FIT-ME: IMAGE-BASED VIRTUAL TRY-ON WITH ARBITRARY POSES

Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Wen-Huang Cheng

National Chiao Tung University

ABSTRACT

The image-based virtual try-on system has raised research attention recently, but it still requires to upload an image of a user with the target pose. We present a novel learning model, Fit-Me network, to seamlessly fit in-shop clothing into a person image and simultaneously transform the pose of the person image to another given one. The proposed Fit-Me network helps users not only save the time used to change clothes physically but also provide comprehensive information about how suitable the clothes are. By facilitating the arbitrary pose transformation, we can generate consecutive poses to help users get more information for deciding whether to buy the clothes or not from different aspects.

Index Terms— Virtual try-on, pose transformation, image synthesis

1. INTRODUCTION

The revenue of e-commerce market has increased recently, where the market's largest segment is Fashion with a market volume of $598,631 million in 2019 and it is expected to present a market volume of US $835,781 million by 2023.[1] Although people get used to going online shopping for fashion items, the average conversion rate of online shopping platforms worldwide in 2018 is only 2.42%.[2] One of the major reasons is that people cannot try fashion items to find whether they are suitable or not when shopping online. Therefore, it is desirable for fashion industry to develop a virtual try-on system for improving the conversion rates and boosting sales.

To facilitate virtual try-on services, image-based virtual try-on [1, 2, 3, 4, 5, 6] and pose transformation [7, 8, 9, 10, 11, 12] are both popular nowadays. For example, Han et al. [3] presented an image-based virtual try-on network (VITON) to put a target clothing item onto the target person. Based on VITON, Wang et al. [4] further presented CP-VTON to preserve the characteristics of clothes by new warping and refining modules. For pose transformation, Balakrishnan et al. [7] separated end-to-end network into four segments to synthesize unseen pose via the modular generative neural network. Liqian Ma et al. [10] constructed a coarse-to-fine architecture to generate images of different posture for people. All the existing works only focus on either virtual try-on or pose transformation but do not consider to try-on and to transform different poses simultaneously.

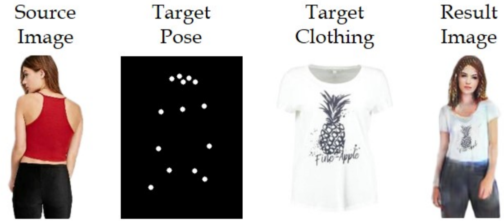


Fig. 1. Examples of virtual try-on with arbitrary poses

When customers go shopping for fashion items, it is natural for them to take arbitrary poses for determining whether it looks good on them from different aspects, e.g., the frontage, side views, or even the reverse side. Therefore, pose transformation plays a key role in improving image-based virtual try-on. However, it is tedious if the customers have to upload their photos with different poses. In this paper, we propose to synthesize the virtual try-on with arbitrary poses defined by joint coordinates which can be predefined according to frequent poses in the database. Take Fig. 1 as an example. Given a source image (depicting a user's details), a target pose (using joint coordinates), and a target image of in-shop clothing (specified by the user), the goal is to generate a result image with the user's details in the target pose wearing the target clothing.

Nevertheless, it is challenging to synthesize virtual try-on images with arbitrary poses. One basic approach is to use a two-phase approach, i.e., apply the existing pose transformation first and then image-based virtual try-on. However, the two-phase approach may suffer from the problem of training data sparsity and generate images with visual artifacts since the combinations of different clothes and poses are much greater than that of try-on only or pose transformation only.

To achieve the goal of virtual try-on with arbitrary poses, we propose a new network, namely, Fit-Me, comprised of four stages: (I) rough pose transformation, (II) clothing warping, (III) precise pose transformation, and (IV) dis-

