High-Resolution Virtual Try-On Network with Coarse-to-Fine Strategy

Qi Lyu$^{1,a}$, Qiu-Feng Wang$^{1,b}$ and Kaizhu Huang$^{1,c}$

$^1$Department of Intelligent Science, School of Advanced Technology, Xi’an Jiaotong-Liverpool University
Email: $^a$Qi.Lyu19@student.xjtlu.edu.cn; $^b$Qiufeng.Wang@xjtlu.edu.cn, $^c$Kaizhu.Huang@xjtlu.edu.cn

Abstract. In this paper, we propose a high-resolution virtual try-on network model based on 2D images, which can seamlessly put on given clothing to a target person with any pose. Under the coarse-to-fine strategy, we firstly transform the given normal clothes to warped clothes to well match the pose of the person by a clothing matching module, then these two generated images are combined to generate one fitting image of the person put on the given clothes by a try-on module, lastly utilize a Very Deep Super Resolution (VDSR) module to refine the generated fitting image. Compared to the 3D based methods that are computationally prohibitive, our method only needs 2D images, which is much faster. We evaluate our proposed model both quantitatively (i.e., in terms of SSIM) and qualitatively on a public virtual try-on dataset (i.e., Zalando). The experimental results demonstrate the effectiveness of the proposed method: generating visually better quality of images, our new method can improve the SSIM by 1.5%.

1. Introduction

With the popularity of Internet, the online shopping has demonstrated booming market. For example, from 2009 to 2020 the volume of business of TianMao in the 'double 11' which is known as Singles’ Day gain from 50 million to 498.2 billion$^1$. One issue in the online shopping is that how to help customers try the product, especially for the clothing sale. Most of current online stores merely show pictures of clothing on models. However, customers may even want to know if the clothing could fit on themselves. Hence, there is a huge demand on developing virtual fitting techniques. For example, virtual fitting rooms emerged for customers to try on clothing, however, most of these rooms need 3D human shape measurements to match well the clothing on customers, which are usually computationally expensive [1]. In this paper, we propose a high-resolution virtual try-on network, facilitating customers to try-on clothing based on 2D images.

With the great success of deep learning [2], virtual fitting techniques have made significant progress [3, 4, 5, 6]. Han et. al [3] proposed a Virtual Try-On Network (VITON) based on 2D images to generate fitting images without replying on 3D information. The fitting image is transformed from two given images (i.e., clothes and person, respectively), where a big challenge is to make the given clothes to match the person’s pose. To obtain appropriate fitting images, the details of clothes and person’s pose and body should be fully utilized. To this end, Wang et.

$^1$https://en.wikipedia.org/wiki/Singles%27_day.
al [6] proposed a Virtual Try-On Network model with Characteristic-Preserving (CP-VTON) by a tailored convolutional neural network (CNN), which will generate better results than VITON, especially on the clothes with rich features and large deformation. However, the performance of CP-VTON is not good for some issues, including massive shape differences for clothes, rare persons’ poses, and neck artifact. Lassner et al [4] integrated the human parser [7] to overcome the issue of neck artifact. In addition, the fitting image of CP-VTON is usually blur. In this paper, we proposed a high-resolution virtual try-on netowrk (HR-VTON) to improve the quality of the synthesized try-on image on the public dataset Zalado.

2. Related Works

2.1. Image synthesis

Generative Adversarial Network (GAN). GAN [8] is one of the most successful application of the idea of adversarial learning [9, 10] in deep learning network for image processing, especially in the area of generating and editing images [11]. Under the framework of image-to-image translation [11], GAN has widely been applied to transform an input image to another one with same content and different styles.

Super-Resolution. To make the image show more details, super-resolution is widely applied. Motivated by the VGG-net, the work of very Deep Super-Resolution convolutional network (VDSR) [12] utilized a convolutional network with many layers for image super-resolution, where small-size filters are cascaded many times to efficiently exploit contextual information over large-size regions. In the super-resolution, it is also necessary to consider the performance of image classification. To this end, Qian et al [13] proposed a framework for super-resolution with three networks of Generator (G), Discriminator (D) and one additional network Classifier (C), where G is used to obtain the image with super-resolution, and D is used to distinguish the high-resolution image whether fake or truth, while C aims to correctly classify generated images with super-resolution.

2.2. Virtual fitting

Fashion analysis. Fashion analysis has been widely studied because it has a huge market potential. But most works mainly focused on clothing recognition and clothing analysis based on attributes [4]. Recently, some works utilized 2D images for virtual try-on clothing, which is very convenient and low-cost [14]. But the performance is not satisfied, because virtual try-on clothing requires to keep as much detail as possible of the target clothing image, including identical styles, embroidery, logos, text, etc [15].

