Dual-Branch Collaborative Transformer for Virtual Try-On

Emanuele Fenocchi[1], Davide Morelli[1], Marcella Cornia[1], Lorenzo Baraldi[1],
Fabio Cesari[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia, Italy [2]YOOX NET-A-PORTER GROUP, Italy
[1]202855@studenti.unimore.it,{name.surname}@unimore.it [2]{name.surname}@ynap.com

Abstract

Image-based virtual try-on has recently gained a lot of attention in both the scientific and fashion industry communities due to its challenging setting and practical real-world applications. While pure convolutional approaches have been explored to solve the task, Transformer-based architectures have not received significant attention yet. Following the intuition that self- and cross-attention operators can deal with long-range dependencies and hence improve the generation, in this paper we extend a Transformer-based virtual try-on model by adding a dual-branch collaborative module that can exploit cross-modal information at generation time. We perform experiments on the VITON dataset, which is the standard benchmark for the task, and on a recently collected virtual try-on dataset with multi-category clothing, Dress Code. Experimental results demonstrate the effectiveness of our solution over previous methods and show that Transformer-based architectures can be a viable alternative for virtual try-on.

1. Introduction

The interest of the fashion industry for Computer Vision applications is consistently growing, as scientific progress allows bringing more effective models into production. In this context, image-based virtual try-on has received considerable attention in the last few years, with the introduction of architectures that can generate a novel image of a target person virtually wearing a given garment [9,20]. According to the current literature, the virtual try-on task requires to tackle a variety of challenges: (1) warping the garment according to the body shape and pose of the target person; (2) transferring the texture of the garment on the target person without losing important details; (3) merging the image of the target person with the warped result in a plausible way; and (4) render light and shades of the final image correctly, to ensure realism.

Since the seminal work of Han *et al*. [9], a lot of progress has been done in the development of virtual try-on architectures. One of the most important advancements is that made by the CP-VTON architecture [20], which introduced a warping module to compute a learnable Thin-Plate Spline (TPS) transformation for warping the in-shop garment in a robust way. This approach, which has later been used in several works [5,6,12,15,22], allows retaining fine-grained details of the clothing, therefore tackling problems (1) and (2). Nevertheless, TPS can fail when the models' poses are too complex. To mitigate this issue, Yang *et al*. [22] proposed a second-order constraint to limit the degree of clothing deformation. While this approach can be beneficial when dealing with extreme poses, it also limits the ability of the warping module to model real-world transformations and hence fails to comply with the first of the aforementioned challenges. Some works have later proposed to hallucinate the label map with in-shop clothes to compensate for this deficiency [1, 4, 8]. In line with Ren *et al*. [17], we instead argue that failures in estimating TPS parameters can be attributed to the inability of the regression network to estimate the long-range dependency between the in-shop garment and the target images.

A second fundamental component of virtual try-on architectures is the try-on module, which aims at fitting the warped garment on the target person. This step usually deals with issues (3) and (4), and also refines the warped clothing shape. A classical approach is that of using a U-Net architecture [18] for the generative task [2, 15, 20, 22], feeding the person image and the warped cloth to the architecture. Other works have also employed a two-branch network, where one branch takes as input the person and the other the in-shop cloth and warping information [5, 11]. Finally, CIT [17] has proposed to apply a Transfomer-based architecture [19] and cross-modal attention mechanisms to the inputs before feeding them to the generative network.

Drawing inspiration from CIT, in this paper we propose a Transformer-based virtual try-on network that employs a dual-branch collaborative module and exploits cross-modal information at generation time. Experimentally, our approach can achieve more realistic and accurate results than other state-of-the-art methods when evaluated on the VI-

TON dataset [9] and when tested on a virtual try-on dataset with multi-category clothing items, Dress Code [16]. We name our approach DBCT, short for *Dual-Branch Collaborative Transformer*.

