Virtual Clothing Try-on Using Generative Adversarial Networks

Rahul Singh[1], Aman Bindal[2], Md Azad Khan[3], Rakshith MR[4], Prof. Divakara N[5]

[1] *BE Student, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[2] *BE Student, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[3] *BE Student, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[4] *BE Student, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[5] *Assistant Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*

ABSTRACT

Virtual clothing try-on aims at transferring a target clothing image onto a reference person, and has become a major topic in recent years. The traditional try-on task aims to align the target clothing item naturally to the given person's body and hence present a try-on look of the person. However, in practice, people may also be interested in their try-on looks with different poses. Therefore, in this work, we introduce a new try-on setting, which enables the changes of both the clothing item and the person's pose. Towards this end, we propose a pose guided virtual try-on scheme based on the generative adversarial networks (GANs). We first predict semantic layout of the reference image that will be changed after try-on, and then determines whether its image content needs to be generated or preserved according to the predicted semantic layout, leading to photo realistic try-on and rich clothing details. In particular this involves three modules. First, a semantic layout generation module utilizes semantic segmentation of the reference image to progressively predict the desired semantic layout after try-on. Second, a clothes warping module warps clothing images according to the generated semantic layout, where a second-order difference constraint is introduced to stabilize the warping process during training. Third, a content fusion module that integrates all information (e.g. reference image, semantic layout, warped clothes) to adaptively produce each semantic part of human body. Experiments on our newly collected dataset demonstrate its promise in the image-based virtual try-on task over state of the art generative models.

Keywords: - *Virtual Try-on; Generative Adversarial Networks, Semantic Layout Generation, Clothes warping, Content Fusion*

1. INTRODUCTION

Recent years have witnessed the increasing demands of online shopping for fashion items with many E commerce websites coming up. The huge economic value of online fashion market demonstrates people's great demand of online fashion shopping. Nevertheless, lacking the physical try-on, online fashion shopping is always criticized for its poor user experience and people still prefer trying on their clothes before purchasing as they would like to see how it looks on them rather than a model. Due to the COVID-19 situation, going to a store and trying on clothes has become an even greater problem and thus transferring of clothing onto a reference person or the customer has received much attention. This, allows consumers to virtually try on clothes, which will not only enhance their shopping experience, transforming the way people shop for clothes, but also save cost for retailers. Owing to the recent advances in computer graphics, there emerge several practical services, such as TriMirror and Fits Me, which work on synthesizing the try-on looks for users based on their 3D body shape measurements, desired poses and target clothing items. Despite that 3D-based methods have achieved promising success, the huge labor costs for 3D data annotation and potential economic costs for scanning equipment largely limit their real-world applications.

Recently virtual try-on methods based solely on RGB images have also been proposed. Conditional Generative Adversarial Networks (GANs), which have demonstrated impressive results on image generation, image-to-image translation and editing tasks, seem to be a natural approach for addressing this problem. In particular, they minimize an adversarial loss so that samples generated from a generator are indistinguishable from real ones as determined by a discriminator, conditioned on an input signal. However, they can only transform information like object classes and attributes roughly, but are unable to generate graphic details and accommodate geometric changes. This limits their ability in tasks like virtual try-on, where visual details and realistic deformations of the target clothing item are required in generated samples. Although several pioneer researches have achieved promising performance, most of existing efforts usually focus on preserving the character of a clothing image (e.g. texture, logo,

embroidery) when warping it to arbitrary human pose and fail to preserve the fine details, such as the clothing of the lower-body and the hair of the person losing the details and style. They also only generate the single-view try-on result, i.e., keeping the person's pose unchanged while simply changing the clothing item. It remains a big challenge to generate photo-realistic try-on images when large occlusions and human poses are presented in the reference person.