Virtual try-on. Currently, most works of virtual fitting are based on 3D virtual fitting technology in Europe and the United States, where both models and clothing are 3D animation [1]. The advantage is that you can set the proportion of the model’s body to make it better match their own figure. The disadvantage is that the cost is too high. Each clothes needs special model to make 3D animation. To overcome this issue, some works [3, 6] proposed 2D image-based virtual try-on network using the deep convolutional neural networks.

3. High-Resolution Virtual Try-On Network

In this section, we will introduce the proposed High-Resolution Virtual Try-On Network (HR-VTON) in details. Figure 1 shows the general framework of HR-VTON, which consists of three sub-modules, i.e., clothing match module, try-on module and refine module. In the following, we will describe each part in details.

3.1. Person Representation

In the virtual fitting, one challenge is to map the clothing into the person posture. To this end, we represent a person as several important features, including posture, body shape, face and
hair, which will be concatenated to constrain the final image synthesis process [16].

Figure 1. An overview of the proposed HR-VTON model consisting of three sub-modules: A Clothing Matching Module to warp clothes, a Try-on Module to put warped clothes onto target person, and a Refine Module to improve resolution.

**Pose map.** Person’s pose is an important factor to decide the clothing on the person. In this paper, we utilized a popular pose estimation method [17] to obtain the person’s pose. Following the parameters setting in the work [17], we calculate person’s pose from 18 key points, and each key point is further transformed into a thermal map as the same image size [5]. Finally, all thermal maps are concatenated to 18 channels.

**Body shape.** The appearance of the clothes depends largely on the body shape, hence the selected fashion items depend on the position and shape of different body parts (e.g., arms or torso). In this paper, we apply the method of human parsing [7] to calculate the segmentation map of the human body, where different regions represent different parts, such as arms, legs, etc. In the segmentation, we convert the person image into a single channel binary mask, where the value of one represents the human body (Here the hair and face are excluded, which will be segmented in the following steps) and zero for others. Furthermore, the binary mask is reduced to a lower resolution to reduce the effect of artifacts when the target clothing and body shape are not matched well.

**Face and hair.** To represent person identity, we integrate physical attributes such as face, skin color and hair style. All of these can be obtained from the face and hair region segmented
by the human parser [7], and represented by RGB 3 channels.

Finally, we concatenate all output to obtain the person representation \( P = m \times n \times k \), where \( n = 192 \) means the width of the output image, and \( m = 256 \) means height of the output image in this work, and \( k \) means the number of channels (here, it is 22, i.e., 18+1+3). This final person representation is independent with the clothing information, and it will be input to the next virtual try-on sub-modules.

3.2. Clothing Matching Module

In the virtual try-on, there are usually different shapes between the target clothing and person. To overcome this issue, we proposed a clothing matching module (CMM) to transform the target clothing to warped clothing related to the person representation.

The paper [18] simulated this process with a learning module (Geometric matching module: GMM), which can be learned by end-to-end optimization. Motivated by this, we proposed a Clothing Matching Module (CMM) which includes three parts: (1) two networks to extract semantic features of given person \( P \) and targeted clothing \( C \), respectively; (2) one correlation matching layer to integrate two features; (3) One regression network to predict the parameters \( \theta \) of the spatial transformation (thin-plate pipeline: TPS [6]), which transforms the original image to the output \( C' = T_\theta(C) \). Finally, we adopted the pixel-wise L1 loss to optimize the module end-to-end, which is evaluated between the warped clothes \( c' \) and ground truth clothes \( c_t \) on the person image, as shown:

\[
L_{CMM} = ||c_t - c'||_1 = ||c_t - T_\theta(c)||_1.
\] (1)

3.3. Try-on Module

Since clothing match module outputs the warped clothing matching the person body shape, the aim of the proposed try-on module is to put the warped clothing on the target person to synthesize the try-on results. A simple solution is to combine the clothing with the target person image directly. Another solution is to model the conversion by a network (e.g., U-Network). By combining the idea from both aforementioned solutions, we firstly utilized a U-network to model the person representation and warped clothing, then estimated a composition mask \( \alpha \) to combine the rendered person \( I_r \) and warped clothing \( c' \) as shown:

\[
I_0 = \alpha \odot c' + (1 - \alpha) \odot I_r,
\] (2)

where \( I_0 \) is the output try-on image, and \( \odot \) represents element-wise matrix multiplication.

