A Segmentation Guided Coarse to Fine Virtual Try-on Network for a new Clothing and Pose

Dashdorj Sandagdorj, Thai Thanh Tuan, Heejune Ahn
Seoul National University of Science and Technology

dashka8009@gmail.com, thaithanhtuan@seoultech.ac.kr, heejune@seoultech.ac.kr

Abstract

Virtual try on is getting interested from researchers these days because its application in online shopping. But single pose virtual try on is not enough, customer may want to see themselves in different pose. Multiple pose virtual try on is getting input as customer image, an in-shop cloth and a target pose, it will try to generate realistic customer wearing the in-shop cloth with the target pose. We first generate the target segmentation layout using conditional generative network (cGAN), and then the in-shop cloth are warped to fit the customer body in target pose. Finally, all the result will be combine using a Resnet-like network. We experiment and show that our method outperforms stage of the art.

1. Introduction

Today, people are buying from offline to online shopping. Because last few years online shopping technology improved quickly most computers and smartphones support AR, and online payment methods easier than ever. However, in most cases, when items delivered costumers unsatisfactory that products because of different online images and real product, especially fashion products. How does the need to consider customers when wearing fashion products on viewers? To solve this problem people are use Virtual try-on. In short, a virtual try-on is the way a customer can "try-on" a product through mobile or other devices equipped with a camera. With a virtual try-on, a customer can see the product and see how they look well with the planned outfit. VITON[1], CPVTON[2], ACGPN[3], these are popular methods of Virtual try-on. These methods create image and change the clothes of source images, but these methods don't work well and only single pose. In this paper, we propose a try on system. It automatically generates source image of person with planned clothing image desired pose. To make it look the most realistic photo, the try-on system needs to function well with a large variation in viewing angles, costumer's distinction as well picture on the planned cloth.

Figure 1. Example pf Multiple pose Virtual Try-on generated by our method.

As shown in Figure 1, from on customer image (I), an inshop cloth (c) and a target pose (pt), the system try to generate try on image (Ito). The try on image capture indentify from customer image such as skin color or pant,… wearing inshop cloth and in new target pose. For this system source image coloring must be natural and realistic because making new perspective requires the system to color itself in angles that cannot be seen in the input image. When looking at panoramic view of old image and unknown parts of new image, people cannot know difference between real image and colored image area. Moreover, when clothes are wearing, icon on the clothes must be natural (I.e shapes, words and logos). The icon on the clothes must be match with the figure and posture of the model image. To solve this problem, we offer a multi-component system to image creation from target pose. source image and sample clothes. The photo changed methodically to the target body figure and make obscure parts in the source image, and clothes are divergence to fit the human's figure, the target pose while keep special. The structure and design on the clothes are most simple kept and divergence on the figure.

2. Related work

2.1. Person image synthesis.

Advances in GAN[7] together with following methods gives creation of realistic images. Randomly generated photos, leades are included to target photo methodically to the user's source in the conditional GAN[8]. Successful information-driven applications named pix2pix do create a realistic photo with obvious information bonding is the semantic segment of photo. In this way generators do not try to faithfully reproduce only the fake image, but also the most realistic dummy image and features. Clearer and more detailed images pix2pixHD[9] can be produced with the application of multiscale and feature matching loss from the discriminator to better train the generator.

2.2. Pose-guided Person Image Generation

Creating different perspectives and poses from the original reference image has also been a focus of recent research. PG2[10] suggestive to make the raw photo first then make the refined photo by GAN style. Methods on this task divided 2 parts. The first part is model image generation conditioning target pose information using data, PG2 suggestive to first merge coarse

result and refine the generated photo with training. Progressive training with attention mechanism has been proposed to focus on each transfers certain regions while generating the person image progressively[5]. Deformable GAN[11] introduce deformable skip connections in the generator to deal with pixel to-pixel misalignment caused by the pose differences. Another group of method is model image synthesis. Semantic parsing transformation network was employed in the unified unsupervised person image generation framework[4] to guide the generator to generate image in the precise region level. Bidirectional generator[6] was utilized in the generator and the whole pipeline could be trained in an unsupervised manner.

