

Towards High-Resolution Virtual Try-On for Fashion Industry

Davide Morelli¹, Marcella Cornia¹, Fabio Cesari², Rita Cucchiara¹

¹University of Modena and Reggio Emilia, ²YOOX NET-A-PORTER GROUP

¹{name.surname}@unimore.it, ²{name.surname}@ynap.com

Abstract

Image-based virtual try-on strives to transfer the appearance of a clothing item onto the image of a target person. Existing literature focuses mainly on upper-body clothes (*e.g.* t-shirts, shirts, and tops) and neglects full-body or lower-body items. This shortcoming arises from a main factor: current publicly available datasets for image-based virtual try-on do not account for this variety, thus limiting progress in the field. In this research activity, we introduce Dress Code, a novel dataset which contains images of multi-category clothes. Dress Code is more than $3\times$ larger than publicly available datasets for image-based virtual try-on and features high-resolution paired images (1024×768) with front-view, full-body reference models. To generate HD try-on images with high visual quality and rich in details, we propose to learn fine-grained discriminating features. Specifically, we leverage a semantic-aware discriminator that makes predictions at pixel-level instead of image- or patch-level.

1 Introduction

Clothes, fashion, and style play a fundamental role in our daily life and allow people to communicate and express themselves freely and directly. With the advent of e-commerce, the variety and availability of online garments have become increasingly overwhelming for the final user. Consequently, user-oriented services and applications such as virtual try-on [Han *et al.*, 2018; Yang *et al.*, 2020; Minar *et al.*, 2020; Choi *et al.*, 2021] are increasingly important for online shopping, helping fashion companies to tailor the e-commerce experience and maximize customer satisfaction. Image-based virtual try-on aims at synthesizing an image of a reference person wearing a given try-on garment. In this task, while virtually changing clothing, the person’s intrinsic information such as body shape and pose should not be modified. Also, the try-on garment is expected to properly fit the person’s body while maintaining its original texture and details. All these elements make virtual try-on a very active and challenging research topic.

Due to the strategic role that virtual try-on plays in e-commerce, many rich and potentially valuable datasets are

proprietary and not publicly available to the research community [Yildirim *et al.*, 2019; Choi *et al.*, 2021; Li *et al.*, 2021]. Public datasets, instead, either do not contain paired images of models and garments or feature a very limited number of images [Han *et al.*, 2018]. Moreover, the overall image resolution is low (mostly 256×192). Unfortunately, these drawbacks slow down progress in the field. To tackle these problems, we present *Dress Code*: a new dataset of high-resolution images (1024×768) containing more than 50k image pairs of try-on garments and corresponding catalog images where each item is worn by a model. This makes Dress Code more than $3\times$ larger than VITON [Han *et al.*, 2018], the most common benchmark for virtual try-on. Differently from existing publicly available datasets, which contain only upper-body clothes, Dress Code features upper-body, lower-body, and full-body clothes, as well as full-body images of human models (Fig. 1, *left*).

Current architectures for virtual try-on are not optimized to work with clothes belonging to different macro-categories (*i.e.* upper-body, lower-body, and full-body clothes) and full-body images. In fact, that would require learning the correspondences between a particular garment class and the portion of the body involved in the try-on phase. In this research activity, we design an image-based virtual try-on architecture that can anchor the given garment to the right portion of the body. As a consequence, it is possible to perform a “complete” try-on over a given person by selecting different garments (Fig 1, *right*). In order to produce high-quality results rich in visual details, we also introduce a novel parser-based discriminator. This component can increase the realism and visual quality of the results by learning an internal representation of the semantics of generated images, which is usually neglected by standard discriminator architectures [Isola *et al.*, 2017]. This component works at pixel-level and predicts not only real/generated labels but also the semantic classes for each image pixel. Extensive experimental evaluation demonstrates that the proposed approach surpasses the baselines and state-of-the-art competitors in terms of visual quality and quantitative results.

2 Dress Code Dataset

Virtual try-on datasets available are often limited by one or more factors such as lack of variety, small size, low-resolution images, privacy concerns, or proprietary license. We identify



Figure 1: Differently from existing publicly available datasets for virtual try-on, Dress Code features different garments, also belonging to lower-body and full-body categories, and high-resolution images.

four main desiderata that the ideal dataset for virtual try-on should possess: (1) it should be publicly available for research purposes; (2) it should have corresponding images of clothes and reference human models wearing them (*i.e.* the dataset should consist of paired images); (3) it should contain high-resolution images and (4) clothes belonging to different macro-categories (tops and t-shirts belong to the upper-body category, while skirts and trousers are examples of lower-body clothes and dresses are full-body garments). In addition to this, a dataset for virtual try-on with a large number of images is more preferable than other datasets with the same overall characteristics but smaller size. To comply with the above desiderata, we collect a new dataset, called Dress Code, which contains high-resolution images and multi-category fashion items taken from different fashion catalogs of YOOX-NET-A-PORTER Group. Overall, the dataset is composed of 53,795 image pairs: 15,366 pairs for upper-body clothes, 8,951 pairs for lower-body clothes, and 29,478 pairs for dresses.