To estimate the parameters in this module, we minimize the L1 loss to evaluate the difference between the output try-on image \( I_0 \) and the ground truth image \( I_t \). To make the try-on image more perceptual, we integrated one VGG perceptual loss between two images:

\[
L_{VGG}(I_0, I_t) = \sum_{i=1}^{5} \lambda_i ||\phi_i(I_0) - \phi_i(I_t)||_1,
\] (3)

In order to preserve the detail characteristic of the target clothes, we prefer the composition mask \( \alpha \) to choose the suitable warped clothing as much as possible by adopting a L1 regularization \( ||1 - \alpha||_1 \). Finally, we get the completed loss function of Try-On Module:

\[
L_{TOM} = \lambda_{L1} ||I_0 - I_t||_1 + \lambda_{vag} L_{VGG}(I_0, I_t) + \lambda_{mask} ||1 - \alpha||_1.
\] (4)

3.4. Refine Module

Although the Try-on module produce the fitting image, there are several issues. For example, the limbs of the person are indistinct, the pants of the target person are difficult to preserve,
and the texture of target clothes is prone to excessive deformation. To this end, we proposed a refine module to improve quality of images generated by try-on module.

Specially, we integrated a very deep super resolution (VDSR) [12] to make the image more photo-realistic. Similar with the Try-on module, we combine both L1 loss and VGG perceptual loss to optimize the parameters:

$$L_{RM} = \lambda L_1||I_1 - I_t||_1 + \lambda_{vgg}L_{VGG}(I_1, I_t).$$  (5)

4. Experiments

In this section, we will firstly introduce the dataset information and experimental setting, then present experimental performance with both quantitative analysis and quantitative evaluation, finally, we will give some error analysis.

4.1. Dataset and experimental setting

**Dataset.** In this paper, we evaluate the proposed method on the public dataset Zalando, which is collected by Han et al[3] from the website[2], After removing those noise images without analytical results, there are 16,253 image pairs (i.e., person and clothes) in total, where 14,211 pairs and 2,032 pairs are used for training set and test set, respectively.

**Training Setup.** We set $\lambda_{L1} = \lambda_{vgg} = 1$ to all the experiments and use $\lambda_{mask} = 1$ in the composition mask. To train Clothing Matching Module, Try-On Module and Refine Module, we run 200K iterations with batch size 4, and apply Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set a fixed learning rate 0.0001 for the first 100K iterations and then decrease it linearly to zero in the remaining iterations. The image size is set as $256 \times 192$.

**Baseline model.** We set the clothing matching module and try-on module following the work of CP-VTON [6], including the representation network for person and clothing (four 2-stride down-sampling convolutional layers and two 1-stride layers) and the regression network for parameters in TPS (two 2-stride convolutional layers, two 1-stride layers and a fully-connected output layer) in the clothing match module, and 12-layer U-network in the try-on module. All details can be found in the work [6].

**Refine Module.** We follow the work [12] to set the structure of VDSR in the refine module. Specially, there are totally 20-layers with the same kernel size=$3 \times 3$, where the first layer operates on the input image, each of the next 18 layers is with 64 channels, and the last layer is used for image reconstruction. To reduce the computation redundancy, we stored the output images from try-on module, which were directly input to the refined module in the training.

Figure 2. Try-on images generated by CP-VTON (a and b) and HR-VTON (c and d).

4.2. Experimental Performance

The proposed High-Resolution Virtual Try-On network (HR-VTON) consists of three sub-modules as shown in Figure 1. In average, processing one sample pair takes 0.512s, 0.594s

---

2 www.zalando.de
and 0.074s for Clothing Matching Module, Try-On Module, and Refine Module, respectively, on the GPU of NVIDIA 1070Ti. Compared to the baseline CP-VTON, we integrated one refine module in the proposed model, but with very limit additional time consuming.

**Qualitative results.** Figure 2 shows a qualitative comparison of the results by the methods of CP-VTON and HR-VTON. Although both models roughly match clothes with target pose of the person, the results of CP-VTON (Fig. 2a and 2b) are a little blur and with some distorted details (e.g., arm). Compared to CP-VTON, the synthesized try-on images (Fig. 2c and 2d) is with higher quality.