2.3. Virtual Try-on:

Virtual try-on networks can be categorized into two kinds of approaches: 1) 3D-model-based approaches and 2) clothing warping-based. Although many studies based on 2D image work on the virtual try-on task, numerous researches aimed to utilize 3D body shape and 4D sequence to make the results more realistic. Thin Plate Spline (TPS) is a spline-based technique that prevails in the non-rigid transformation of images without going through any generator. Therefore, warping clothes directly by TPS is widely used in many try-on researches since it warps clothes and preserves patterns, texture, and logos.

2.4. Multiple pose virtual try on:

Virtual try-on only shows one pose of user. In MGVTON[13], to generate try-on image from users' image with clothing arbitrary pose. Try to generate synthesize human segmentation from input cloth and target pose, then after generate try-on blurry. Using refinement network for improvement the details on image to get more realistic image. But in their segmentation, they failed to generate bottom parts like legs and shoes labels, because lacking of information to guide the network. Because of unbalanced dataset, some small label require the network to train for a long time for better recognition, in this paper, show a way to train the network with equalized for faster and higher in accuracy. After having human segmentation, they use extracted body shape to warped the cloth, by this way, warped clothes are deformed strongly. We apply some improvemtn in CPVTON+ [14] such as regularization loss for better preserve the detail of clothes.

3. Proposed Method

We propose a refined pineline to synthesize a new realistic image for virtual try on in which not only changing cloth but also changing human pose. From an uploaded human image by the customer, the system can generate a new image which preserves the identity of uploaded human image, wearing new in-shop desired cloth in a new pose. We use synthesized semantic segmentation guidance for the coarse to fine generated human image. The networks are composed of four stages, a new Synthesize human segmentation(SHS), Generate Warped Cloth(GWC) and Try-On Generation(TOG) networks as shown in Figure 2.

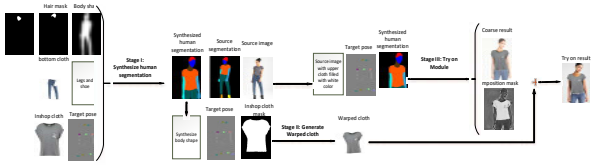

Figure 2. Our Methods.

3.1. Learning to synthesize human segmentation:

From customer image I, we use a human segmentation to extract face mask(Mf), hair mask (Mh), body shape (Ms), left leg(Ll) right leg(Lr), bottom cloth(Cb), then combine with inshop cloth(C) and target pose (P). To easy feed pose to the network, for each keyjoint, we build a heatmap with value of one near in neighbor of 4 pixels the keyjoint and zero at other places. Then the synthesize human segmentation(SHS) is to generate the target human segmentation (St) as in Figure 3.

$$St = SHS(Mf,Mh,Ms,Ll,Lr,Cb,C,P) \quad (1)$$

We add bottom part of customer image (I) to the input of SHS to guide the segmentation for generating the output for accuracy. Moreover, we notice that some small area on human such as shoes and legs,… are rarely appear in the dataset, this lead to train the network more time. So handle this problem, we apply a special loss named 'equalized' label weighted loss. The idea is to increase the weight of small areas so that they can have same probabilities in the networks. We do statistical on the dataset to calculate the weight. But this make the true probabilities of each label to be the same for all the pixels. So, after training with 'equalized' label weighted loss in 3 epochs, we change back to normal cross entropy loss.

