Dress Code: High-Resolution Multi-Category Virtual Try-On

Davide Morelli[1], Matteo Fincato[1], Marcella Cornia[1], Federico Landi[1], Fabio Cesari[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it
[2]YOOX NET-A-PORTER GROUP, Italy
[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. The Dress Code dataset is publicly available at* https://github.com/aimagelab/dress-code.

1. Introduction

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 [3, 7, 20, 28] are increasingly important for online shopping. Due to the strategic role that virtual try-on plays, many rich and potentially valuable datasets are proprietary and not publicly available to the research community [3, 15, 16, 21, 29]. Public datasets, instead, either do not contain paired images of models and garments or feature a very limited number of images [7]. Moreover, the overall image resolution is low (mostly 256×192). Unfortunately, these drawbacks slow down progress in the field. In this paper, 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 [7], 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. Unfortunately, these works employ non-public datasets to train and test the proposed architectures [3, 29].

Figure 1. Differently from 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.

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 [5, 7, 12, 19, 20, 26, 28, 28, 31]. 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 work, 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). In order to produce high-quality results rich in visual details, we introduce a 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 [10, 27]. This component works at pixel-level and predicts not only real/generated labels but also the semantic classes for each image pixel.

2. Dress Code Dataset

We identify four main desiderata that the ideal dataset for virtual try-on should possess: (1) it should be publicly avail-

Figure 2. Sample image pairs from the Dress Code dataset with pose keypoints, dense poses, and segmentation masks of human bodies.

able for research purposes; (2) it should have corresponding images of clothes and reference human models wearing them (3) it should contain high-resolution images and (4) clothes belonging to different macro-categories (*i.e.* upper body, lower body, dresses). By looking at Table 1, we can see that Dress Code complies with all of the above desiderata, while featuring more than three times the number of images of VITON [7]. To the best of our knowledge, this is the first publicly available virtual try-on dataset comprising multiple macro-categories and high-resolution image pairs. Additionally, it is the biggest available dataset for this task at present, as it includes more than 100k images evenly split between garments and human reference models.

Image collection and annotation. All images are collected from fashion catalogs of YOOX-NET-A-PORTER GROUP, containing both casual clothes and luxury garments. To create a coarse version of the dataset, we select images of different categories for a total of 250k fashion items, each containing 2-5 images of different views of the same product. Using a human pose estimator, we select only those products where the front-view image of the garment and the corresponding full figure of the model are available. After this automatic stage, we manually validate all images and group the products into three categories: upper-body clothes (composed of tops, t-shirts, shirts, sweatshirts, and sweaters), lower-body clothes (composed of skirts, trousers, shorts, and leggings), and dresses. 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. To further enrich our dataset, we use OpenPose [2] to extract 18 keypoints for each human body, DensePose [6] to compute the dense pose of each reference model, and SCHP [17] to generate a segmentation mask of model body parts and clothing items. All model images are anonymized. Sample human model and garment pairs from our dataset with the corresponding additional information are shown in Figure 2.

Comparison with other datasets. Table 1 reports the main characteristics of the Dress Code dataset in comparison with existing datasets for virtual try-on and fashion-related tasks. Although some proprietary and non-publicly available datasets have also been used [15, 16, 29], almost all virtual try-on literature employs the VITON dataset [7] to train the proposed models and perform experiments. We believe

Dataset	Public	Multi-Cat	# Images	# Garments	Resolution
VITON-HD [3]	✗	✗	27,358	13,679	1024×768
O-VITON [21]	✗	✓	52,000	-	512×256
TryOnGAN [15]	✗	✓	105,000	-	512×512
Revery AI [16]	✗	✓	642,000	321,000	512×512
Zalando [29]	✗	✓	1,520,000	1,140,000	1024×768
FashionOn [9]	✓	✗	32,685	10,895	288×192
DeepFashion [19]	✓	✗	33,849	11,283	288×192
MVP [4]	✓	✗	49,211	13,52	256×192
FashionTryOn [32]	✓	✗	86,142	28,714	256×192
LookBook [30]	✓	✓	84,748	9,732	256×192
VITON [7]	✓	✗	32,506	16,253	256×192
Dress Code	✓	✓	107,584	53,792	1024×768

Table 1. Comparison between Dress Code and the most widely used datasets for virtual try-on and other related tasks.

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. Proposed Model

To tackle the virtual try-on task, we start 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. To generate high quality results, we introduce 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. Our complete model is shown in Fig. 3 and detailed in the following.

Warping Module. We follow the warping module proposed in [26]. To train this network, we minimize the L_1 distance between the warped result $\tilde{c}$ and the cropped version of the garment $\hat{c}$ obtained from I. In addition, to reduce visible distortions in the warped result, we employ the second-order difference constraint introduced in [28].