To address the above limitations, we propose a virtual try-on network, which first predicts the semantic layout of the reference image and then adaptively determines the content generation or preservation according to the predicted semantic layout. Specially, the network consists of three major modules as shown in Fig. 1. The first one is the Semantic Generation Module (SGM), which uses the semantic segmentation of body parts and clothes to progressively generate the mask of exposed body parts (i.e. synthesized body part mask) and the mask of warped clothing regions. SGM generates semantic masks in a two-stage fashion to generate the body parts first and synthesize clothing mask progressively, which makes the original clothes shape in reference image completely agnostic to the network. The second part is the Clothes Warping Module (CWM), which is designed to warp clothes according to the generated semantic layout. A second-order difference constraint is also introduced to the Warping loss to make the warping process more stable, especially for the clothes with the complex texture. Finally, the Content Fusion Module (CFM) integrates the information from the synthesized body part mask, the warped clothing image, and the original body part image to adaptively determine the generation or preservation of the distinct human parts in the synthesized image.

The main contributions can be summarized as follows:

(1) We propose a new image-based virtual try-on network, which greatly improves the try-on quality in semantic alignment, character retention and layout adaptation.

(2) We take the semantic layout into consideration to generate photo-realistic try-on results.

(3) A second-order difference constraint makes the training process of warping module more stable, and improves the ability of our method to handle complex textures on clothes.

(4) Experiments demonstrate that the proposed method can generate photorealistic images that outperform the state-of-the-art methods.

2. LITERATURE SURVEY

Generative Adversarial Networks. The name "GAN" was introduced by Ian Goodfellow et al [1] in 2014. Their paper popularized the concept and influenced subsequent work. It introduced a new framework for estimating generative models via an adversarial process, in which they simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. There was no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrated the potential of the framework through qualitative and quantitative evaluation of the generated samples.

Radford et al [4] proposed the most famous GAN, the DCGAN, which adopted the use of a Convolutional Neural Network (CNN), a model predominantly used in supervised learning, in an unsupervised learning scenario. This model performs well in image synthesis in the early work of the research. In order to control generated result, different GAN such as CGAN, InfoGAN, and ACGAN are proposed. Some methods have been proposed for solving the model collapse problem by designing a new loss function such as mini-patch feature and WGAN.

Fashion Analysis. Fashion related tasks recently have received considerable attention due to their great potential in real-world applications. Most of the existing works focus on clothing compatibility and matching learning [9], clothing landmark detection [8], and fashion image analysis [10, 11]. Virtual try-on is among the most challenging tasks in fashion analysis.

Image Synthesis. In the image synthesis domain, GANs have achieved compelling success in various tasks, ranging from the general image generation [6, 7], to the pose-guided person synthesis [12, 13]. In particular, one family of its derivatives, conditional GANs, have been extensively studied recently, especially in image-to-image translation tasks, like the style transfer and virtual try-on, where certain conditional images should be given. In the context of image synthesis for fashion applications, Yoo et al [14] generated a clothed person conditioned on a product image and vice versa regardless of the person's pose. A more related work is FashionGAN [15], which replaced a fashion item on a person with a new one specified by text descriptions. However, all those works fail to preserve the texture details consistency corresponding with the pose.

Virtual Try-on. Existing deep learning based methods on virtual try-on can be classified as 3D model based approaches and 2D image based ones. In contrast to relying on 3D measurements to perform precise clothes simulation, in our work, we focus on synthesizing a perceptually correct photo-realistic image directly from 2D images, which is more computationally efficient. In computer vision, limited work has explored the task of virtual try-on. Jetchev et al. [16] presented a conditional analogy GAN (CAGAN), which casts the try-on task as an image analogy problem. As a matter of fact, CAGAN overlooks the clothing item deformation according to the user's body shape, and hence suffers from the unsatisfactory try-on performance. To address this issue, several efforts have been dedicated to synthesizing the virtual try-on images with the geometric alignment, such as VITON [2] and CP-VTON [3]. VITON exploits a Thin-Plate Spline (TPS) [17] based warping method to first deform the in shop clothes and map the texture to the refined result with a composition mask. CP-VTON adopts a similar structure of VITON but uses a neural network to learn the transformation parameters of TPS warping rather than using image descriptors, and achieves more

accurate alignment results. CP-VTON and VITON only focus on the clothes, leading to coarse and blurry bottom clothes and posture details. VTNFP [5] alleviates this issue by simply concatenating the high-level features extracted from body parts and bottom clothes, thereby generating better results than CP-VTON and VITON. However, blurry body parts and artifacts still remain abundantly in the results because it ignores the semantic layout of the reference image.