We train the network follow GAN style with generator and discriminator. We follows the pix2pixHD to build the discriminator and Resnet like to build the generator. The GAN loss can be as follow:

$$L_cGAN(Gt,Dt) = E_{SHSin}\left[logD\left((SHSin),S_t\right)\right] + E_{SHSin}\left[\log\left(1 - D\left((SHSin),G_t(SHSin)\right)\right)\right], \quad (2)$$

where $SHSin$=(Mf,Mh,Ms,Ll,Lr,Cb,C,P). and loss of synthesized label is:

$$L_{BCE}^{G_t}(G_t) = -\sum_{n_c} S_t \log\left(G_t,(SHSin)\right) + (1 - S_t)\log\left(1 - G_t(SHSin)\right), \quad (3)$$

For 'equalized' label weighted loss, we put a weighted on corresponding label. And the whole SHS will try to optimize the following equation:

$$arg \min_{G_t} \max_{D_t} L_cGAN\,(G_tD_t) + \lambda_{bce}L_{bce}^{G_t}(G_t). \quad (4)$$

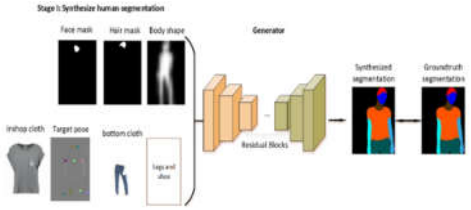

Figure 3. Synthesize human segmentation.

Figure 5 compares MG-VTON segmentation results and ours, where MP-VTON fails in generating the bottom parts correctly. With hair and face mask as input, MG-VTON can generate hair style is same with input image but without bottom parts of source image, MG-VTON cannot preserve correctly source parts, and from statistic learning from data, MG-VTON can only generate pant as bottom parts.

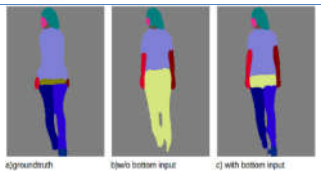

Figure 5: Comparison of Synthesized segmentation.

3.2. Generate warped cloth:

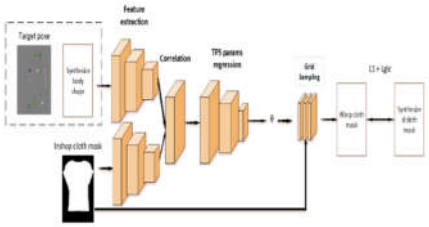

Figure 6: Generate warped cloth.

Spatial transformation network(STN) is a good tool to generate warped cloth (C_{warped}) which is widely use in CPVTON[2], CPVTON+ [14] and MGVTON[13]. The inshop cloth mask(Mc) feature and target body {target pose(Pt), synthesized body shape(Bs')} feature are extracted separately by two feature extraction networks. Then, correlation between the two features is calculated based on dot matrix multiplication. This can be considered as a calculation of cosin angle between each vector in feature source to every target features. This correlation is then fed into a regression network to estimate the transformation parameters between source and target. By applying this regressed transformation to the in-shop cloth, we can generate warped cloth which is fit to the target body. Unlike MGVTON, we apply the improved version of STN to warp cloth form CPVTON+ [14]. Finally, the experiments above reveal that the warped clothing is often severely distorted. We conclude that the TPS parameters estimation needs the regularization, to take into account the restriction of clothing textures. Our grid warping regularization is defined on the grid deformation and not directly on the TPS parameters for the purpose of easy visualization and understanding, which means that less different warping between two grids in equation 7. The total loss for training GWC is:

$$L_{GMM}^{CP_{VTON}+} = \lambda_1 L1(C_{warped}, I_C) + \lambda_{reg} L_{reg}(G_X, G_Y) \quad (5)$$

where λ_1, λ_{reg} are hyper parameters and L_{reg} can be calculate as:

$$L_{reg}(G_x, G_Y) = \sum_x \sum_y |G_x(x+1,y) - G_x(x,y) - |G_x(x,y) - G_x(x-1,y)| + |G_Y(x,y+1) - G_y(x,y)| - |G_Y(x,y) - G_y(x,y-1)| \quad (6)$$

where Gx, Gy is the location of grid of pixel after being warped.

