Stage Geometric Transformation Module

In this module, we employ two different geometric transformations, namely affine and thin-plate spline, to warp the in-shop image c of a particular garment.

Affine Transformation. We start from an image I representing a clothed person, and from the image c of the same garment taken from the catalog. From I, we extract the 18 keypoints describing the person pose using a state-of-the-art pose-estimator [24], thus obtaining p_k. In this step, information

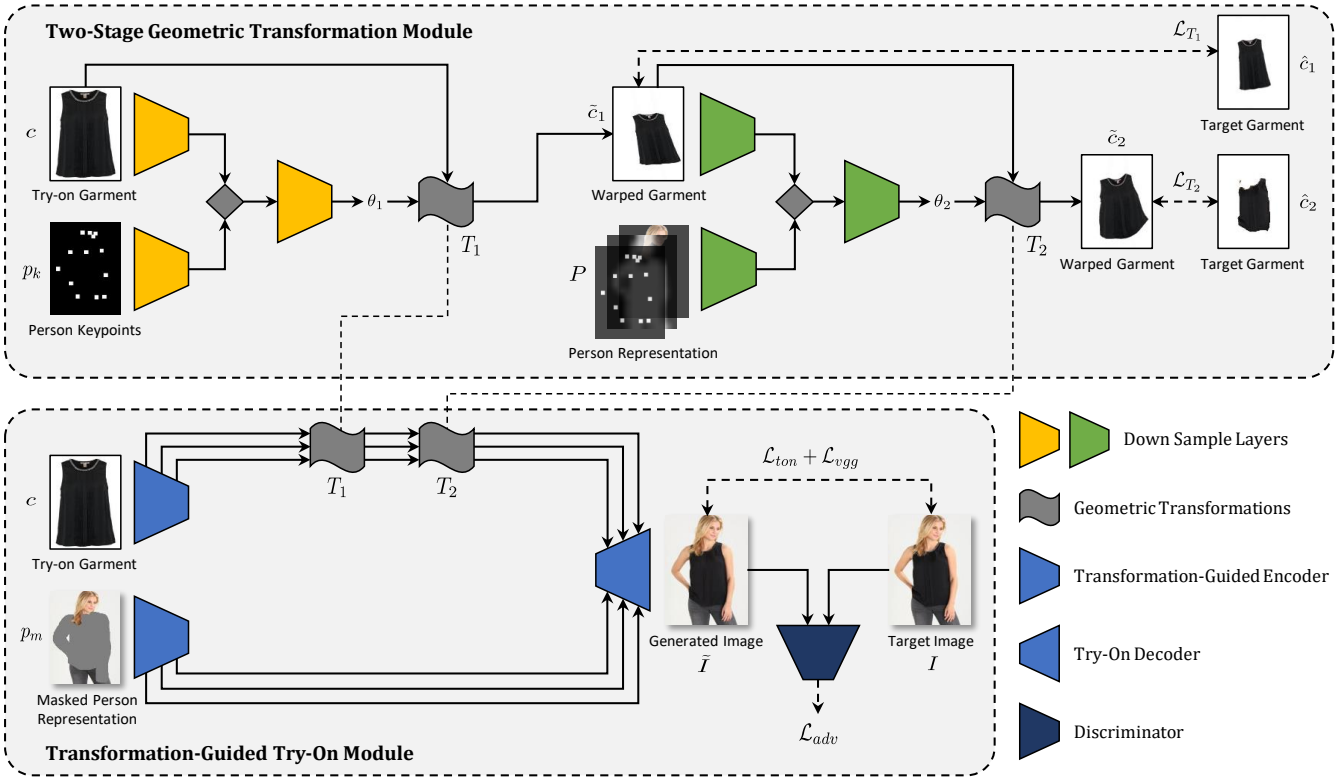

Fig. 2: Overview of the proposed architecture. VITON-GT is composed of a two-stage geometric-transformation module that learns two different geometric transformations, and a transformation-guided try-on module that synthesizes the try-on image exploiting the previously learned transformations.

about the pose is crucial to align the in-shop piece of clothing c with the person displayed in I. As a first step, we encode c and p_k using two convolutional networks: F_{c,χ_c} and F_{p,χ_p}, with parameters χ_c and χ_p respectively. According to [25], we perform ℓ_2-normalization on the vectors and compute a correlation map C as follows:

$$C_{ijz} = F_{c_{i,j}}(c, \chi_c) \cdot F_{p_{m,n}}(p_k, \chi_p) \tag{1}$$

where z is the index for the position (m, n). Finally, we employ a 5-layer neural network to compute the parameters θ_1 for the affine transformation T_1.