2. Proposed Model

Our approach is based on the method presented in [17]. To enhance the image generation, we modify the inputs of both the warping and try-on modules, and propose architectural improvements for the try-on module. In the following section, we describe our approach and detail its improvements with respect to previous literature.

2.1. Warping Module

Most of the methods trained with the VITON dataset [9] fail to consider two fundamental aspects of virtual try-on generation, *i.e*. the preservation of background and non-target clothing in the generated image. In order to meet these requirements, we employ a new person representation obtained by subtracting only the body parts and garment of interest and retaining the unchangeable parts (*i.e*. hands, feet, head, hairs). Our complete person representation is composed by the three channels masked image and by the 18 channel keypoints representing the body pose of the reference person.

Matching Block. The proposed person representation p along with the try-on in-shop garment c are then individually processed by a Transformer encoder, *i.e*.

$$\begin{aligned} p' &= \text{Encoder}\left(p\right), \\ c' &= \text{Encoder}\left(c\right). \end{aligned} \tag{1}$$

Then, cross-modal attention is performed in a symmetric manner between the refined representations of the person and the in-shop garment. To do so, in the first cross-modal block the person representation branch is used as key while the clothing branch as query and value, while in the second cross-modal block the order is inverted. This allows to fully exploit the interaction between the inputs and capture long-range dependencies:

$$X_{cross} = \text{CrossAttn}\left(p', c'\right) \cup \text{CrossAttn}\left(c', p'\right)), \tag{2}$$

where $\cup$ indicates concatenation and $\text{CrossAttn}(x, y)$ indicates a cross-attention operation in which x is employed to form queries and y is employed to form keys and values. After concatenating the output, a linear projection is applied to obtain the TPS parameters θ.

This module is trained by evaluating how close the newly warped garment is to the garment worn by the reference person, applying an L_1 loss together with the grid regularization loss proposed in [22].

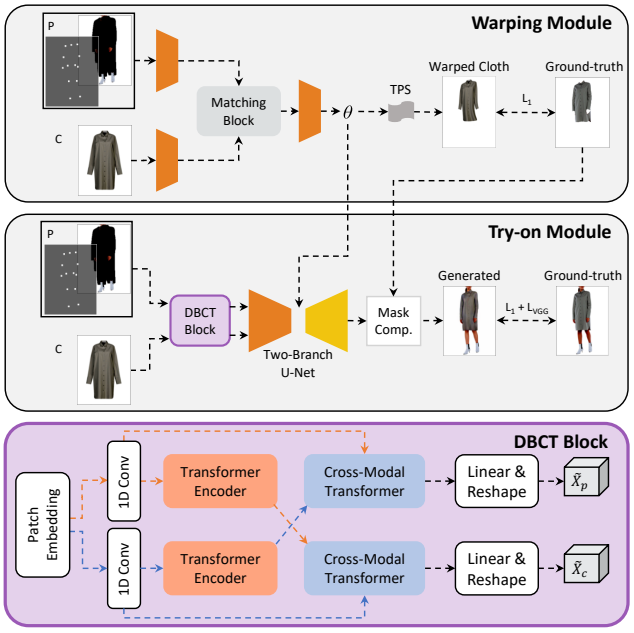

Figure 1. Overview of the proposed architecture.

2.2. Try-On Module

In the try-on module, we propose three major modifications to improve the quality of the generated images while reducing the memory occupancy of the overall network. Firstly, we propose to reduce and modify the input of the module to reduce complexity and memory consumption. Secondly, we design a novel Dual-Branch Collaborative Transformer block (DBCT) that improves the quality and realism of the results. Finally, we propose to substitute the standard U-Net used in some previous methods [17, 22] with a two-branch U-Net. An overview of the resulting architecture is shown in Figure 1.

Dual-Branch Collaborative Transformer. Instead of using as input the person representation, the cloth mask, and the warped cloth used in [17], we employ the person representation p employed in the warping module and the in-shop garment c. This modification allows reducing the number of cross-modal Transformer blocks from three to two thus reducing the network size and fastening the training.