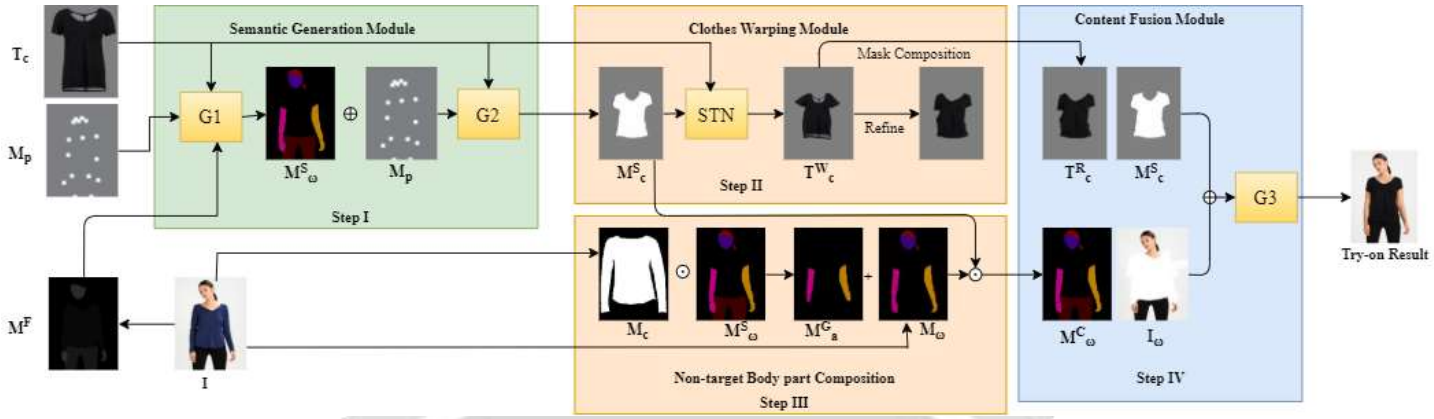

Figure 1: The overall architecture of our network. G1, G2 and G3 stand for conditional GANs. (1) In step I, the SGM outputs synthesized body part mask M^S_ω and target clothing mask M^S_c; (2) In step II, the CWM warps the target clothing image to T^R_c; (3) In step III and IV, the CFM gives us the final try-on image.

3. PROPOSED METHOD

The proposed network consists of three modules as shown in Fig. 1. First, the Semantic Generation Module (SGM) progressively generates the mask of the body parts and the mask of the warped clothing regions via semantic segmentation, yielding semantic alignment of the spatial layout. Second, the Clothes Warping Module (CWM) is designed to warp target clothing image according to the warped clothing mask, where we introduce a second-order difference constraint on Thin-Plate Spline (TPS) to produce geometric matching yet character retentive clothing images. Finally, the Content Fusion Module (CFM), integrates the information from previous modules to adaptively determine the generation or preservation of the distinct human parts in the output synthesized image.

3.1 Semantic Generation module (SGM)

Most previous works in virtual try-on focus more on the target clothes and do not consider the human body generation. They only use the coarse body shape directly in the network, which leads o loss of the details in the human body parts. To overcome this, the mask generation mechanism that is adopted in this module generates semantic segmentation of body parts and target clothing region precisely. The semantic generation module (SGM) is proposed to separate the target clothing region as well as to preserve the body parts of the person, without changing the pose and the rest of the details of the human body.