We recall that an affine transformation is an automorphism of an affine space. It is a composition of two functions: a translation and a linear mapping. Using an augmented vector, it is possible to represent both of them using a single matrix:

$$\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{0} & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

We decide to adopt an affine transformation and not a complete perspective homography because, in virtual try-on, the deformation in the third dimension is negligible *w.r.t.* the normal distance between the camera and the target. In such a case, affine transformation is very suitable. Additionally, since affine transformation preserves lines and parallelism, we find it very effective to maintain perceptual details such as parallel strips in clothes or logos.

In our architecture, $\theta_1 = \{\boldsymbol{A}, \boldsymbol{b}\}$. Hence, we predict 6 elements in total. After we obtain θ_1, we apply the affinity T_1 to the in-shop garment c, thus getting $\tilde{c}_1$. We aim at bringing the in-shop clothes as close as possible to the position of the clothes worn by the target person. For this reason, we compare $\tilde{c}_1$ with $\hat{c}_1$, computed by applying the ground-truth homography matrix to the in-shop garment c, using a L_1 loss:

$$\mathcal{L}_{T_1} = \|\tilde{c}_1 - \hat{c}_1\|_1 = \|T_1(c, \theta_1) - \hat{c}_1\|_1 . \tag{3}$$

This loss acts as a preliminary supervision for our affinity module. During training, additional signal is given by the objective function outlined in the following paragraph.

Thin-Plate Spline Transformation. The thin-plate spline (TPS) is a commonly used basis function to represent coordinate mappings from $\mathbb{R}^2$ to $\mathbb{R}^2$. The idea of using the TPS to warp the in-shop clothes for virtual try-on was first presented in [8]. Since it can easily modify the aspect of a garment to make it fit a human body in terms of shape and convexity while operating on a 2D image, it is the *standard the facto* in this task. However, our finding is that the results from our previous module can greatly ease the warping task. For this reason, the main input for this block is the transformed garment $\tilde{c}_1$ from the affine transformation module. We also employ a person representation P, which is a 22-channel structure composed of an 18-channel feature map of human keypoints, an RGB agnostic representation of the target person, and a binary mask

of the body shape. The three networks and the correlation function used in this building block are analogous to the ones used for the affine transformation. Here, we perform regression to predict the spatial transformation parameters θ_2. These parameters are used, along with the transformed clothes $\tilde{c}_1$, in the TPS transformation module to generate the final output $\tilde{c}_2 = T_2(\tilde{c}_1, \theta_2)$. The loss function used in this step is a L_1 distance between the obtained warped result $\tilde{c}_2$ and the cropped version of the garment $\hat{c}_2$ obtained from the ground-truth image:

$$\mathcal{L}_{T_2} = \|\tilde{c}_2 - \hat{c}_2\|_1 = \|T_2(\tilde{c}_1, \theta_2) - \hat{c}_2\|_1 . \tag{4}$$

B. Transformation-Guided Try-On Module

The goal of the previous module is to warp the in-shop clothing item c according to the pose of the reference person depicted in I. In this stage, we want to generate an output image $\tilde{I}$ representing the reference person wearing c. To that end, we employ a U-Net architecture [26] consisting in two main components: a transformation-guided encoder composed of two different branches (one for try-on clothes and one to encode person representations), and a try-on decoder.

