Paper:

An Improved Two-stage Network for Video Virtual Try-on

Yongfeng Zhang*, Zhongjian Dai, Zhiyang Jia, Yaping Dai

School of Automation, Beijing Institute of Technology, Beijing 100081,China
Key laboratory of intelligent control and decision of complex systems
zyf1274006166@163.com

[Received 00/00/00; accepted 00/00/00]

In order to solve the problem that the picture-based virtual try-on model provides consumers with limited information, an improved two-stage video virtual try-on scheme based on deep neural networks is proposed. The network consists of two parts. One is a deep network used to learn Thin Plate Spline (TPS) deformation parameters of clothing, the other is a U-Net module used to generate the try-on effect. In the latter, the network extracts optical flow from the input data, performs DensePose pose annotations, trains on the improved U-Net, and outputs the frame-by-frame try-on effect. Compared with the existing work, the proposed solution obtains a higher quality video, which is embodied in that the facial details retained are more adequate and the edges of the clothing generated are better. The proposed approach is evaluated on VVT and VITON dataset. In the same test sample, the proposed network has a better visual effect than the current neural network-based try-on work.

Keywords: virtual try-on, deep neural network, image synthesis

1. Introduction

In the Internet age, a large number of consumers have moved from physical stores to online stores. However, due to the inability to try on online clothing, a large number of return and exchange issues have been caused. This wastes a lot of materials and time, which is harmful to consumers, stores and environment. In order to reduce the occurrence of the above-mentioned situations, virtual try-on has gradually become an important research direction in the field of image synthesis. Before neural networks were widely used in the field of virtual try-on, most studies used the strategy of constructing 3-dimension(3D) models [1, 2]. Although these technologies can perform realistic dress simulations, the high cost of installing hardware and collecting 3D annotation information hinders their deployment on the user side.

More and more researches in this field have used deep learning methods to generate try-on effects [3]. Existing works have achieved good results in the picture-based virtual try-on work. However, in actual use, the static pictures generated by these jobs provide very limited information for clothing purchase. At present, most of the research is focused on improving the effect of virtual try-on based on pictures [4, 5]. Until recently, some researches focused on try-on based on video [6]. In the videos generated by the existing work, there is still a lot of room for improvement in terms of clothing detail retention and clothing edge definition.

In this paper, a two-stage virtual try-on network based on video is proposed. The goal is to output a series of realistic continuous pictures for composite video under the premise of realizing a given character image, target clothing and a series of poses. The first stage is a network that obtains Thin Plate Spline (TPS) [7] parameters required for clothing deformation from the human body posture. The input standard clothing and human posture are extracted and merged, and then sent to the regression network for clothing transformation training. Since the datasets of different clothing are not easy to extract, the training method adopted at this stage is to reconstruct the original samples. The second stage takes four kind of input to generate continuous try-on images on the improved U-Net network. They are the deformed clothing obtained in the first stage, the optical flow information extracted from the frame-by-frame pictures, the Densepose information and optical flow. The experiment is tested under the video dressing dataset VVT. Compared with the typical virtual try-on model based on image CP-VTON, the video try-on effect synthesized by the model in this paper is significantly improved. Compared with the typical video virtual try-on model FW-GAN, the video synthesized by the model in this paper achieves better results in terms of portrait preservation and clothing edge definition.

The paper is structured as follows. In Section 2, we give an overview over related to the work presented in this paper. In Section 2, some work related to this paper will be introduced. Section 3 introduces the proposed method of deep neural network for virtual try-on in detail. In Section 4, the presented approaches are evaluated and compared to related methods experimentally. We conclude a discussion and an outlook to future work in Section 5.

2. Related work

2.1. Image synthesis

Generative adversarial networks (GANs) [8] had recently achieved impressive results on image synthesis. There are a few recent studies investigating the problem of image-to-image translation using conditional GANs [9], which transform a given input image to another one with a different representation [10, 11]. Most of image-to-image translation tasks conditioned on unaligned images [12], adopted a coarse-to-fine manner to enhance the quality of final results.

For person image generation, Lassner et al. proposed a generative model of people in clothing for the full body [13], but the fashion attributes are not controllable in this method. Zhao et al. proposed an image generation model to generate multi-view cloth images from only a single view input [14, 15].