**Quantitative results** In order to show the effectiveness of the proposed model, we compared it with several alternative methods on the quantitative metric. We adopt the Structural Similarity(SSIM) [19] to evaluate the similarity of the output try-on image and the ground-truth image, and the SSIM is higher then method is better. Table 1 shows the SSIM values of different methods on the Zalando dataset. We can see that the proposed model achieves the highest SSIM value, which demonstrates its effectiveness.

| Method   | PRGAN[20] | CAGAN[21] | CRN[22] | VITON[3] | CP-VTON[6] | HR-VTON |
|----------|-----------|-----------|---------|----------|------------|---------|
| SSIM     | 27.3%     | 21.8%     | 69.1%   | 77.2%    | 80.3%      | 81.8%   |

**4.3. Failure cases**

![Failure Cases](image)

**Figure 3.** Some failure cases and their possible reasons: (a)Rare poses, (b)Massive shape differences, (c)Neck artifacts.

Although the proposed HR-VTON improves the quality of synthesized try-on image effectively, there are still some failure cases as shown in Figure 3. The reason may include rare poses of the person (Figure 3.a) and massive shape differences (Figure 3.b) between target clothing and clothing in the person. In addition, Figure 3. (c) shows some failure cases caused by the neck artifacts. In the original clothing picture, we can usually see the size tag in the neck position, but in the actual fitting, this tag will be blocked by the human body.

**5. Conclusions**

In this paper, we proposed a High-Resolution Virtual Try-On network (HR-VTON) model to synthesize virtual fitting images, which consists of three sub-modules, namely, a clothing matching module, a try-on module and a refine module. The clothing matching module aims to warp clothes to match the shape of person body. With the warped clothes, the try-on module learns to put clothes on the body, and the output will be further optimized by the refine module. We evaluated this proposed method on the public try-on dataset Zalando. Both qualitative analysis and quantitative metric demonstrate the effectiveness of the proposed method.
Acknowledgement

The work was partially supported by the following: National Natural Science Foundation of China under no.61876154 and 61876155; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no. BK20181189, BK20181190, BE2020006-4; Key Program Special Fund in XJTLU under no. KSF-T-06, KSF-E-26, and KSF-A-10, and XJTLU Research Development Fund RDF-16-02-49;

Reference

[1] Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B.: Synthesizing training images for boosting human 3d pose estimation. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 479-488. IEEE (2016)
[2] Huang, K.Z., Hussain, A., Wang, Q.F., Zhang, R., : Deep Learning: Fundamentals, Theory, and Applications, Springer, ISBN 978-3-030-06072-5, 2019.
[3] Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In CVPR 2018, pp. 7543 - 7552.
[4] Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. In: ICCV 2017, pp. 853 - 862.
[5] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in Neural Information Processing Systems. pp. 405-415, 2017.
[6] Wang, B.C. Zheng, H.B. Liang, X.D., Chen, Y.M., Lin, L., Yang, M.: Toward Characteristic-Preserving Image-based Virtual Try-On Network. in IEEE ECCV (2018)
[7] Gong, K., Liang, X., Shen, X., Lin, L.: Look into person: Self-supervised structure sensitive learning and a new benchmark for human parsing. In: CVPR 2017, pp. 932 - 940.
[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672 - 2680 (2014)
[9] Yang, G.Y., Huang, K.Z., Zhang, R., Goulermas, J., Hussain, A.: Inductive Generalized Zero-shot Learning with Adversarial Relation Network, European Conference on Machine Learning (ECML), 2020.
[10] Lv, C.C., Huang, K.Z., Liang, H.N.: A Unified Gradient Regularization Family for Adversarial Examples, In IEEE Fifteen Conference on Data Mining (ICDM), 2015.
[11] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR 2017, pp. 5967 - 5976.
[12] Kim, J., Lee, J.K., Lee, K.M.: Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In IEEE CVPR 2016.
[13] Qian Z., Huang K.Z., Wang Q.F., Xiao J.M., Zhang R., Generative Adversarial Classifier for Handwriting Characters Super-Resolution, Pattern Recognition, 2020, vol. 107: 107453
[14] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: Highresolution image synthesis and semantic manipulation with conditional gans. In CVPR 2018, pp. 8798 - 8807.
[15] Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. In ICCV 2017, pp. 1689 - 1697.
[16] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE T-PAMI, vol. 24, No. 4, pp. 509-522, 2002.
[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. In CVPR, 2017.
[18] Ignacio Rocco, Belja Arandjelovic, Josef Sivic: Convolutional neural network architecture for geometric matching. In: CVPR 2017, pp. 6148 - 6157.
[19] Wang Z. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing, 2004.
[20] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool. Pose guided person image generation. In NIPS, 2017.
[21] N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In ICCVW, 2017.
[22] Q.F. Chen, V. Koltun : Photographic image synthesis with cascaded refinement networks. In: ICCV 2017, pp. 1511 - 1520.