Transformation-Guided Encoder. In this block we aim at creating two different representations, one for try-on clothes c and one for person representations, that our decoder will then use to synthesize a proper try-on image. To achieve this goal, we divide our U-Net encoder in two branches with different learnable parameters. When connecting the clothes branch to the decoder, we apply the previously learned spatial transformations. Formally, the skip connections typical of the U-Net architecture no longer perform an identity mapping, but compute instead:

$$T(E^i(c), \theta_1, \theta_2) = T_2(T_1(E^i(c), \theta_1), \theta_2), \tag{5}$$

where $E^i(c)$ indicates the encoded features for the garment c extracted from the i-th layer of the U-Net encoder.

As input for the person representation branch, we employ a masked image of the target person p_m. To obtain p_m, we remove the information regarding the clothes and the upper part of the body from I. In this way, the U-Net generator can only have access to the face, the hair and the lower part of the body of the target person. Finally, we align and concatenate the feature maps coming from the two branches.

Try-On Decoder. The goal of our decoder is to generate the final image of the person wearing the in-shop item of clothing. To guide the generation of the final image $\tilde{I}$, we use three different losses. The first loss is a pixel-level L_1 loss measuring the distance of the generated and the ground-truth images:

$$\mathcal{L}_{ton} = \left\| \tilde{I} - I \right\| . \tag{6}$$

The second objective is the perceptual loss, also known as *VGG loss* [27]. The perceptual loss computes the distance of the two images (generated and ground-truth) using the features extracted with a VGG-19 [28] pretrained on ImageNet [29]:

$$\mathcal{L}_{vgg} = \sum_{i=1}^{5} \left\| \phi_i(\tilde{I}) - \phi_i(I)) \right\| \tag{7}$$

where $\phi_i(I)$ are the feature maps of an image I extracted at the i-th layer of the VGG-19 network. Finally, the try-on decoder is trained adversarially. We use a conditional GAN approach similar to the one presented by Wang *et al.* [30] for image-to-image translation. In this context, the generator G and the discriminator D are competing in a process that aims to optimize a min-max loss $\mathcal{L}_{adv}$ [31]. The generator gets better and better at synthesizing fake images and the discriminator at distinguishing real images to fake ones. In this case, the discriminator is conditioned by both the agnostic person representation and the in-shop clothes.

C. Training

To train VITON-GT, we adopt a two-stage strategy. First, we train our geometric transformation module to produce adequate warping results. This stage acts as a sort of pretraining and helps avoiding unstable training dynamics when learning the two geometric transformations from scratch. Then, we use the learned projections to train our try-on module from scratch. In this phase, it is possible to finetune the parameters of the previous block to further improve the efficacy of the geometric transformations.

Two-Stage Geometric Transformation. The final loss used to train this module is a weighted sum of Eq. 3 and Eq. 4:

$$\mathcal{L}_{GT} = \lambda_1 \mathcal{L}_{T_1} + \lambda_2 \mathcal{L}_{T_2}, \tag{8}$$

where λ_1 and λ_2 act as weighting coefficients balancing the contribution of the two losses to the final training signal.

It is worth noting that in this training step, the affine module is encouraged to perform not only the *oracle* projection (the one aligning the clothes with the targets), but also the projection that helps the most during the generation of the final warped result. In this scenario, the error signal coming from $\mathcal{L}_{T_2}$ is free to influence the parameters of the affine module.

Transformation-Guided Try-On. In this second training step, we aim to learn the optimal parameters to employ in our U-Net architecture for image synthesis. The objective function employed in this step is given by a weighted combination of Eq. 6, Eq. 7, and the adversarial loss:

$$\mathcal{L}_{TON} = \rho_1 \mathcal{L}_{ton} + \rho_2 \mathcal{L}_{vgg} + \rho_3 \mathcal{L}_{adv}, \tag{9}$$

where ρ_1, ρ_2, and ρ_3 are the weighting coefficients for the three losses. Additionally, since the try-on module is driven by the set of parameters θ_1 and θ_2 predicted in the previous block, we are able to back-propagate the error signal in the geometric transformation module to further improve the overall quality of the generated images.

