Deep Virtual Try-on with Clothes Transform

Szu-Ying Chen(✉), Kin-Wa Tsoi(✉), and Yung-Yu Chuang(✉)

Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
{clairecat,rance1108}@cmlab.csie.ntu.edu.tw, cyy@csie.ntu.edu.tw

Abstract. The goal of this work is to enable users to try on clothes by photos. When users providing their own photo and photo of intended clothes, we can generate the result photo of themselves wearing the clothes. Other virtual try-on methods are focused on the front-view of the person and the clothes. Meanwhile, our method can handle front and slightly turned-view directions. The details of the clothes are clearer. In the user study, about 90% of the cases, respondents chose our results over others.

Keywords: Convolutional neural network · Virtual try-on · Perceptual loss

1 Introduction

In recent years, the demand for online shopping has been increased dramatically, one of the best-selling commodities is clothes. However, the problem of online purchasing clothes is people cannot try the clothes and see if they are suitable for them until they buy them. Therefore, it is very convenient for consumers to try on clothes virtually.

There are some researches on image-based virtual try-on. However, most of them focus on front-view clothes and people. We want to expand the clothes and people to different viewing angles.

Therefore, we want to develop a system that uses images for virtual try-on, which allows trying on clothes without limiting the view direction of people and target clothes. In this way, consumers can try on clothes more easily which promotes desire to purchase and keep the cost down for clothes stores.

2 Related Work

Recently, plenty of researches have been conducted on fashion-related works. Virtual try-on is the more challenging task among them, which should preserve more details of a target clothes as output. Two papers have been conducted on this task for the recent two years, VITON [3] and CAGAN [4], which deliver good results on image-based virtual try-on with aids of deep learning.

C.-Y. Chang et al. (Eds.): ICS 2018, CCIS 1013, pp. 207–214, 2019.
https://doi.org/10.1007/978-981-13-9190-3_22

2.1 VITON

VITON consists of two stages. In the encoder-decoder generator stage, the network generates a coarse result. While in the refinement stage, target clothes is warped with the mask by estimating a thin plate spline (TPS) transformation with shape context matching [1]. After warping, a network composites the coarse image with the warped clothes and generates a final output.

However, VITON requires accurate poses and segmentations, which needs manually fine tuning and are difficult to be obtained.

2.2 CAGAN

Conditional Analogy Generative Adversarial Network (CAGAN) is based on Conditional GAN (CGAN) [6]. Given a person image and a target clothes image, CAGAN can output an image with the target clothes worn by the person. Moreover, CAGAN can also learn a segmentation mask of difference between the input and output person image without labeling data.

However, the output from CAGAN is blurry and preserves less details than the target clothes.

3 Methodology

3.1 Data Collection

Since we require different viewing directions of clothes, we select the MVC dataset [5] as our training data, which is used in clothing retrieval and clothing style recognition. It contains 161260 images in 9 categories. Since our task is focused in upper-body clothes, images of followed three categories are selected from the dataset, *Shirts and Tops, Sweaters and Cardigans, Coats and Outerwear*. Meanwhile, the datasets include 4 different viewing directions of people wearing the same clothes and 3 of them are used in our task (front, left, and right views).

3.2 The Proposed Approach

Our proposed method contains four steps. Firstly, given a person image and a target clothes image, CAGAN is used to generate a preliminary result and a binary mask of where to change. Secondly, a transform network is used to extract the clothes only. Meanwhile, the mask and output from previous step are used in segmentation step for a better mask to indicate where the clothes should be transformed to. Next, the mask is used to transform the target clothes. Lastly, the transformed clothes and the output from CAGAN are combined which becomes our final output. Figure 1 shows our overall architecture.

Figure 2 shows the generator and discriminator of CAGAN. Given a person image x_i and a clothes image y_j as inputs, it gives a intermediate output x'' and a alpha mask M'. Then the input person x_i and the intermediate output x''_j are merged according to the mask M' and generate the final output of that person wearing the clothes x'_j.