In this block, the two inputs described above are first passed each through a patch embedding module and then a 1D temporal convolution, in order to extract relevant features. Then the Transformer encoder and subsequently the cross-modal Transformer is applied on each branch. The first performs self-attention to exploit long-range dependencies, while the second enhances the person information with the garment one and vice-versa. We name the output of the Transformer encoder respectively $X_p^{'}$ for the person branch and $X_c^{'}$ for the garment branch. Differently from the cross-

Model	Upper-body			Lower-body			Dresses			All		
	SSIM ↑	FID ↓	KID ↓	SSIM ↑	FID ↓	KID ↓	SSIM ↑	FID ↓	KID ↓	SSIM ↑	FID ↓	KID ↓
CP-VTON [20]	0.812	46.99	3.236	0.782	54.66	3.656	0.816	34.95	1.759	0.803	35.16	2.245
CP-VTON+ [15]	0.863	28.93	1.856	0.819	41.37	2.506	0.826	32.27	1.630	0.836	25.19	1.586
CIT [17]	0.860	26.41	1.496	0.834	31.77	1.753	0.810	35.58	1.734	0.835	21.99	1.313
CP-VTON† [20]	0.898	23.03	1.338	0.887	26.96	1.409	0.838	33.04	1.668	0.874	18.99	1.117
CIT† [17]	0.884	24.01	1.323	0.859	28.11	1.466	0.815	41.09	2.055	0.853	20.92	1.120
DBCT	**0.911**	**17.30**	**0.893**	**0.895**	**22.93**	**0.991**	**0.855**	**22.66**	**0.876**	**0.887**	**13.46**	**0.780**

Table 1. Try-on results on the Dress Code test set [16]. † denotes the improved version of CP-VTON [20] and CIT [17].

modal attention performed in the warping module, we respectively use for each stream the Transformer encoder output $X_p^{'}$ and $X_c^{'}$ as key and value and the 1D convolution output X_p and X_c as query:

$$\hat{X}_p = \text{CrossAttn}\left(X_p, X_c'\right) \tag{3}$$

$$\hat{X}_c = \text{CrossAttn}\left(X_c, X_p'\right) \tag{4}$$

The features are then linearly projected, reshaped, and fed to the two-branch U-Net.

Two-Branch U-Net. We finally propose to substitute the standard single-branch U-Net used in previous methods [17] with a two-branch U-Net where the former branch takes into account the person information while the latter considers the in-shop clothing one, both pre-processed as described above. Having as input the in-shop garment, the warping is performed injecting the warping parameters θ on the residual connection of the in-shop clothing branch. The network outputs both the generated image and the warped clothing mask, which are used along with the warped clothing to perform a mask composition [17, 20]. Finally, an L_1 loss and a perceptual L_{VGG} loss [13] are applied between the generated image and the ground-truth image.

3. Experimental Evaluation

Datasets. We perform experiments on two virtual try-on datasets: VITON [9], that can be considered as the most widely used dataset for the task, and Dress Code [16], a recently proposed benchmark with multi-category clothing items and large image variety. VITON is composed of 16,253 model-garment pairs and presents only upper-body garments (*i.e*. manly t-shirts). Images are divided in training and test split with 14,221 and 2,032 image pairs respectively. Dress Code instead contains 53,795 model-garments pairs divided in three different clothing categories (*i.e*. upper-body, lower-body, and dresses). In our experiments, we use 5,400 image pairs for test (1,800 for each category) and the rest for training.

Training and Implementation Details. Each module of our architecture is trained separately. In particular, we train the warping and try-on modules for 200k and 250k iterations, respectively. In all experiments, we use a batch size of 4 and Adam as optimizer [14] with a learning rate equal to 0.0001. For both datasets, experiments are performed using an image resolution of 256×192.