[1] `https://www.statista.com/outlook/244/100/fashion/worldwide`

[2] `https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/`

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

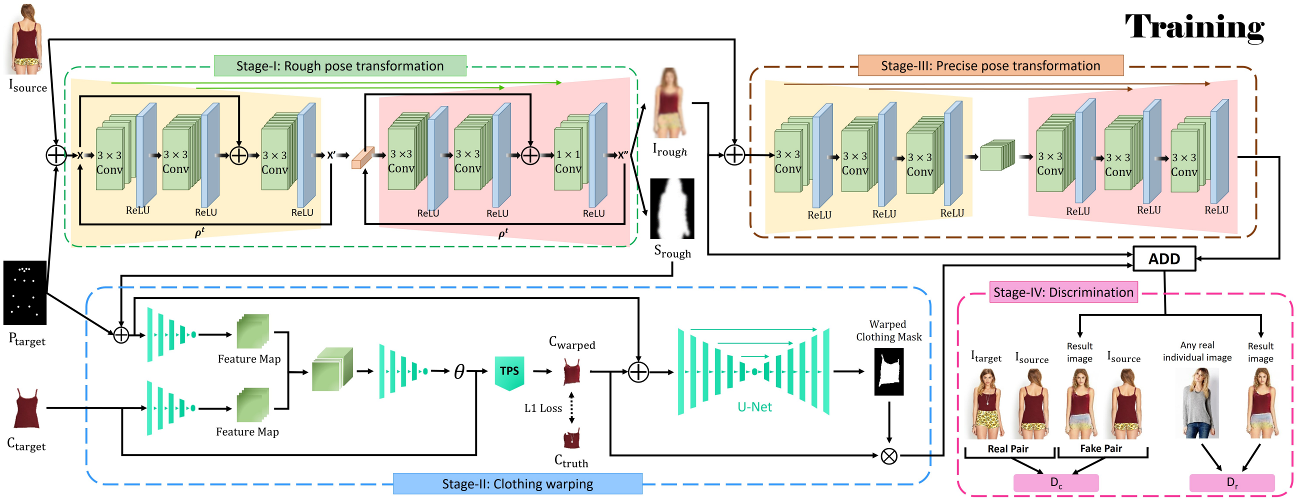


Fig. 2. **The training overview of our Fit-Me network**, which consists of four stages. Stage-I (*Rough pose transformation*) generates the rough image according to the target pose, where ρ^t represent t layers of convolutional neural layers. Stage-II (*Clothing warping*) warps the clothes to fit the target pose, stage-III (*Precise pose transformation*) generates detailed appearance, and stage-IV (*Discrimination*) learns to differentiate the real/fake images and image pairs to help stage-II and stage-III generate better results.

crimination. Specifically, Fit-Me generates the rough target pose transformation from the source image and target pose in stage-I. Afterward, we exploit thin plate spline to warp the clothing for fitting the target pose in stage-II, and then generates detailed appearance in stage-III. Lastly, Fit-Me adopts two discriminators to 1) select the most realistic image, for which a lot of training data exist and thus alleviate the data sparsity problem, and 2) select the real try-on image, which helps our model to precisely generate a detailed and refined result.

The contributions of this paper are summarized as follows. (1) We have identified an unprecedented task of virtual try-on with arbitrary poses, which can be a value-added service for fashion e-commerce platforms and improve the conversion rates. (2) A novel Fit-Me network is proposed to solve the complex task of virtual try-on with arbitrary poses, while two discriminators are utilized to alleviate the problem of training data sparsity. (3) Experimental results show that Fit-Me generates results with greater inception scores.

2. FIT-ME NETWORK

Given a source image I_{source} of a user, a target pose P_{target} defined by joint coordinates[3], and target in-shop clothing C_{target}, our goal is to generate a new image of the user virtually trying on target clothing and holding the same posture. A naïve way is to train a deep neural network with the dataset including triplet images: one in-shop clothing image, one image of a user wearing different clothing, and one image of a user wearing the in-shop clothing in different posture. However, such dataset does not exist. Therefore, our training dataset only contains the images of a person wearing the same clothing with different poses and in-shop clothing images.

2.1. Rough pose transformation

The first task of Fit-Me is to transform the pose of the person in the source image into the target pose. To achieve the transformed task, conditional GANs[15, 16] can be utilized to generate a realistic image. Since the task is challenging, in stage-I, we use a generative model G_r to generate a rough result I_{rough}. Specifically, we concatenate the $I_{source} \in \mathbb{R}^{w \times h \times 3}$ and the $P_{target} \in \mathbb{R}^{w \times h \times J}$ (J confidence maps for joints) as input for stage-I. Afterward, we construct G_r via an U-Net-like architecture [17] to: 1) encode I_{source} and P_{target} and transform the pose based on the encoded vector, and 2) highlight minor changes by sequential convolution-ReLU with local skip connections, which improve the generator performance [18]. Let I_{target} denote the groundtruth image of the user in the target pose, the loss for G_r is derived as follows.