Although some proprietary and non-publicly available datasets have also been used [Lewis *et al.*, 2021; Li *et al.*, 2021; Yildirim *et al.*, 2019], almost all virtual try-on literature [Wang *et al.*, 2018a; Yang *et al.*, 2020; Ge *et al.*, 2021; Choi *et al.*, 2021] employs the VITON dataset [Han *et al.*, 2018] to train the proposed models and perform experiments. We believe that the use of Dress Code could greatly increase the performance and applicability of virtual try-on solutions. In fact, when comparing Dress Code with the VITON dataset, it can be seen that our dataset jointly features a larger number of image pairs (*i.e.* 53,792 vs 16,253 of the VITON dataset), a wider variety of clothing items (*i.e.* VITON only contains t-shirts and upper-body clothes), and a greater image resolution (*i.e.* 1024×768 vs 256×192 of VITON images).

3 Virtual Try-On with Pixel-level Semantics

Virtual try-on models address the task of generating a new image of the reference person wearing the input try-on garment.

Given the generative nature of this task, virtual try-on methods are usually trained using adversarial losses that typically work at image- or patch-level and do not consider the semantics of generated images. Differently from previous works, we design a novel Pixel-level Semantic Aware Discriminator (PSAD) that can build an internal representation of each semantic class and increase the realism of generated images.

Baseline Architecture. To tackle the virtual try-on task, we begin by building a baseline generative architecture that performs three main operations: (1) garment warping, (2) human parsing estimation, and finally (3) try-on. First, the warping module employs geometric transformations to create a warped version of the input try-on garment. Then, the human parsing estimation module predicts a semantic map for the reference person. Last, the try-on module generates the image of the reference person wearing the selected garment.

Pixel-level Semantic-Aware Discriminator. We draw inspiration from semantic image synthesis literature [Park *et al.*, 2019; Liu *et al.*, 2019] and train our discriminator to predict the semantic class of each pixel using generated and ground-truth images as fake and real examples respectively. In this way, the discriminator can learn an internal representation of each semantic class (*e.g.* tops, skirts, body) and force the generator to improve the quality of synthesized images. Specifically, our discriminator is built upon the U-Net model [Ronneberger *et al.*, 2015], which is used as an encoder-decoder segmentation network. For each pixel of the input image, the discriminator predicts the corresponding semantic class and an additional label (real or generated). Overall, we have $N + 1$ classes (*i.e.* N classes corresponding to the ground-truth semantic classes plus one class for fake pixels) and thus we train the discriminator with a $(N + 1)$ -class pixel-wise cross-entropy loss. In this way, the discriminator prediction shifts from a patch-level classification, typical of standard patch-based discriminators [Isola *et al.*, 2017; Wang *et al.*, 2018b], to a per-pixel class-level prediction.

Experimental Results. We compare our model with CP-

VTON [Wang *et al.*, 2018a], CP-VTON+ [Minar *et al.*, 2020], VITON-GT [Fincato *et al.*, 2020], WUTON [Isenhuth *et al.*, 2020], and ACGPN [Yang *et al.*, 2020], that we re-train from scratch on our dataset using source codes provided by the authors, when available, or our implementations. In addition to these methods, we implement an improved version of [Wang *et al.*, 2018a] (*i.e.* CP-VTON[†]). To validate the effectiveness of our Pixel-level Semantic Aware Discriminator (PSAD), we also test a model trained with a patch-based discriminator [Isola *et al.*, 2017] (Patch) and a baseline trained without the adversarial loss (NoDisc). Following recent literature, we employ evaluation metrics that either compare the generated images with the corresponding ground-truths, *i.e.* Structural Similarity (SSIM), or measure the realism and the visual quality of the generation, *i.e.* Frechét Inception Distance (FID) [Heusel *et al.*, 2017], Kernel Inception Distance (KID) [Bińkowski *et al.*, 2018], and Inception Score (IS) [Salimans *et al.*, 2016].

In Table 1, we report numerical results on the Dress Code test set at different image resolutions. As it can be seen, our model obtains better results than competitors on all image resolutions in terms of almost all considered evaluation metrics. Quantitative results also confirm the effectiveness of PSAD in comparison with a standard patch-based discriminator, especially in terms of the realism of the generated images (*i.e.* FID and KID). PSAD is second to the Patch model only in terms of SSIM, and by a very limited margin. Both model configurations outperform the NoDisc baseline, thus showing the importance of incorporating a discriminator in a virtual try-on architecture.