Model	SSIM ↑	FID ↓	KID ↓
CP-VTON [20]	0.798	19.06	0.906
CP-VTON+ [15]	0.828	16.31	0.784
SieveNet [12]	0.766	14.65	-
ACGPN [22]	0.845	-	-
DCTON [7]	0.830	14.82	-
CIT [17]	0.827	15.82	0.626
DBCT	**0.892**	**13.68**	**0.452**

Table 2. Try-on results on the VITON test set [9].

Evaluation Protocol and Metrics. According to the virtual try-on literature, evaluation is performed both in paired and unpaired settings. In the paired setting, the input try-on garment corresponds to the one originally worn by the reference model, while for the unpaired one, we rearrange images to form unpaired pairs of clothes and target models. For evaluation, we use the Structural Similarity (SSIM) [21] to compute the perceived quality of the generated image with respect to the ground-truth, and the Frechét Inception Distance (FID) [10] and Kernel Inception Distance (KID) [3] to measure the realism and the visual quality of the generation.

Results on VITON Dataset. We first evaluate our approach on the VITON dataset. In this case, we compare our results with those generated by CP-VTON [20], CP-VTON+ [15], SieveNet [12], ACGPN [22], DCTON [7], and CIT [17], using the results reported in the original papers or, when available, the official source codes to extract the generated images and compute the evaluation metrics. Results are reported in Table 2 in terms of SSIM, FID, and KID. As it can be seen, our model achieves the best results according to all evaluation metrics. In particular, compared to the CIT model, the proposed DBCT solution can reduce FID and KID scores by 2.14 and 0.17 points respectively, thus confirming the effectiveness of our approach.