$$\mathcal{L}_{G_r} = \sum_{i=1}^{n} \|G_r(I_{source}, P_{target}) - I_{target}\|_1 , \qquad (1)$$

[3] Pre-trained joint detector [13, 14] can be applied to estimate (x, y) joint coordinates from images. In our experiment, the following joints are used: nose, eyes, ears, neck, shoulders, elbows, wrists, hips, knees, and ankles.

where n is the number of image pairs with a user in two different poses. Note that L1 loss is adopted here since it measures the difference at low frequencies well. The details of clothing are refined in stage-II, while the edges of the body and the facial details are improved in stage-III.

2.2. Clothing warping

To warp the in-shop clothing for better fitting the target shape, inspired by [4, 19], we adjust their framework for stage-II of Fit-Me network. As illustrated in Fig.2, stage-II consists of: (1) two extraction modules: one extracts features from P_{target} and S_{rough} to retrieve the details of human body, and the other extracts the high-level information of C_{target}, (2) correlation matching module: combining two feature tensors to a single correlation tensor, (3) estimating network: predicting the parameter, θ, of spatial transformation from the correlation tensor, and (4) thin plate spline transformation $\mathcal{T}$: warping the target clothing C_{target} to the warped clothing C_{warped} with θ, i.e., $C_{warped} = \mathcal{T}(C_{target}, \theta)$. The loss is then computed as pixel-wise L1 loss between C_{warped} and C_{truth} as follows:

$$\mathcal{L}_{clothing}(C_{warped}, C_{truth}) = \|\mathcal{T}(C_{target}, \theta) - C_{truth}\|_1 \tag{2}$$

Moreover, we put the concatenated P_{target}, S_{rough} and C_{warped} into U-Net [20] for generating more detailed warped clothing mask to fit people more properly. The final clothing, denoted as $\hat{C}$, is obtained by compositing C_{warped} with warped clothing mask.

2.3. Precise pose transformation

To alleviate the blurry artifacts on I_{rough} derived from G_r, as well as integrate the results from the warped clothing C_{warped}, we construct a network G_p in stage-III to generate the absent information and to rectify the erroneous pixels in I_{rough}. We modify the network architecture of G_r for G_p by removing residual blocks to reduce the quantity of information and focus on preserving details. With I_{rough} and I_{source} as input, G_p generates a residual output $r = G_p(I_{source}, I_{rough})$. After processing images through stage-I to stage-III, we derive the final result by adding r, I_{rough} and $\hat{C}$ together.

2.4. Discrimination

In traditional GAN [21, 22, 23, 24, 25], the training loss is minimized by alternatively optimizing the discriminator and generator. However, it is challenging to synthesize virtual try-on images with arbitrary poses since it requires pose transformation and clothing warping at the same time. Therefore, we adopt a similar idea but construct two discriminators to complete the GAN architecture and fine-tune the final result from our network. Our conditional GAN consists of two generative networks and two discriminators to achieve a natural and convincing result image. Specifically, the first discriminator is denoted as D_c for examining the correctness of our result comparing to the groundtruth I_{target}. The first adversarial loss $\mathcal{L}_{correct}(G_p, D_c)$ is derived as follows.

$$\begin{aligned}\mathcal{L}_{correct}(G_p, D_c) = \mathbb{E}_{(I_{source}, I_{target})}[logD_c(I_{source}, I_{target})] \\ +\mathbb{E}_{(I_{source}, \hat{x})}[log(1 - D_c(I_{source}, G_p(\hat{x}) + I_{rough} + \hat{C})],\end{aligned} \tag{3}$$

where $\hat{x}$ represents I_{source} concatenating with I_{rough}.