IV. Experimental Evaluation

In this section, we first present the datasets and evaluation metrics used in our experiments. Then, we describe all implementation and training details, and evaluate our solution in comparison with different baselines and previous methods.

A. Datasets

To train and test our model, we first employ the most widely used dataset for virtual try-on proposed in [7]. Then, we test the generalization capabilities of our solution on a subset of upper-body clothes of different categories extracted from the YOOX fashion catalog[1].

VITON Dataset [7]. This dataset contains $16,253$ image pairs composed of an upper-body garment (*i.e.* typically a t-shirt) and a front-view woman model wearing it, both with a resolution of 256×192. Images are divided into training and test set with $14,221$ and $2,032$ image pairs, respectively. During evaluation, the image pairs of the test set are rearranged to form unpaired pairs of clothes and front-view models.

Out-of-Domain Upper-Body Clothes. We collect a proprietary set of upper-body clothes divided in five different categories: *short-sleeve t-shirts*, *long-sleeve t-shirts*, *sleeveless t-shirts*, *shirts*, and *sweatshirts*. Overall, we collect $5,000$ images, $1,000$ for each category. To create unpaired image pairs composed of an upper-body garment and a front-view model, we randomly select $1,000$ front-view model images from the VITON dataset that we use to condition our architecture. Also in this case, all images have a size of 256×192.

B. Evaluation Metrics

We quantitative evaluate our model by using different evaluation metrics that either compare the generated images with the corresponding ground-truths (*i.e.* Structural Similarity) or measure the realism and the visual quality of the generation (*i.e.* Frechét Inception Distance, Kernel Inception Distance, and Inception Score).

Structural Similarity (SSIM, MS-SSIM) estimates the similarity between two images. In addition to the standard structural similarity score (SSIM), we also compute its multi-scale version (MS-SSIM). In both cases, we compute the SSIM scores on the VITON paired test set.

Frechét Inception Distance (FID) measures the difference of two Gaussian distributions [32]. In our experiments, the two distributions are fitted on Inception-v3 [33] activations of real and generated images, respectively.

Kernel Inception Distance (KID) measures the squared maximum mean discrepancy between Inception-v3 representations of real and generated images. Following [34], the final KID values are averaged over 100 different splits of size 100, randomly sampled from each image set. For ease of the reader, averaged KID scores and their standard deviations are multiplied by a factor of 100.

Inception Score (IS) estimates the output statistics of Inception-v3, pre-trained on real images and applied to generated ones [35]. Although it has demonstrated to be less reliable than FID and KID metrics [36], it is widely used to evaluate the image quality of virtual try-on architectures [16], [9].

[1]https://www.yoox.com

TABLE I: Warping quantitative results on VITON test set [7]. We compare our model with CP-VTON that exploits only the TPS transformation to generate the warped clothes.

Model	FID	KID	IS
CP-VTON [8] (TPS only)	101.12	6.80±0.67	3.31±0.35
VITON-GT (Affine + TPS)	**59.53**	**3.27±0.48**	**3.40±0.22**

Fig. 3: Warping qualitative results on VITON test set [7].

C. Implementation Details

Two-stage Geometric Transformation Module. Both affine and TPS components include two feature extraction networks, containing four 2-strided down-sampling convolutional layers with a kernel size of 4 followed by two 1-strided ones with a kernel size of 3. Then, a correlation map is computed and fed to a convolutional network composed of two 2-strided convolutional layers with a kernel size of 4 followed by two 1-strided ones with a kernel size of 3. The output is then passed through a fully connected layer that predicts the parameters of the two geometric transformations. In the case of the affine transformation, the fully connected layer predicts the 6 parameters of the homography matrix defined in Eq. 2 and its weights are initialized to obtain an identity mapping at the beginning of the training. For the TPS, the fully connected layer predicts the coordinate offsets of the TPS anchor points and has an output size of $2 \times 5 \times 5 = 50$. All convolutional layers are followed by batch normalization.