2.2. Virtual try-on

Virtual try-on has been an attractive topic even before the renaissance of deep learning. Most previous works were based on computer graphics. Guan et al. designed a framework for synthesizing clothes on 3D bodies, with ignoring the shape and pose [16]. Pons-Moll solved the problem of capturing multiple clothes of a neatly dressed person in motion by using a multi-part 3D model of the clothes bodies [2].

There are also a few works based on image-based generative models which aim to synthesize perceptually correct images from real 2D images. Jetchev introduced a conditional analogy GAN to swap fashion items [17]. Methods such as VITON [3], CP-VTON [4] use coarse human shape and pose map as the input to generate a clothed person. While methods such as VTNFP [5] and ACGPN [15] adopt semantic segmentation as input to synthesize clothed person.

Not satisfied with the limited information provided by the picture try-on, some work started on the video-based try-on. Dong et al. proposed a strategy to use the previous frame information to assist the current frame's try-on image, and provided a VVT dataset for video try-on [6].

3. Improved two-stage network

The network proposed in this paper is divided into two parts and the two parts are trained separately. The first part generates an image that distorts the target clothing into an image that approximately fits the current person's posture. The second part generates masks for each part from the clothing, posture information and optical flow data, and outputs the try-on effect of the current frame. This section will introduce these two parts in detail.

3.1. Clothes warping module

The input of this part is the target clothing c, the current human body posture information p, and the clothing part in the original image c_t. The goal is to train a network that can generate clothing pictures matched the human posture. The most ideal training data requires the same person to maintain the exact same posture under multiple sets of different clothing, which is very difficult for a single set of samples, let alone a huge number of datasets. Therefore, we use the clothing currently worn as the target clothing, and train in the form of reconstructing the original image to reduce the difficulty of dataset construction. This kind of method is a common for some image synthesis training in previous works.

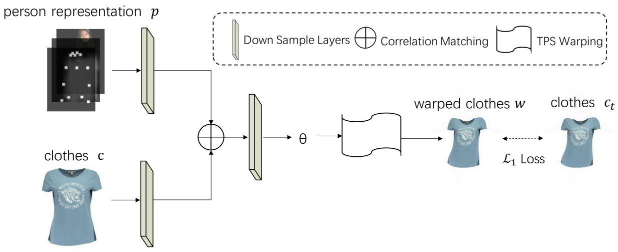

Fig. 1. Clothes warping module. The target clothes c and person representation p are aligned via a learnable matching module.

We extract high-level features from the human body pose and clothing through a sampler, and then combine them into a single tensor related layer as the input of the regression network. The predicted spatial transformation parameters are obtained through the regression network, and the clothing picture c is processed with the parameters to obtain w. This part of network is shown in Fig 1. Use w and the clothing part in the original image c_t to calculate the L_1 loss as the equation 1.

$$L_1(\theta) = \|w - c_t\|_1 = \|T_\theta(c) - c_t\|_1 \tag{1}$$

where θ is the TPS parameter learned by the network, w is the output of the distorted network under the current parameters. w is the output of TPS warping net while input parameter is θ and input clothes is c.

3.2. Try-on module

The try-on results produced by previous work often have texture artifacts and irrelevant area information lost during the generation process. The U-Net architecture is very effective in the synthesis of cloth and human body. In order to obtain a better try-on effect, we further preprocess the input Mask sent to U-Net to generate images.

The network structure of this part is shown in Fig 2. The main body of the network is a U-Net network. In order to learn features better, we use a self-attention layer in the deep feature layer. We will use four parts as input to this network. The inputs include the current frame, the distorted clothing generated by the first stage network, the DensePose posture information, and the combination of optical flow information.