However, when only using the pairwise discriminator D_c, D_c can differentiate the real and fake pairs by the correctness of clothing and pose warping, as well as the realness of images, which makes D_c too strong for the generator. Moreover, since real individual images are much more accessible, we construct the other auxiliary discriminator D_r to enhance the realness of our results via distinguishing our results with real individual images. Here, the second adversarial loss $\mathcal{L}_{real}(G_p, D_r)$ is derived as follows.

$$\begin{aligned}\mathcal{L}_{real}(G_p, D_r) = \mathbb{E}_{I_{real}}[logD_r(I_{real})] \\ + \mathbb{E}_{\hat{x}}[log(1 - D_r(G_p(\hat{x}) + I_{rough} + \hat{C})],\end{aligned} \tag{4}$$

where I_{real} represents any real individual image. Afterward, two discriminators of the proposed Fit-Me network are optimized with a hyperparameter α as follows.

$$\underset{G_p}{min}\ \underset{D_c, D_r}{max}\left(\frac{\alpha}{2} L_{correct}(G_p, D_c) + \frac{1-\alpha}{2} L_{real}(G_p, D_r)\right)$$

First, we set a larger α to encourage our discriminators focusing on D_c dwindling the difference between our result and I_{target} instead of naturalness. Then, we gradually decrease α to improve the significance of D_r to concentrate more on the realness of the result image.

3. EXPERIMENTS

In this section, we evaluate the Fit-Me network on two real datasets. We first present the details of the datasets and implementation. Afterward, quantitative and qualitative analysis are conducted with baselines. Finally, the limitation of Fit-Me are discussed for future research.

3.1. Experiment setup

To evaluate the proposed Fit-Me network, two real datasets are used in the experiment. The first dataset is DeepFashion[26] containing about 140K pairs of the same person with different poses and 50K in-shop clothing images which are required to train with our Fit-Me network. The second dataset is collected by VITON [3] containing about 16K frontal-view

Table 1. **Comparison on our virtual try-on dataset between different methods.**

Method	CP-VTON[4]	Pose-Guide[28]	Fit-Me (Ours)	Real Images
Mean	2.5733	3.0064	3.3360	3.6744
Variance	0.0050	0.0094	0.0162	0.0171

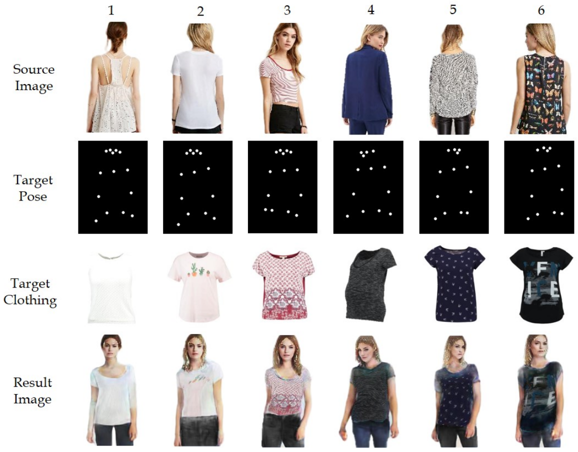
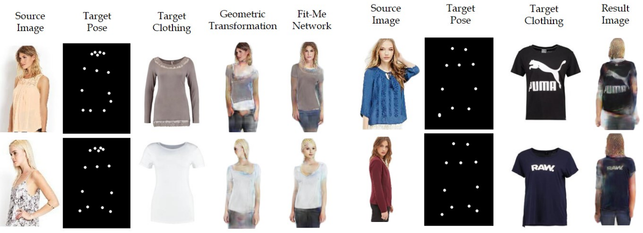


Fig. 3. **Experiment results of virtual try-on with arbitrary poses.** Fit-Me network precisely transforms the source image to the target pose with detailed target clothing.

woman and top clothing image pairs. All the images are cropped and resized to 256×192. We train our Fit-Me network with the DeepFashion dataset because two different poses of the same person and one in-shop clothing image are required, and use 5-fold cross-validation. Moreover, the second dataset (VITON) is used to test that our model has the ability to work well on the different dataset. We use the Adam [27] as the optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to be 0.00002 for 200K steps with a batch size of 8. We use PyTorch for the network implementation, where all convolutional layers are followed by a ReLU except the output convolutional layer.

3.2. Quantitative Analysis

Since none of the previous works can perform virtual try-on with arbitrary poses, we first compare Fit-Me with CP-VTON [4] and Pose-Guide [28] in terms of Inception Score (IS)[23] to evaluate the generated image quality. Table 1 shows that the proposed Fit-Me network outperforms other methods on two datasets because we extract human and clothing information separately and make them focus on their respective priorities to achieve comparable results. The IS value of Fit-Me is the closest to that of real images among all methods.