Transformation-Guided Try-On Module. The two encoders contain four U-Net blocks, each composed of two convolutional layers with a kernel size of 3 and a 2-strided max pooling layer with a kernel size equal to 2. The decoder has a symmetric structure where each max pooling is replaced with a 2-strided transposed convolutional layer with a kernel size of 2 that upsamples the feature maps. After each up-sampling, the feature maps are concatenated with the feature maps passed through skip connections. Each convolutional layer is followed by instance normalization.

Fig. 4: Qualitative results on VITON test set [7]. We report the results generated by our complete model, our model without adversarial training, our model without finetuning and adversarial training, and CP-VTON [8].

TABLE II: Quantitative results on VITON test set [7]. Differently from the other evaluation metrics, SSIM and MS-SSIM scores are reported on the paired test set of the dataset.

Model	SSIM	MS-SSIM	FID	KID	IS
CP-VTON [8]	0.789	0.838	19.04	0.93±0.18	2.61±0.14
VITON-GT (no FT, no Adv. Loss)	0.879	0.919	15.32	0.58±0.19	2.72±0.14
VITON-GT (no Adv. Loss)	0.879	0.921	13.01	0.36±0.12	2.73±0.09
VITON-GT	**0.886**	**0.925**	**12.45**	**0.32±0.12**	**2.76±0.11**

The discriminator is composed of three 2-strided and two 1-strided down-sampling convolutional layers, all with a kernel size of 4. After each layer, we use instance normalization and apply Leaky ReLU activation.

Training. To compute the ground-truth predictions $\hat{c}_1$ for the affine transformation module, we first extract four vertices from the garment c, thus approximating a parallelogram. Then, we take the four keypoints of the body pose corresponding to the shoulders and the hips, and compute the reference homography matrix. We first pre-train the two-stage geometric transformation module for 300k iterations. Then, we train the transformation-guided try-on module and finetune the rest for other 300k iterations. For all experiments, we use a batch size of 8 and Adam optimizer [37] with a learning rate of 0.0001. We set λ_1, λ_2, ρ_1, and ρ_2 equal to 1, while we set the the weight of the adversarial loss ρ_3 to 0.125. We run all experiments on an NVIDIA 2080 Ti GPU taking one day to train the two-stage geometric transformation module and around two days to finetune the entire model.

D. Experiments on VITON Dataset

Warping results. To validate the effectiveness of our two-stage geometric transformation module, we evaluate the visual quality of the generated warped clothes. Table I reports the quantitative results in terms of FID, KID and Inception Score on VITON test set. For FID and KID metrics, we extract the Inception-v3 representations of the generated clothes and those of the target clothes cropped from real images. For this experiment, we compare our solution with CP-VTON [8] that only uses TPS transformation to generate the warped clothes. This comparison can be regarded as an ablation study in which we remove the affine module from our architecture. As it can be seen, our model outperforms CP-VTON by a large margin on all the evaluation metrics, thus demonstrating the advantages of including the affine geometric transformation.

Fig. 3 shows some qualitative results for this experiment. Again, we compare the warped clothes generated by our model with those generated by CP-VTON. The results confirm the effectiveness of our solution also from a qualitative point of view. The affine transformation module helps the TPS generating better warped clothes that are closer to the target body pose while reducing artifacts and distortions.

Fig. 5: Failure cases on VITON test set.

Try-on results. In Table II, we report the quantitative results

TABLE III: Quantitative results on different categories of out-of-domain upper-body clothes. Results are reported in terms of Frechét Inception Distance, Kernel Inception Distance, and Inception Score for each category of the dataset.