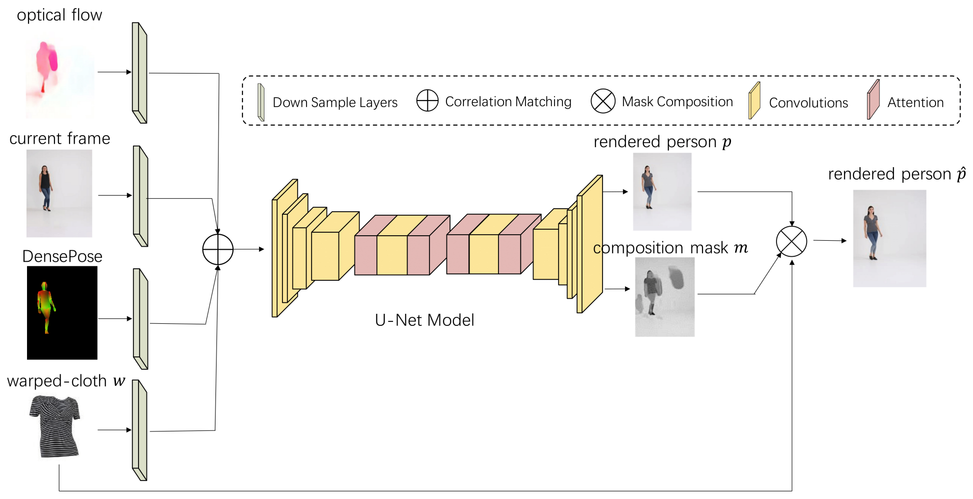

Fig. 2. Try-on module. Densepose and optical flow information are obtained by preprocessing the used VVT dataset . Person representation and warped cloth w are fed into the U-Net model with self-attention. The output of the U-Net is the person p in target cloth .

The U-Net model outputs a rendered person p, and composition mask m. The person try-on, $\hat{p}$ is obtained as equation 2.

$$\hat{p} = w * m + p * (1 - m) \tag{2}$$

where w is the warped cloth obtained from the first stage.

4. Experiments

In this section, we introduce the training details of the network, and construct a suitable evaluation method to analyze the experimental results.

4.1. Implementation details

In the training of the first part, the loss function attenuation coefficient=1, the batch size is set to 4, the number of training times is 20K, and the Adam optimizer is used in the training process. The learning rate is set to 0.0001 in the initial test, and linearly decays to 0 after half of the training times.

In the second part of training the U-Net model, the batch size is set to 32, the number of training is 1K, and the Adam optimizer is also used. The design of the learning rate is the same as the previous part.

4.2. Dataset

The two parts of the network training use two different datasets. Currently, there are two datasets commonly used in the field of try-on, namely VITON and VVT. The former includes more than 16,000 sets of samples, each of which includes a picture of clothing and a picture of a model trying on. The latter includes 791 samples. Each set of samples includes a clothing picture and a series of consecutive frames that the model tries on against a white background. The number of frames for each sample is between 200 and 300.

In the experiment, the training goal of the wrapping-net model distorted clothing is to generate clothing that fits the current posture as much as possible, and the demand for dynamic information of the characters is relatively low, so VITON with more samples is used as the dataset. In the try-on network, we import optical flow and the previous frame to generate better continuous pictures. For the need of time-relevant information, we use VVT as dataset in this stage.

4.3. Poseture expression and activation function

The human pose representation used in the experiment is different from the CocoPose annotation commonly used in previous work. A UV coordinate map obtained by DensePose includes more 3D human body information. The difference of them is shown in Fig 3.

Comparing the training effects under the two kinds of annotations, the results are shown in third and fourth columns as shown in Fig 4. It can be seen that although the processing of the half-length image of CocoPose annotation is satisfactory, it is still inferior to another one in details and the effect of full-body processing.

In the course of the experiment, it was found that using GELU as the activation function of the second stage network can obtain better results than others such as ReLu and Swish. The comparison results are shown in the fifth to seventh columns of Fig 4.

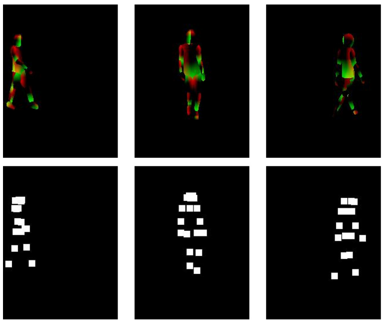

Fig. 3. Visual comparison between DensePose and CocoPose. DensePose annotations (first row) contain dense 3D body information in the form of UV coordinates, where CocoPose annotations (second row) are sparse and 2D keypoints.

4.4. Experimental result

After the above attempts, the network obtains the best output result when Densepose is used as the pose information and GELU is used as the activation function. The try-on problem is the same as many image generation problems that it is difficult to evaluate the results through several specific numerical indicators. Here we will qualitatively evaluate the model by analyzing the test pictures. Comparing the results with CP-VTON and FW-GAN, they are currently the two best jobs in the field of virtual try-on. In order to better present the try-on effect, we take a picture of the walking process and a picture of the closest position of the model to the picture as a set for comparison, as shown in Fig 5.