(a) Comparison (b) Failure cases

Fig. 4. (a) **Comparison against geometric transformation.** The three columns at left are inputs. The fourth column presents results with geometric transformation, and the rightmost column presents the results by Fit-Me network (b) **Failure cases of Fit-Me network**.

3.3. Qualitative Analysis

Fig. 3 shows the results of the Fit-Me network. Our network accurately transforms slight posture changes (column 3) as well as dramatic posture changes, e.g., turning from back to the front (columns 1, 2, and 5). Moreover, it is conspicuous that Fit-Me network not only succeeds in transforming poses but also preserves the detailed characteristics of clothing such as texture (lace in column 3) and patterns (camouflage, flecks and alphabets in columns 4, 5 and 6).

Additionally, we further replace G_r in Fit-Me network with spatial transformation [7]. We translate, rotate, and scale each limb pixel-to-pixel via geometric transformation. Because geometric transformation only considers the two-dimension transformation of body parts, it fails to warp the head naturally to the target position. As shown in the 4-th column of Fig. 4a, the result images show that geometric transformation network is able to change body poses with the target clothing successfully but fail to precisely fit the clothing to the target pose. Whereas our Fit-Me network preserves facial details, especially in the fifth column of Fig. 4a. Fig. 4b shows our failure cases due to a few training data of transforming the posture to the backside. Therefore, it always considers backside to be front when warping clothing.

4. CONCLUSION

In this work, we propose a new virtual try-on system (Fit-Me) that can generate virtual try-on images with arbitrary customer poses. By first transforming the pose of the user to a target pose according to the joints of the target pose, we warp the in-shop clothing to fit the target pose and preserve characteristics of both customers and clothing. Experimental results show that the proposed Fit-Me network outperforms the state-of-the-art methods both qualitatively and quantitatively. In the future, we plan to take the human segmentation information into consideration for generating better results.

5. ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-105-2628-E-009-008-MY3, MOST-107-2218-E-009-062, MOST-108-2634-F-007-009, MOST-107-2218-E-002-054, MOST-107-2221-E-182-025-MY2 and MOST-108-2634-F-009-006.

6. REFERENCES

[1] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu, "Swapnet: Image based garment transfer," in *ECCV*, 2018.

[2] Nikolay Jetchev and Urs Bergmann, "The conditional analogy gan: Swapping fashion articles on people images," *ICCV*, pp. 2287–2292, 2017.

[3] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis, "Viton: An image-based virtual try-on network," in *CVPR*, 2018.

[4] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin, "Toward characteristic-preserving image-based virtual try-on network," in *ECCV*, 2018, pp. 589–604.

[5] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai, "M2e-try on net: Fashion from model to everyone," *CoRR*, vol. abs/1811.08599, 2018.

[6] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler, "A generative model of people in clothing," in *ICCV*, 2017.

[7] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fredo Durand, and John Guttag, "Synthesizing images of humans in unseen poses," in *CVPR*, 2018.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros, "Everybody dance now," *CoRR*, vol. abs/1808.07371, 2018.

[9] Luan Tran, Xi Yin, and Xiaoming Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017.

[10] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *NIPS*, 2017.

[11] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng, "Multi-view image generation from a single-view," in *ACM*, 2018.

[12] Patrick Esser, Ekaterina Sutter, and Björn Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018.

[13] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[14] Zhe Cao, Tomas Simon, Shih-En Wei, , and Yaser Sheikh, "Realtime multiperson 2d pose estimation using part affity fields," in *CVPR*, 2017.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[16] 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.

[17] Tran Minh Quan, David Grant Colburn Hildebr, and Won-Ki Jeong, "Fusionnet: a deep fully residual convolutional neural network for image segmentation in connectomics," *CoRR*, vol. abs/1612.05360v2, 2016.

[18] Stanisaw Jastrzbski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio, "Residual connections encourage iterative inference," in *ICLR*, 2018.

[19] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[22] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," *NIPS*, 2015.

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," *NIPS*, 2016.

[24] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[25] Junbo Zhao, Michael Mathieu, and Yann LeCun, "Energy-based generative adversarial network," *ICLR*, 2017.

[26] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, June 2016.

[27] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *NIPS*, 2017, pp. 405–415.