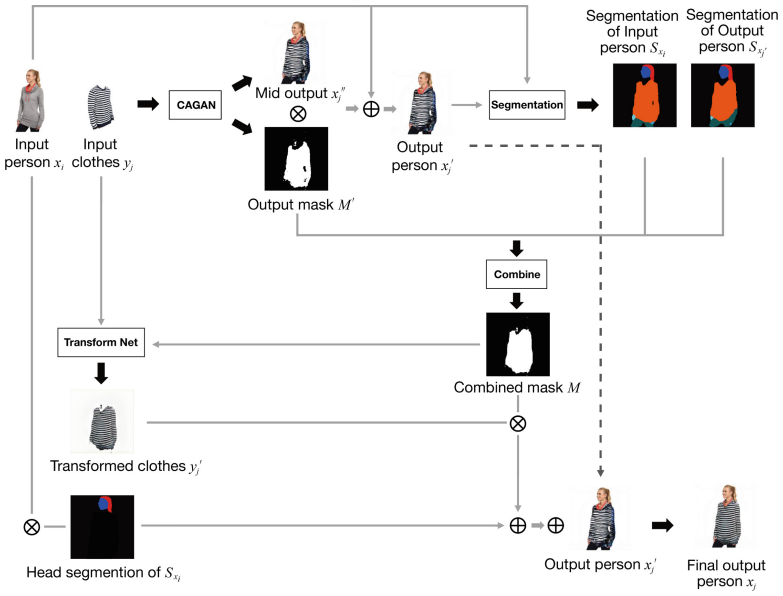


Fig. 1. Overall architecture of our method

CAGAN [4] is trained with our own dataset to suit our task. During training, image from the generator is used to fool the discriminator while discriminator try to discriminate it. Which can result in an improvement of the generated image.

The loss of CAGAN is defined as:

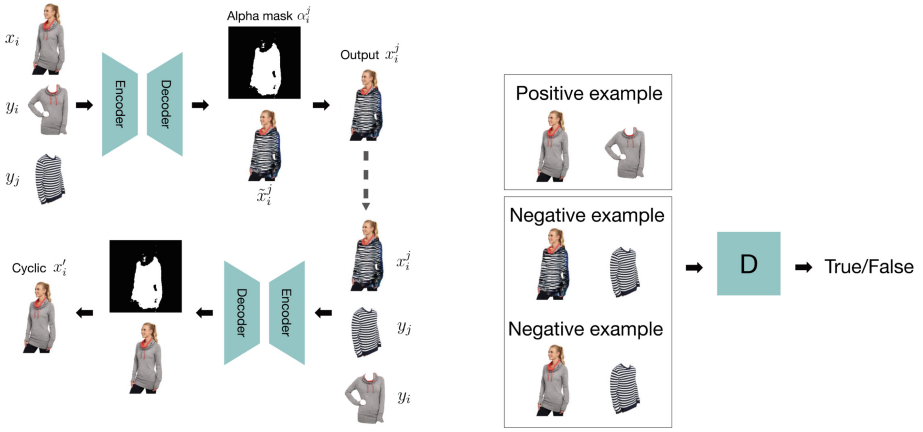
$$\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \gamma_i \mathcal{L}_{id}(G) + \gamma_c \mathcal{L}_{cyc}(G), \tag{1}$$

where consists of three terms: the adversarial loss of CAGAN, the regularization of generator and the cycle loss of generator, which is weighted by γ_i and γ_c.

The adversarial loss of CAGAN is different from typical GAN as the discriminator have to discriminate both the rationality of the image and the correctness of the changed clothes, as shown in Fig. 2b. The regularization term restricts the output mask from the generator from getting much difference compared with the original image. The cycle loss of the generator keeps the other parts remain unchanged except the clothing part.

Segmentation. Since the output masks of CAGAN are often shattered, It is difficult to transform the target clothes into masks. To solve this problem, modification of the mask is needed to improve the integrality of it.

The clothing part of the person image is the mask region. A state-of-the-art human parser, LIP-SSL [2] is used to capture the clothes worn by people. The human parser network is used to obtain a segmentation map with 19 human parts with clothes.