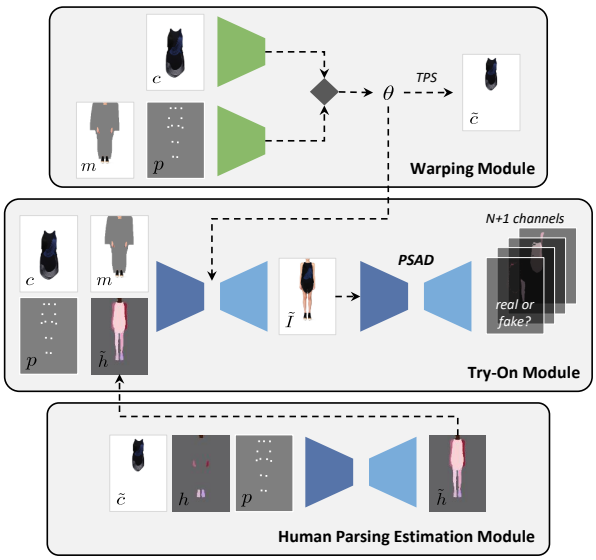

Figure 3. Overview of the proposed architecture.

Human Parsing Estimation Module. This module, based on the U-Net architecture [23], takes as input a concatenation of the warped try-on clothing item $\tilde{c}$, the pose image p, and the masked semantic image h, and predicts the complete semantic map $\tilde{h}$ containing the human parsing for the reference person. This module is trained using a pixel-wise cross-entropy loss between the generated semantic map $\tilde{h}$ and the ground-truth $\hat{h}$.

Try-On Module. Finally, the try-on module produces the image $\tilde{I}$ depicting the reference person described by the triple $(p, m, \tilde{h})$ wearing the input try-on clothing item c. To this end, we employ a modified U-Net model [23] featuring a two-branch encoder and a decoder. The input of the first branch is the original try-on garment c, while the input of the second branch is a concatenation of the pose image p, the masked person representation m, and the one-hot semantic image obtained by taking the pixel-wise argmax of $\tilde{h}$. In the skip connection of the first branch, we apply the previously learned TPS transformation. During training, we exploit a combination of three different loss functions: an L_1 loss between the generated image $\tilde{I}$ and the ground-truth image I, a perceptual loss [13] to compute the difference between the feature maps of $\tilde{I}$ and I, and the adversarial loss $\mathcal{L}_{adv}$ defined below.

Pixel-level Semantic-Aware Discriminator. Most of the existing discriminator architectures work at image- or patch-level, thus neglecting the semantics of generated images. To address this issue, we draw inspiration from semantic image synthesis literature [18, 22, 25] 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. Our discriminator is built upon the U-Net model [23]. For each pixel of the input image, the discriminator predicts the corresponding N semantic class and an additional label (real or generated). 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 [10, 27], to a per-pixel class-level prediction. Due to the unbalanced nature of the semantic classes, we weigh the loss class-wise using the inverse pixel frequency of each class. Formally, the loss function used to train this Pixel-level Parsing-Aware Discriminator (PSAD) can be defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &-\mathbb{E}_{(I,\hat{h})}\left[\sum_{k=1}^{N} w_k \sum_{i,j}^{H\times W} \hat{h}_{i,j,k} \log D(I)_{i,j,k}\right] \\ &-\mathbb{E}_{(p,m,c,\hat{h})}\left[\sum_{i,j}^{H\times W} \log D(G(p,m,c,\hat{h}))_{i,j,k=N+1}\right],\end{aligned} \tag{1}$$

where I is the real image, $\hat{h}$ is the ground-truth human parsing, p is the model pose, m and c are respectively the person representation and the try-on garment given as input to the generator, and w_k is the class inverse pixel frequency.

4. Experimental Evaluation

Dataset and Evaluation Metrics. We perform experiments on our newly proposed dataset using 48,392 image pairs as training set and the remaining 5,400 pairs as test set. During evaluation, the test set is rearranged to form unpaired pairs of clothes and front-view models. We use three different image resolutions: 256×192 (*i.e.* the one typical used by virtual try-on models), 512×384, and 1024×768. To evaluate the results, we employ Structural Similarity (SSIM), Frechét Inception Distance (FID) [8], Kernel Inception Distance (KID) [1], and Inception Score (IS) [24].

Training. We train the three modules separately. Specifically, we first train the warping module and then the human parsing estimation module for 100k and 50k iterations respectively. Finally, we train the try-on module for other 150k iterations. We set the weight of the second-order difference constraint λ_{const} to 0.01 and the weight of the adversarial loss λ_{adv} to 0.1. All experiments are performed using Adam [14] as optimizer and a learning rate equal to 10^{-4}.