3.3. Try on Generation:

Fashion on[12], MGVTON[13] is used without cloth reference image provided guarantee cloth agnostic because we train on image from same model wearing the in-shop cloth. Consequently, there are some information from source reference image cloth can be inferred from the shape of cloth on the reference image, such as cloth type. The network can also know the information about the shape of cloth such as body fit cloth or loose cloth, then, it can be hard for the network to learn the new shape of new in-shop cloth while testing. Motivating from CPVTON+ , we estimate composition mask at the same time with the coarse result while MGVTON makes use of a new separate refine network to render generator to do such a task.

From source image, we colorize upper cloth and background to white color and get (I_{woc}),

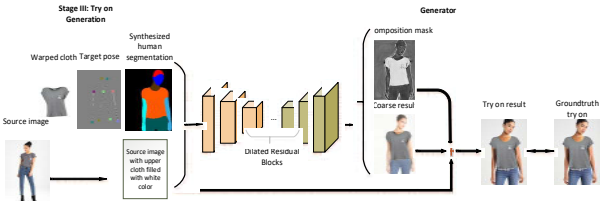

Figure 9: Try on Generation network.

We use Resnet-like network to implement try on generation network. To increase the receptive field of the network to transfer information from source image to target location, we make use of dilated convolutional layer in the middle layer of TOG. We train the network with GAN style network to generate composition mask($M_{composition}$) and coarse results(I_{coarse}^t), which is firstly used in CP-VTON[2] to preserve the texture details of cloth. This stage can be formulated as:

$$(M_{composition}, I_{coarse}^t) = TOG(S_t, Pt, C_{warped}, I_{woc}) \quad (7)$$

Where $M_{composition}$ is the composition mask and I_{coarse}^t is the coarse virtual try on result. Then these two are fused with warped cloth to generate refined final try on It final.

$$I' = M_{composition} * C_{warped} + (1 - M_{compostion}) * I_{coarse}^t \quad (8)$$

To handle mis-alignment between source image and target location, we make use of dilated convolutional layer to increase the receptive field for better handle mis-alignments. We use dilated convolution layer at the innermost layer of Unet generator. We train the network in supervised way with the following loss function:

$$L = \lambda_1 |I' - I_{GT}| + \lambda_{VGG} L_{VGG}(I', I_{GT}) + \lambda_{GAN} L_{GAN}(I', I_{GT}) \quad (9)$$

Where λ_1, λ_{VGG}, λ_{GAN} are hyper-parameters, L_{VGG} is a perception loss between final try on (I') and ground-truth image(I_{GT}), L_{GAN} is GAN loss of the pix2pix model.

4. Experiment and Result

4.1. Datasets and Settings.

The dataset used for experiments is the MPV dataset [13], which consists of 35,687/13,524 person/clothes images, with the resolution of 256x192. Each person has different poses. We split them into train/test set 52,236/10,544 three tuples, respectively. We further generate 10,544 three tuples of different clothes for testing. By default, the learning rates for the discriminator and the generator are 0.0002. We adopt ADAM optimizer to train our network withβ1=0.5, β2=0.999.

4.2. Results and Comparison

Table 1 compares the intersections over union (IOU) and pixel accuracy of the synthesize human segmentation using our networks and MP-VTON's, i.e., with and without bottom parts as inputs and the equalized cross-entropy loss. Our methods show significant improvement over MGVTON's. Especially for some regions which are small or rarely appeared in the dataset and bottom cloth area. Other benefits of our training methods with equalized cross entropy shows faster (50 epochs over 200 epochs).

Table 1. Comparison of synthesize human segmentation result.