	Short-Sleeve T-Shirts			Long-Sleeve T-Shirts			Sleeveless T-Shirts			Shirts			Sweatshirts		
Model	**FID**	**KID**	**IS**	**FID**	**KID**	**IS**	**FID**	**KID**	**IS**	**FID**	**KID**	**IS**	**FID**	**KID**	**IS**
CP-VTON [8]	23.81	0.86±0.16	2.41±0.21	31.92	1.85±0.33	2.66±0.18	31.50	1.98±0.34	2.36±0.20	35.38	2.33±0.38	2.43±0.14	31.89	1.57±0.28	2.63±0.15
VITON-GT (no FT, no Adv. Loss)	22.11	0.76±0.16	2.54±0.12	23.74	0.89±0.22	2.69±0.09	27.52	1.42±0.24	2.47±0.18	28.85	1.49±0.27	2.65±0.18	27.00	1.11±0.21	2.63±0.11
VITON-GT (no Adv. Loss)	20.95	0.61±0.16	2.63±0.17	**20.02**	**0.62±0.16**	2.79±0.16	24.30	1.16±0.30	2.47±0.10	25.67	1.18±0.27	2.60±0.17	**24.30**	**0.90±0.17**	2.70±0.14
VITON-GT	**20.73**	**0.57±0.15**	**2.65±0.14**	20.83	0.64±0.17	**2.81±0.18**	**22.88**	**1.01±0.24**	**2.56±0.16**	**25.22**	**1.17±0.27**	**2.62±0.10**	25.59	1.04±0.19	**2.76±0.10**

Fig. 6: Qualitative results on different categories of out-of-domain upper-body clothes. For each category, we report three sample try-on images generated by our model in comparison with those generated by CP-VTON [8].

for the try-on generation task. We compare our complete model with CP-VTON. As an ablative analysis, we progressively remove the finetuning of the geometric transformation module, and the contribution given by adversarial training. In particular, we show the results by training the two modules of our model separately and by removing the adversarial loss (*i.e.* VITON-GT (no FT, no Adv. Loss)) and without adversarial loss only (*i.e.* VITON-GT (no Adv. Loss)). While SSIM and MS-SSIM scores are computed on the paired test set, all other metrics are evaluated on the unpaired setting. As it can be seen, all versions of our model achieve better results than CP-VTON according to all evaluation metrics. In particular, the superior performance of VITON-GT (no FT, no Adv. Loss) compared to CP-VTON further demonstrates the effectiveness of our two-stage geometric transformation module that can help to improve the results of the entire pipeline. Moreover, both finetuning and adversarial training give a significant contribution to the final results.

Fig. 4 shows some qualitative results sampled from VITON test set. By comparing the images generated by VITON-GT with those obtained with CP-VTON, we can notice that the proposed solution can reduce distortions while maintaining textures and details of the original clothes, thus increasing the realism of the generation. Additionally, adversarial training helps to produce lighter and less blurred images (*e.g.* first row-left image, third row-left image, and fourth row-right image) and to avoid loosing important elements of the try-on clothes (*e.g.* the left sleeve in the first row-right image).

Failure cases. For a fully insightful analysis, we report some failure cases in Fig. 5. In these examples, VITON-GT partially fails to generate correct try-on images. In particular, errors can be due to different sleeve lengths between try-on and target model clothes (*i.e.* first row), to challenging details or shapes in the original try-on clothes (*i.e.* second row), or to body poses that partially overlap with the worn clothes (*i.e.* third row).

E. Experiments on Out-of-Domain Upper-Body Clothes

To assess whether our model can generalize to different piece of clothing from the one contained in the VITON dataset, we collect a dataset with diverse items and test VITON-GT on this set of out-of-domain upper-body clothes. In particular, we experiment with five different categories: short-sleeve t-shirts,

long-sleeve t-shirts, sleeveless t-shirts, shirts, and sweatshirts. Table III shows the quantitative results on the five considered categories. Since in this setting there are no image pairs with try-on garment and corresponding target model, we only report the results in terms of FID, KID, and Inception Score. We compare our model with CP-VTON and the two versions of our model without finetuning and adversarial loss. Overall, our model obtains the best results according to all evaluation metrics, surpassing the CP-VTON method on all categories. When comparing the different versions of our model, it can be seen that the adversarial training slightly degrades the performance on long-sleeve t-shirts and sweatshirts while improving the final results on all other categories. This can be partially due to the diversity of try-on clothes in these two categories, which contain almost only long-sleeved clothes and differ from image pairs used to train the networks.