As an additional experiment on the Dress Code dataset, we present a novel setting in which the try-on is performed twice: first with an upper-body garment, and then with a lower-body item. This fully-unpaired setting aims to push further the difficulty of image-based virtual try-on, as it doubles the number of operations required to generate the resulting image. We remind that this experiment would have not been possible on the standard VITON dataset [Han *et al.*, 2018], as it contains only upper-body clothes. In Figure 2, we report high-resolution qualitative try-on results on sample image pairs extracted from upper-body clothes, lower-body clothes, and dresses, showing also the ability of the proposed model to deal and perform well in a multi-garment setting.

Finally, while quantitative metrics used in the previous experiments can capture fine-grained variations in the generated images, the overall realism and visual quality of the results can be effectively assessed by human evaluation. To further evaluate the quality of generated images, we conduct a user study measuring both the realism of our results and their coherence with the input try-on garment and reference person. In the first test (Realism test), we show two generated images, one generated by our model and the other by a competitor, and ask to select the more realistic one. In the second test (Coherency test), in addition to the two generated images, we include the images of the try-on garment and the reference person used as input to the try-on network. In this case, we ask the user to select the image that is more coherent with the given inputs. All images are randomly selected from the Dress Code test set. Overall, this study involves a

Model	Resolution	SSIM \uparrow	FID \downarrow	KID \downarrow	IS \uparrow
CP-VTON	256 \times 192	0.803	35.16	2.245	2.817
CP-VTON+	256 \times 192	0.902	25.19	1.586	3.002
CP-VTON [†]	256 \times 192	0.874	18.99	1.117	3.058
VITON-GT	256 \times 192	0.899	13.80	0.711	3.042
WUTON	256 \times 192	0.902	13.28	0.771	3.005
ACGPN	256 \times 192	0.868	13.79	0.818	2.924
Ours (NoDisc)	256 \times 192	0.907	13.51	0.704	3.041
Ours (Patch)	256 \times 192	0.909	12.53	0.666	3.043
Ours (PSAD)	256 \times 192	0.906	11.40	0.570	3.036
<hr/>					
CP-VTON	512 \times 384	0.831	29.24	1.671	3.096
CP-VTON [†]	512 \times 384	0.896	10.08	0.425	3.277
Ours (NoDisc)	512 \times 384	0.906	10.32	0.430	3.290
Ours (Patch)	512 \times 384	0.923	9.44	0.246	3.310
Ours (PSAD)	512 \times 384	0.916	7.27	0.394	3.320
<hr/>					
CP-VTON	1024 \times 768	0.853	36.68	2.379	3.155
CP-VTON [†]	1024 \times 768	0.912	9.96	0.338	3.300
Ours (NoDisc)	1024 \times 768	0.908	16.58	0.763	3.121
Ours (Patch)	1024 \times 768	0.922	9.99	0.370	3.344
Ours (PSAD)	1024 \times 768	0.919	7.70	0.236	3.357

Table 1: Try-on results on the Dress Code test set using three different image resolutions.

	CP-VTON	VITON-GT	WUTON	ACGPN	Ours (Patch)
Realism	10.1 / 89.9	46.4 / 53.6	47.2 / 52.8	35.9 / 64.1	34.8 / 65.2
Coherency	11.5 / 88.5	32.1 / 67.9	46.9 / 53.1	23.1 / 76.9	36.9 / 63.1

Table 2: User study results. Our model is always preferred more than 50% of the time.

total of 30 participants, including researchers and non-expert people, and we collect more than 3,000 different evaluations (*i.e.* 1,500 for each test). Results are shown in Table 2. For each test, we report the percentage of votes obtained by the competitor / by our model. We also include a comparison with the Patch baseline. Our complete model is always selected more than 50% of the time against all considered competitors, thus further demonstrating the effectiveness of our solution.

4 Challenges and Future Works

In this research activity, we presented Dress Code: a new dataset for image-based virtual try-on. Dress Code, while being more than 3 \times larger than the most common dataset for virtual try-on, is the first dataset for this task featuring clothes of multiple macro-categories and high-resolution images. We proposed a Pixel-level Semantic-Aware Discriminator (PSAD) that improves the generation of high-quality images and the realism of the results. Through extensive experiments on our newly proposed dataset, we demonstrated the effectiveness of the proposed solution in comparison to different baselines and the current state-of-the-art for virtual try-on. Technological development in several fields (*e.g.* augmented reality, metaverse, computational power in mobile devices, etc.) are increasing the attention of e-commerce and fashion industry and the practical applications of this research activities, requiring at the same time more complex scenarios. For these reasons, further possible lines of work will comprehend multi-modal, 3D, and video settings.