	MG-VTON 200 epochs		Our proposed method (After 50 epochs)	
	Accuracy	IOU	Accuracy	IOU
background	0.97827595	0.95317669	**0.98575394**	**0.96492789**
hat	0.	0.	**0.33436449**	**0.32636166**
hair	0.72218297	0.61588937	**0.77150404**	**0.69145163**
glove	NA	NA	NA	NA
sunglasses	0.	0.	0.	0.
upper-clothes	0.91791136	0.78900368	**0.9343231**	**0.8771459**
dress	0.02783092	0.02575886	**0.84958977**	**0.71720552**
coat	0.01427718	0.01365604	**0.59081235**	**0.54606807**
socks	0.03688724	0.03158321	**0.17987053**	**0.16224863**
pants	0.8552432	0.662966	**0.90734984**	**0.83266766**
neck	0.66336829	0.54690293	**0.81447017**	**0.6758383**
scarf	0.	0.	0.	0.
skirt	0.09012766	0.07712106	**0.88980229**	**0.72361228**
face	0.89533577	0.8294081	**0.93819425**	**0.85506298**
left-arm	0.78642302	0.67887898	**0.87036264**	**0.77587756**
right-arm	0.75983363	0.64760119	**0.86240729**	**0.76454793**
left-leg	0.13981736	0.11598884	**0.67977078**	**0.58171521**
right-leg	0.13294737	0.11021019	**0.65307499**	**0.57050878**
left-shoe	**0.63280849**	0.44877137	0.61813586	**0.48990337**
right-shoe	0.6285375	0.43084955	**0.66402908**	**0.50732609**

Figure 10 shows the comparison between our method and state of the art MGVTON, due to the accuracy of synthesize human segmentation, we can generate clear boundary. In MGVTON, warped cloth is strongly deformed while our warped clothes can give better result.

Figure 10: Comparison between MGVTON try on result and our proposed method.

5. Conclusions

In this paper, we proposed a method for transfer both cloth and pose of reference image. By investigate each intermediate output from each step, we can modify reasonable network structure and loss. In these methods, more attentive to bottom part of reference cloth give better human parsing synthesis. Base on improvement of human parsing synthesize, generated warped cloth base on CPVTON+ is generate. Inside try-on network, a dilated convolutional layer is used for fixing unperfect aligned from previous step. Experiments shows our proposed method significant improvement to state of the art MGVTON. Improved version of GMM is still based on 2D image and it still gets limitation of strong 3D deformation. In future work, a hybrid method combines between 2D and 3D can help to improve the performance.

Acknowledgement:

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2018R1D1A1B07043879)

References

1.Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. Viton: An image-based virtual try-on network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7543-7552.

2. Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L.; Yang, M. Toward characteristic-preserving image-based virtual try-on network. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 589-604.

3.Yang, H.; Zhang, R.; Guo, X.; Liu, W.; Zuo, W.; Luo, P. Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7850-7859.

4. Song, S.; Zhang, W.; Liu, J.; Mei, T. Unsupervised person image generation with semantic parsing transformation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2357-2366.

5. Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; Bai, X. Progressive pose attention transfer for person image generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2347-2356.

6. Pumarola, A.; Agudo, A.; Sanfeliu, A.; Moreno-Noguer, F. Unsupervised person image synthesis in arbitrary poses. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8620-8628.

7. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 2014, pp. 2672-2680.

8. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125-1134.

9. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis andsemantic manipulation with conditional gans. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798-8807.

10. Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; Van Gool, L. Pose guided person image generation.Advances in neural information processing systems, 2017, pp. 406-416.

11. Siarohin, A.; Sangineto, E.; Lathuiliere, S.; Sebe, N. Deformable gans for pose-based human image generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3408-3416.

12. Hsieh, C.W.; Chen, C.Y.; Chou, C.L.; Shuai, H.H.; Liu, J.; Cheng, W.H. Fashion On: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 275-283.

13. Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; Yin, J. Towards multi-pose guided virtual try-on network. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9026-9035.

14. Minar, M.; Tuan, T.; Ahn, H.; Rosin, P.; Lai, Y. CP-VTON+ : Clothing Shape and Texture Preserving Image-Based Virtual Try-On. IEEE CVPR Workshop on CVFAD (Computer Vision for Fashion and Art Design, 2020, pp. 123-125.

15. Rocco, I.; Arandjelovic, R.; Sivic, J. Convolutional neural network architecture for geometric matching. CVPR, 2017, pp. 6148-6157.