The try-on effect of CP-VTON is satisfactory when the upper body is big enough, especially the preservation of the clothing pattern, but the generation effect during the walking process is much weaker, and the continuity of the adjacent two frames of image also poor. The model proposed in this paper can reach a level equivalent to CP-VTON when it is closest to the screen, and the generated image will not have unexpected highlights in the distorted position of the clothing. FW-GAN introduces the optical flow method to obtain more realistic and smooth video. In the first row of each sample, our model and FW-GAN can guarantee the natural lifelikeness of the generated video. Compared with FW-GAN, our model can obtain dynamic continuous frame images of comparable quality, and the preservation of clothing details when closest to the lens is better than the latter. Our improvement to the try-on network can take into account the smoothness of the output image and the undistorted clothing pattern.

5. Conclusion

We propose an improved two-stage video virtual try-on network. We have imported Densepose, a better human pose representation method, and selected GELU as the activation function through comparative experiments. In the experiment, the test results have been qualitatively analyzed with two excellent models CP-VTON and FW-GAN in the virtual try-on field. Among them, DensePose and the self-attenton layer improve the face and body details of the generated pictures. The addition of optical flow information increases the continuity between images, but it is found that the output images have undesired artifacts in some areas of the result.

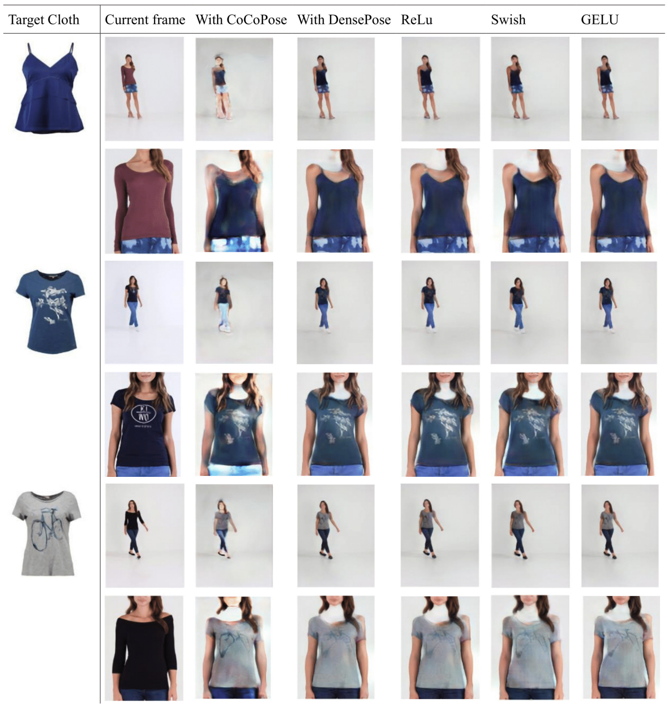

Fig. 4. Comparison between different annotations and activation function. The first column is the target clothing, and the second column is the current frame of the input. The third and fourth columns are the comparison between CoCoPose and Densepose. The fifth to seventh columns are comparisons of try-on results under three different activation functions.

At present, due to performance and hardware limitations in the video-based virtual try-on, it is still very difficult to achieve the detail retention capability of the image-based virtual try-on. How to find a balance between dynamic try-on and high precision to meet the needs of actual use as much as possible is still a problem that needs to be explored. Another problem is the model cannot process the geometric information of the clothing (such as distinguishing the inside and outside of the clothing), which will result in the inability to perform better try-on simulations for certain types of clothing. This may also be a direction of our further research.

References:

[1] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Bjorn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *Proceedings of the 5th International Conference on 3D Body Scanning Technologies, Lugano, Switzerland, 21-22 October 2014*, Ascona, Switzerland, 2014. Hometrica Consulting - Dr.