Finally, Fig. 6 shows some qualitative results on different categories of out-of-domain upper-body clothes. Each row of the figure reports three sample results using try-on clothes of a category of the dataset. As it can be seen, our model is able to better preserve the original content of try-on clothes and reduce the distortions caused by a wrong geometric transformation. Additionally, our generated images are in general more realistic and maintain both the sleeve length of the input garment and the body pose of the target model.

V. Conclusion

We have presented VITON-GT, a new image-based virtual try-on model that integrates multiple geometric transformations of the input clothes during the generation of the try-on result. Specifically, our model includes a two-stage transformation of the input that can reduce any distortions caused by a wrong warping operation, and a generative network that exploits the previously learned transformations to generate high-quality and more realistic images. Through extensive experiments on two different datasets, we have demonstrated the effectiveness of our solution *w.r.t.* previously proposed methods. Additionally, we have proposed a novel challenging setup with out-of-domain items, in which the proposed model is shown to outperform other models and baselines.

References

[1] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *ICCV*, 2015.

[2] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016.

[3] M. Manfredi, C. Grana, S. Calderara, and R. Cucchiara, "A complete system for garment segmentation and color classification," *Machine Vision and Applications*, vol. 25, no. 4, pp. 955–969, 2014.

[4] W.-L. Hsiao and K. Grauman, "Creating capsule wardrobes from fashion images," in *CVPR*, 2018.

[5] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, "Learning type-aware embeddings for fashion compatibility," in *ECCV*, 2018.

[6] G. Cucurull, P. Taslakian, and D. Vazquez, "Context-aware visual compatibility prediction," in *CVPR*, 2019.

[7] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An Image-based Virtual Try-On Network," in *CVPR*, 2018.

[8] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *ECCV*, 2018.

[9] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert, "Image Based Virtual Try-On Network From Unpaired Data," in *CVPR*, 2020.

[10] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "Drape: Dressing any person," *ACM Trans. on Graphics*, vol. 31, no. 4, pp. 1–10, 2012.

[11] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. on Graphics*, vol. 36, no. 4, pp. 1–15, 2017.

[12] H. Jae Lee, R. Lee, M. Kang, M. Cho, and G. Park, "LA-VITON: A Network for Looking-Attractive Virtual Try-On," in *ICCV Workshops*, 2019.

[13] T. Issenhuth, J. Mary, and C. Calauzènes, "End-to-End Learning of Geometric Deformations of Feature Maps for Virtual Try-On," *arXiv preprint arXiv:1906.01347*, 2019.

[14] R. Yu, X. Wang, and X. Xie, "VTNFP: An Image-based Virtual Try-on Network with Body and Clothing Feature Preservation," in *ICCV*, 2019.

[15] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, "CP-VTON+: Clothing Shape and Texture Preserving Image-Based Virtual Try-On," in *CVPR Workshops*, 2020.

[16] S. Jandial, A. Chopra, K. Ayush, M. Hemani, B. Krishnamurthy, and A. Halwai, "SieveNet: A Unified Framework for Robust Image-Based Virtual Try-On," in *WACV*, 2020.

[17] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content," in *CVPR*, 2020.

[18] G. Yildirim, N. Jetchev, R. Vollgraf, and U. Bergmann, "Generating high-resolution fashion model images wearing custom outfits," in *ICCV Workshops*, 2019.

[19] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *ICCV*, 2019.

[20] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, and W.-H. Cheng, "Fit-me: Image-based virtual try-on with arbitrary poses," in *ICIP*, 2019.

[21] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *ACM Multimedia*, 2019.

[22] X. Han, X. Hu, W. Huang, and M. R. Scott, "ClothFlow: A flow-based model for clothed person generation," in *ICCV*, 2019.

[23] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "FW-GAN: Flow-navigated Warping GAN for Video Virtual Try-on," in *ICCV*, 2019.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[25] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs," in *CVPR*, 2018.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *NeurIPS*, 2014.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a Nash equilibrium," *NeurIPS*, 2017.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[34] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *ICLR*, 2018.

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," in *NeurIPS*, 2016.

[36] S. Barratt and R. Sharma, "A Note on The Inception Score," in *ICML Workshops*, 2018.

[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.