Experimental Results. We compare with CP-VTON [26], VITON-GT [5], WUTON [11], and ACGPN [28], that we re-train from scratch on our dataset using source codes provided by the authors, when available, or our re-implementations. In addition to these methods, we implement an improved version of [26] (*i.e.* CP-VTON†) in which we use the masked person m as an additional input

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

Table 2. Try-on results on the Dress Code test set using three different image resolutions.

to the model. To validate the effectiveness of our Pixel-level Semantic Aware Discriminator (PSAD), we also test a model trained with a patch-based discriminator [10] (Patch) and a baseline trained without the adversarial loss (NoDisc).

In Table 2, 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. In Fig. 4, we report a qualitative comparison between the results obtained with our Patch model and the proposed PSAD. In Fig. 5, we compare our results with those obtained by state-of-the-art competitors. Overall, our model with PSAD can better preserve the characteristics of the original clothes such as colors, textures, and shapes, and reduce artifacts and distortions, increasing the realism and visual quality of the generated images.

To further evaluate the quality of generated images, we conduct a user study. In the first test (Realism), we show one image generated by our model and the other by a competitor, and ask to select the more realistic one. In the second test (Coherency), we include also 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.

Figure 4. Qualitative comparison between Patch and PSAD.

Figure 5. Sample try-on results on the Dress Code test set.

	CP-VTON	VITON-GT	WUTON	ACGPN	Ours (Patch)
Realism	10.1 / **89.9**	46.4 / **53.6**	42.0 / **58.0**	35.9 / **64.1**	34.8 / **65.2**
Coherency	11.5 / **88.5**	32.1 / **67.9**	41.6 / **58.4**	23.1 / **76.9**	36.9 / **63.1**

Table 3. User study results. Our model is always preferred more than 50% of the time.

Overall, this study involves a 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 3. For each test, we report the percentage of votes obtained by the competitor / by our model. Our complete model is always selected more than 50% of the time against all considered competitors.

5. Conclusion

In this paper, we presented Dress Code, a new dataset for image-based virtual try-on that, while being more than 3× larger than the most common dataset for virtual try-on, is the first publicly available dataset for this task featuring clothes of multiple macro-categories and high-resolution images. We also introduced a Pixel-level Semantic-Aware Discriminator (PSAD) that improves the generation of high-quality images and the realism of the results.

References

[1] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *CVPR*, 2017. 2

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *ICCV*, 2021. 1, 2

[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 *ICCV*, 2019. 2

[5] Matteo Fincato, Federico Landi, Marcella Cornia, Cesari Fabio, and Rita Cucchiara. VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In *ICPR*, 2020. 1, 3

[6] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In *CVPR*, 2018. 2

[7] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-based Virtual Try-On Network. In *CVPR*, 2018. 1, 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. *NeurIPS*, 2017. 3

[9] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *ACM Multimedia*, 2019. 2

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *CVPR*, 2017. 1, 3, 4

[11] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do Not Mask What You Do Not Need to Mask: a Parser-Free Virtual Try-On. In *ECCV*, 2020. 3, 4

[12] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. SieveNet: A Unified Framework for Robust Image-Based Virtual Try-On. In *WACV*, 2020. 1

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[14] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 3

[15] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. TryOnGAN: Body-Aware Try-On via Layered Interpolation. *ACM Trans. Gr.*, 40(4), 2021. 1, 2

[16] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward Accurate and Realistic Outfits Visualization with Attention to Details. In *CVPR*, 2021. 1, 2

[17] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-Correction for Human Parsing. *arXiv preprint arXiv:1910.09777*, 2019. 2

[18] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 3

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2

[20] 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 *CVPR Workshops*, 2020. 1

[21] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image Based Virtual Try-On Network From Unpaired Data. In *CVPR*, 2020. 1, 2

[22] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 3

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *NeurIPS*, 2016. 3

[25] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021. 3

[26] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 3, 4

[27] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *CVPR*, 2018. 1, 3

[28] 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 *CVPR*, 2020. 1, 2, 3, 4

[29] Gokhan Yildirim, Nikolay Jetchev, Roland Vollgraf, and Urs Bergmann. Generating high-resolution fashion model images wearing custom outfits. In *ICCV Workshops*, 2019. 1, 2

[30] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *ECCV*, 2016. 2

[31] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An Image-based Virtual Try-on Network with Body and Clothing Feature Preservation. In *ICCV*, 2019. 1

[32] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually Trying on New Clothing with Arbitrary Poses. In *ACM Multimedia*, 2019. 2