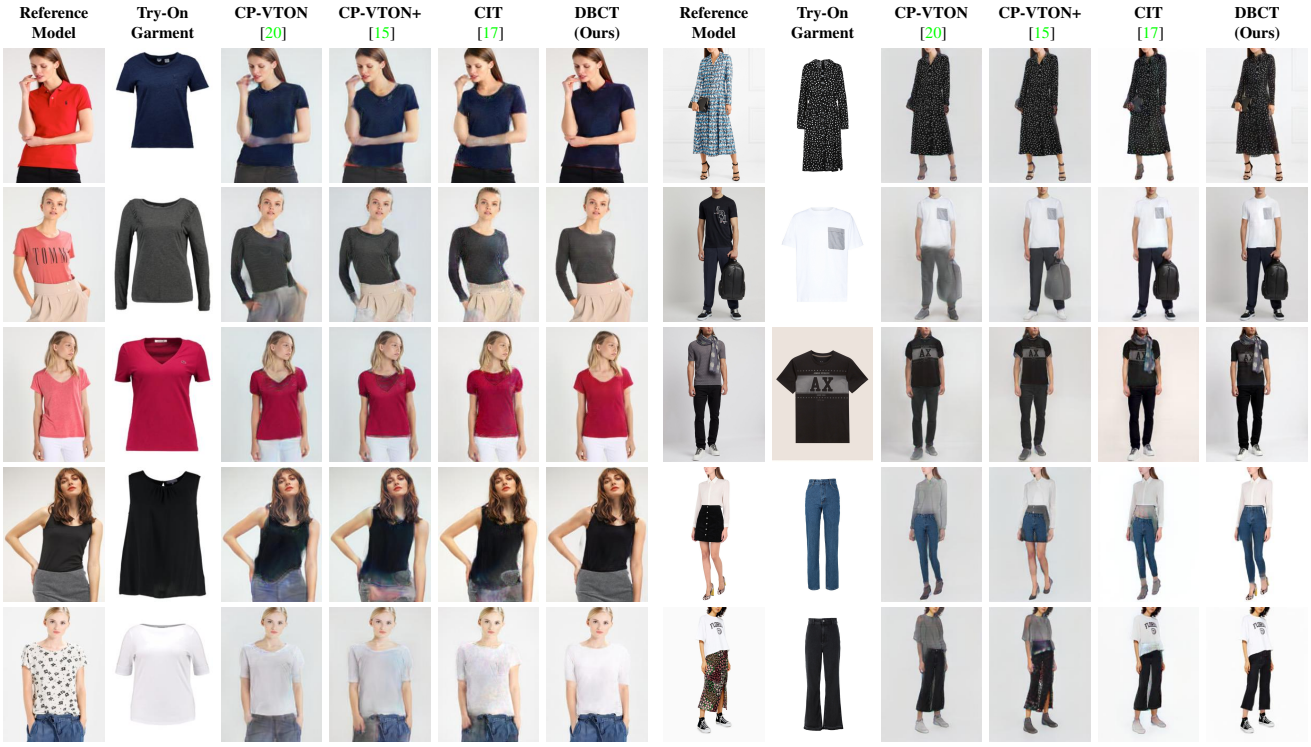

Figure 2. Qualitative results generated by our model and competitors on VITON (left) and Dress Code (right) image pairs.

	Two-Branch U-Net	Cross-Modal Transformer	Transformer Encoder	FID ↓
	✓			13.97
	✓	✓		13.86
DBCT	✓	✓	✓	**13.46**

Table 3. Ablation study on the Dress Code test set [16].

Results on Dress Code Dataset. To further validate our solution, in Table 1 we report experimental results on the Dress Code test set. Results are shown on the entire test set and on three subsets obtained by filtering the test image pairs of each of the three categories. In this experiment, we compare with CP-VTON [20], CP-VTON+ [15], and CIT [17] that we re-train from scratch of this dataset. For fair comparison, we also implement a variant of the CP-VTON and CIT models (*i.e.* CP-VTON† and CIT†) by using the masked person representation as input to the try-on module. In this way, also the information about the background and other model's body parts can be preserved in the generated images. Also in this case, our model surpasses other methods according to all evaluation metrics and all clothing categories.

To analyze the contribution of each component of the DBCT block, we also perform an ablation study. As it can be seen from Table 3, pre-processing the two-branch U-Net features with the cross-modal Transformer allows a reduction of 0.11 points in terms of FID. However, the major improvement is brought by the interaction of Transformer encoder and cross-modal attention which leads to a further reduction of the FID score of 0.40 points.

Qualitative Results. Finally, Figure 2 shows some qualitative results on VITON (left) and Dress Code (right). Overall, DBCT can better preserve details and shapes of the original try-on garments and the body poses of the reference models, thus further demonstrating the appropriateness of the proposed dual-branch collaborative module to increase the realism of generated images.

4. Conclusion

In this paper, we have presented DBCT, a Transformer-based architecture for virtual try-on that can jointly exploit cross-modal information of in-shop garments and reference models at generation time. Through experimental analyses on the VITON and Dress Code datasets, we have demonstrated the effectiveness of the proposed solution in comparison to both standard pure convolutional approaches and previous Transformer-based proposals for the task.

References

[1] Kumar Ayush, Surgan Jandial, Ayush Chopra, and Balaji Krishnamurthy. Powering virtual try-on via auxiliary human

segmentation learning. In *ICCV*, 2019. 1

[2] Kumar Ayush, Surgan Jandial, Ayush Chopra, and Balaji Krishnamurthy. Powering virtual try-on via auxiliary human segmentation learning. In *ICCV Workshops*, 2019. 1

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

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

[5] Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Transform, Warp, and Dress: A New Transformation-guided Model for Virtual Try-on. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2):1–24, 2022. 1

[6] 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

[7] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled Cycle Consistency for Highly-realistic Virtual Try-On. In *CVPR*, 2021. 3

[8] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-Free Virtual Try-on via Distilling Appearance Flows. In *CVPR*, 2021. 1

[9] 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, 3

[10] 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

[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. 1

[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, 3

[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] 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, 3, 4

[16] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. *arXiv preprint*, 2022. 2, 3, 4

[17] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *arXiv preprint arXiv:2104.05519*, 2021. 1, 2, 3, 4

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 1

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1

[20] 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, 3, 4

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 3

[22] 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