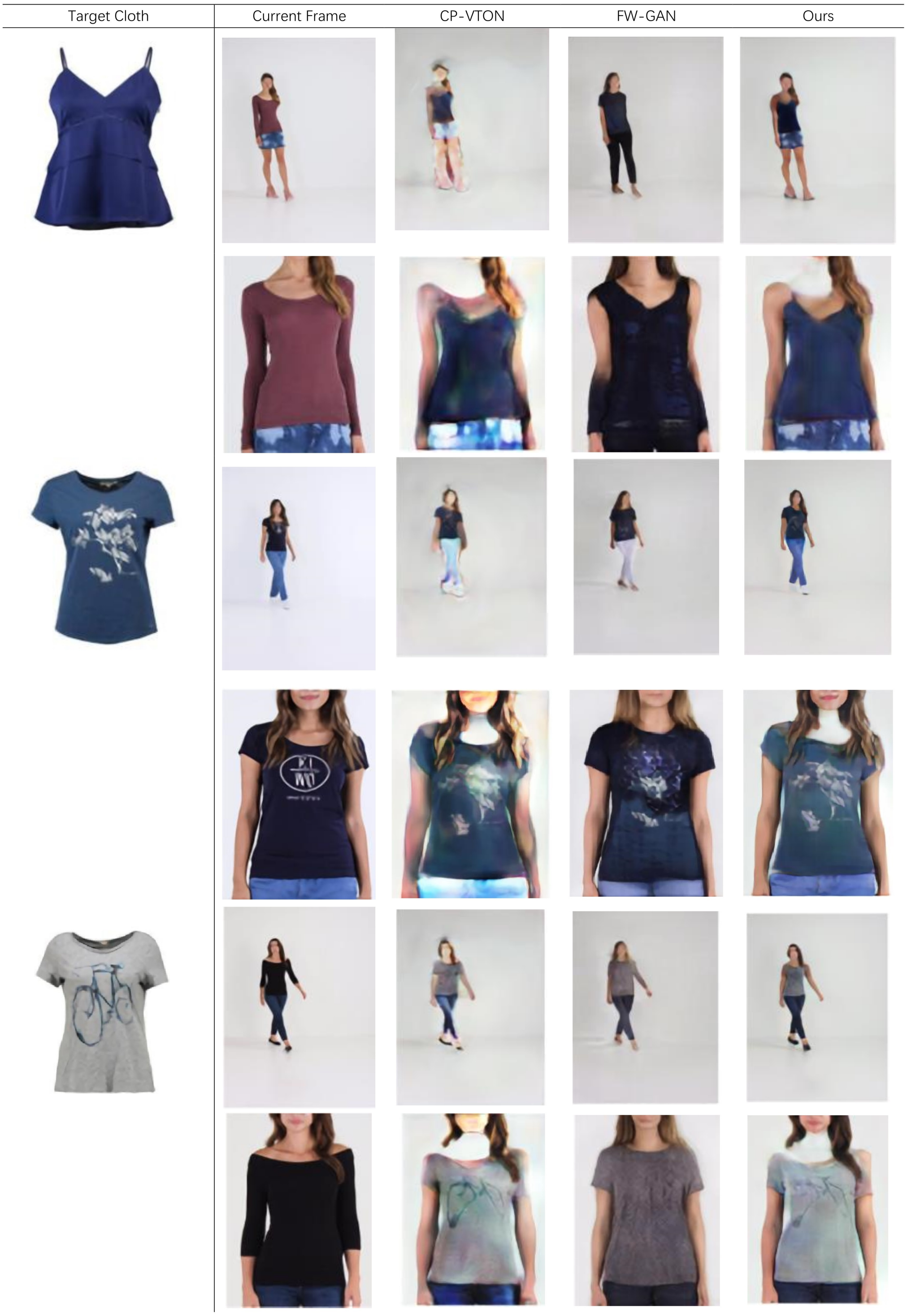

Fig. 5. Try-on results comparing FW-GAN, CP-VTON and ours

Nicola D'Apuzzo.

[2] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Trans. Graph.*, 36(4):1–15, 2017.

[3] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An image-based virtual try-on network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

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

[5] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An image-based virtual try-on network with body and clothing feature preservation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[6] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. FW-GAN: Flow-navigated warping GAN for video virtual try-on. In *2019 IEEE/CVF International*

Conference on Computer Vision (ICCV). IEEE, 2019.

[7] F L Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989.

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014.

[9] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. 2014.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[12] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[13] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[14] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-View image generation from a Single-View. In *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*, New York, New York, USA, 2018. ACM Press.

[15] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[16] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. DRAPE: DRessing any PErson. *ACM Trans. Graph.*, 31(4):1–10, 2012.

[17] Nikolay Jetchev and Urs Bergmann. The conditional analogy GAN: Swapping fashion articles on people images. 2017.