In the SGM, given a reference image I, and its corresponding mask M, we take the fused map M^F shown in Fig. 1 as one of the inputs to SGM. This module consists of two steps, the try-on mask generation module first synthesizes the masks of the body parts M^S_ω ($\omega = \{h, a, b\}$ (h: head, a: arms, b: bottom clothes)), which helps to adaptively preserve body parts instead of coarse feature in the subsequent steps. As shown in Fig. 1, we train a body parsing conditional GAN G_1 to generate M^S_ω by using the information from the fused map M^F, the pose map M_p, and the target clothing image T_c. In the second step, M^S_ω, M_p and T_c are combined to generate the synthesized mask of the clothes M^S_c by conditional GAN G_2. For training this module, both steps use the conditional generative adversarial network (CGAN), in which a U-Net structure is used as the generator while a discriminator given in pix2pixHD [18] is deployed to distinguish generated masks from their ground-truth masks. For each of the stages, the CGAN loss can be formulated as

$$\begin{aligned}\mathcal{L}_1 =& \mathbb{E}_{x,y}\left[\log\left(\mathcal{D}\left(x,y\right)\right)\right] \\ &+ \mathbb{E}_{x,z}\left[\log\left(1-\mathcal{D}\left(x,\mathcal{G}\left(x,z\right)\right)\right)\right],\end{aligned}$$

where x indicates the input and y is the ground-truth mask. z is the noise which is an additional channel of input sampled from standard normal distribution.

The SGM can serve as a core component for accurate understanding of body-parts and clothes layouts in virtual try-on and preserving of image content by composition. SGM is also effective for other tasks that need to partition semantic layout.

3.2 Clothes Warping Module (CWM)

Clothes warping aims to fit the clothes into the shape of target clothing region with visually natural deformation according to human pose as well as to retain the character of the clothes. We use a second-order difference constraint on the clothes warping network to realize geometric matching and character retention. As shown in Fig. 2, compared to the result with our proposed constraint, target clothes transformation without the constraint shows obvious distortion on shape and unreasonable mess on texture.

Specifically, given T_c and M^S_c as the input, we train the Spatial Transformation Network (STN) to learn the mapping between them. The warped clothing image T^W_c is transformed by the learned parameters from STN, where we introduce the following constraint L_3 as a loss term,

$$\mathcal{L}_3 = \sum_{p \in \mathbf{P}} \lambda_r \left| \|pp_0\|_2 - \|pp_1\|_2 \right| + \left| \|pp_2\|_2 - \|pp_3\|_2 \right| + \lambda_s \left(\left| S(p, p_0) - S(p, p_1) \right| + \left| S(p, p_2) - S(p, p_3) \right| \right),$$

where λ_r and λ_s are the trade-off hyper-parameters. As illustrated in Fig. 2, p(x, y) represents a certain sampled control point and $p_0(x_0, y_0)$, $p_1(x_1, y_1)$, $p_2(x_2, y_2)$, $p_3(x_3, y_3)$ are the top, bottom, left, right sampled control points of p(x, y), respectively in the whole control points set P; S (p, p_i) is the slope between two points. The warping loss can be represented as L_w, which measures the loss between the warped clothing image T^W_c and its ground-truth I_c,

$$\mathcal{L}_w = \mathcal{L}_3 + \mathcal{L}_4,$$

where $L_4 = \| T^W_c - I_c \|_1$. The warped clothes are then fed into a refinement network, where a learned matrix α ($0 \le \alpha_{ij} \le 1$) is then utilized to finally combine the two clothing images as the refined clothing image T^R_c by

$$\mathcal{T}^R_c = (1 - \alpha) \odot \mathcal{T}^W_c + \alpha \odot \mathcal{T}^R_c,$$

where $\odot$ denotes element-wise multiplication. Thus, the refined clothing image can fully retain the character of the target clothes.

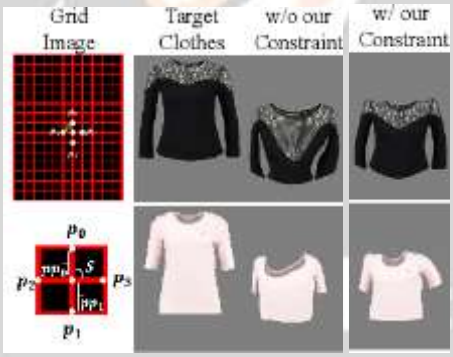

Figure 2: Comparison of warping results with and without the second-order difference constraint.

3.3 Content Fusion Module (CFM)

Most existing works usually adopt the coarse body shape as a cue to generate the final try-on images, and fail to reconstruct fine details where the target clothing region is required to clearly rendered, and body parts are needed to be adaptively preserved. Whereas, the content fusion module (CFM) is composed of two main steps, i.e. Steps 3 and 4 in Fig. 1. Step 3 fully maintains the untargeted body parts as well as preserves the changeable body part (i.e. arms). Step 4 fills in the changeable body part by utilizing the masks and images generated from previous steps accordingly by an in painting based fusion GAN G_3 in Fig. 1.

Non-target Body Part Composition. The composited body mask M^C_ω is composed by original body part mask M_ω, the generated body mask M^G_a which is the region for generation, and synthesized clothing mask M^S_c according to

$$\mathcal{M}^G_a = \mathcal{M}^S_\omega \odot \mathcal{M}_c,$$
$$\mathcal{M}^C_\omega = (\mathcal{M}^G_a + \mathcal{M}_\omega) \odot (1 - \mathcal{M}^S_c),$$
$$\mathcal{I}_\omega = \mathcal{I}_{\omega'} \odot (1 - \mathcal{M}^S_c),$$

where $\odot$ denotes element-wise multiplication. $I_{\omega'}$ is the original image I subtracting clothing region M_c. M^C_ω preserves the non-target body part by combining the two masks (i.e. M^S_ω and M_ω), which are used to recover the non-targeted details in the

following step to preserve I_ω and generate coherent body parts with the guidance of M^G_a . It can also deal with different cases. For example, when transferring a T-shirt to a person in long-sleeve only the within region of M^G_a will perform generation and preserve all the others, while in the opposite case, remaining body parts will be shaded by clothes.

Mask Inpainting. CFM removes part of the arms in the body images I_ω, making it possible to separate the regions of preservation and generation. To combine the semantic information, composited body mask M^C_ω and synthesized clothing mask M^S_c are concatenated with the body part image I_ω and refined clothing image T^R_c as the input. Therefore, in the inference stage, the network can adaptively generate the photo-realistic try-on image with rich details via the proposed CFM.

4. EXPERIMENTS

4.1 Dataset

Experiments are conducted on the dataset that used in VITON [2] and CP-VTON [3]. It contains about 19,000 image pairs, each of which includes a front-view woman image and a top clothing image. After removing the invalid image pairs, it yields 16,253 pairs, further splitting into a training set of 14,221 pairs and a testing set of 2,032 pairs.

4.2 Implementation Details

Architecture. The network contains three modules SGM, CWM and CFM. All the generators in SGM and CFM have the same structure of U-Net and all the discriminators are from pix2pixHD [18]. The structure of STN in CWM begins with five convolution layers followed by a max-pooling layer with stride 2. Resolution for all images in training and testing is 256 × 192. The architecture is shown in Fig. 1, we first predict the semantic layout of the reference image, and then decide the generation and preservation of image content.

Training. We train the three modules separately and combine them to eventually output the try-on image. Target clothes used in the training process are the same as in the reference image since it is intractable to have the ground-truth images of try-on results. Each module in the proposed method is trained for 20 epochs by setting the weights of losses $\lambda r = \lambda s = 0.1$, $\lambda 1 = \lambda 2 = 1$. The learning rate is initialized as 0.0002 and the network is optimized by Adam optimizer with the hyper-parameter $\beta 1 = 0.5$, and $\beta 2 = 0.999$. All the networks are implemented using the deep learning toolkit PyTorch and NVIDIA GeForce RTX 2080 GPU is used in our experiments.

Testing. The testing process follows the same procedure of training but is only different with that the target clothes are different from the ones in the reference images.

Figure 3: Examples of images generated by our network. The images in order, correspond to the body part mask, synthesized clothing mask, target clothing, try-on image generated by the network and the reference person image.

5. TESTING AND ANALYSIS

The network after training is able to generate photo-realistic try-on images when a target clothing and reference person are given as input. A few samples of the images generated along with the body part mask M^{S}_{ω} and synthesized clothing mask M^{S}_{c} can be seen in the following Fig. 3. Our network performs much better in simultaneously preserving the character of clothes and the body part information. Benefited from the proposed second-order spatial transformation constraint in CWM, it prevents Logo distortion and realizes character retention, making the warping process to be more stable to preserve texture and embroideries. As shown in the first example of the second row in Fig. 3, Logos on the clothes are clear and undistorted. Fig. 4 shows some more results of the try-on networks. It can be seen here that our network performs robustly with various poses including occlusions and cross-arms and for long-sleeve clothes to short-sleeve reference image and short-sleeve clothes to long-sleeve reference image, which demonstrates the generality of our method.

Figure 4: Extensive try-on results with different types of clothes and different reference people with varying poses.

6. CONCLUSION AND FUTURE WORK

In this work, we propose a network, which aims at generating photo-realistic try-on result while preserving both the character of clothes and details of human identity (posture, body parts and bottom clothes). We present three carefully designed modules, i.e. Semantic Generation Module (SGM), Clothes Warping Module (CWM), and Content Fusion Module (CFM). We evaluate our network on the VITON [2] dataset. The results clearly show the great superiority of our network.

In our future work, we look to train the model on other dataset as well, such as for men's clothing. We also look to improve on the image generation for different poses and views of the reference person (side view, back view). Another factor that could be considered in future work and that will be helpful for consumers would be to take the size of the clothing and reference person into consideration.

We conclude by noting the robustness of the implemented model and the high quality of the images generated. This model has potential to be employed in generating photo-realistic try-on images in online shopping and other websites. The model has a long way to go before it can generate satisfactory results for all kinds of clothing and reference poses. But the field of machine learning is doing very well with the ever decreasing cost of computational power and ever increasing brain power invested in its research.

7. ACKNOWLEDGEMENT

Firstly, we would like to express our sincere gratitude to Dr. M. P. Pushpalatha, HOD of Dept. of Computer Science and Engineering, JSS STU for providing an excellent environment for our education and encouraging us throughout our stay in college. We extend our heartfelt gratitude to our guide Prof. Divakara N, Assistant Professor of Dept. of Computer Science and Engineering, JSS STU who has supported us throughout our project with his patience and knowledge whilst allowing us the room to work in our own way.

8. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Ward Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

[2] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: an image-based virtual try-on network. In CVPR, pages 7543–7552. IEEE Computer Society, 2018.

[3] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristicpreserving image-based virtual try-on network. In ECCV (13), volume 11217 of Lecture Notes in Computer Science, pages 607–623. Springer, 2018.

[4] Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, abs/1511.06434, 2015.

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

[6] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Semantics disentangling for text-to-image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2327–2336

[7] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision. IEEE, 5907–5915.

[8] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 5337– 5345, 2019.

[9] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. IEEE Trans. Multimedia, 19(8):1946–1955, 2017.

[10] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In Proceedings of the IEEE International Conference on Computer Vision, pages 4481–4491, 2019.

[11] Jingyuan Liu and Hong Lu. Deep fashion analysis with feature map upsampling and landmark-driven attention. In Proceedings of the European Conference on Computer Vision (ECCV), pages 0–0, 2018.

[12] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 8340–8348.

[13] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In Advances in Neural Information Processing Systems. MIT Press, 406–416.

[14] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixellevel domain transfer. In ECCV, 2016.

[15] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. L. Chen. Be your own prada: Fashion synthesis with structural coherence. In ICCV, 2017.

[16] Nikolay Jetchev and Urs Bergmann. 2017. The conditional analogy gan: Swapping fashion articles on people images. In Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2287–2292

[17] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In Constructive theory of functions of several variables, pages 85–100. Springer, 1977.

[18] 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, pages 8798–8807. IEEE Computer Society, 2018.