(a) Generator of CAGAN (b) Discriminator of CAGAN

Fig. 2. Network of CAGAN

The final combined mask M is then calculated by:

$$M = (S_{x_i}(clothes) \cup S_{x'_j}(clothes) \cup M') - S_{x_i}(head) - S_{x'_j}(hands). \quad (2)$$

Where S_{x_i} and $S_{x'_j}$ are the segmentation maps from the original image x_i and CAGAN x'_j respectively.

Figure 3 shows some results of our map is cleaner and more integral.

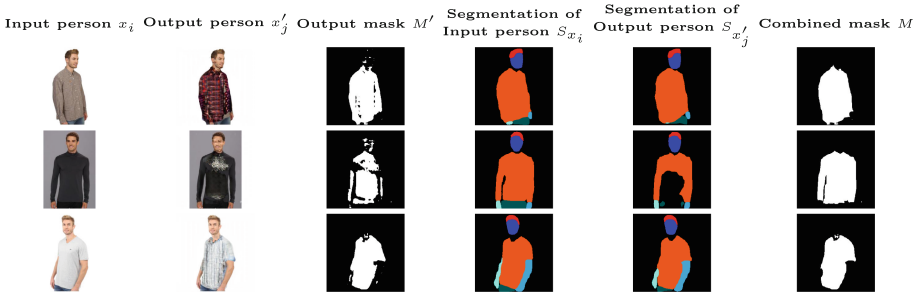


Fig. 3. Results of the segmentation step

Transform. The combined mask M is then used for transformation. The target clothes y_j and the mask M act as an input for the Transform Net, which gives an output of transformed clothes y'_j.

The architecture of the Transform Net is based on U-net, with skip connections to directly share information between layers.

Supervised learning is used to train the Transform Net. Two clothes with different viewing directions and a binary mask of one of the clothes are needed. Let $\hat{y}$ be the ground truth of the clothes image, y be the output of the Transform Net, and $\phi_j(y)$ be the feature map from the jth layer of the network VGG16 [7] for the input y. So is for $\hat{y}$. The loss function is defined as:

$$\mathcal{L}_T = \lambda_p \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2 + \lambda_m \|\hat{y} - y\|_2^2, \tag{3}$$

where is weighted by λ_p and λ_m. Since perceptual loss is computed on higher frequency information, the details and patterns of the inputs are preserved in the output image. While MSE loss requires lower frequency information, which results in blurry output but preserving more correct color.

Combination. The outputs from the Transform Net and CAGAN are combined in this step.

Let $\otimes$ denotes element-wise product, $\oplus$ denotes combining the contents. As above, with the input person image x_i, the transformed clothes y'_j, the head part of the input image $S_{x_i}(head)$, the combined mask M, the output from CAGAN x'_j. Then the final output with changed clothes x_j can be solved as follows:

$$x_j = (x_i \otimes S_{x_i}(head)) \oplus (y'_j \otimes M) \oplus ((1 - S_{x_i}(head) - M) \otimes x'_j). \tag{4}$$

Figure 4 shows some results of the above-mentioned steps.

3.3 Experiments

Implementation Details. CAGAN and the Transform Net are implemented on Keras with Tensorflow backend. Parameters used for training CAGAN are the same as original paper, where the learning rate and batch size is set at 0.0002 and 16 respectively and it is trained with our 18573 image pairs. 15000 training steps are used rather than 10000 steps from the original paper, as a better results can be obtained in theory.

The learning rate and batch size is set at 0.0001 and 4 respectively for the Transform Net and it is trained for 30 epochs with 14858 image pairs while validating with 3715 image pairs. The weights of loss equation 3 are $\lambda_p = 0.9999$ and $\lambda_m = 0.0001$ respectively. Fifth layer of VGG16 ($j = 5$) is used.

4 Evaluation

4.1 Qualitative Evaluation

Comparison with VITON and CAGAN. Since VITON [3] also contains a procedure of warping clothes to the mask, it is used to compare with our proposed method in terms of warping stage performance.

VITON [3] uses shape context matching. With our unclean masks, estimation errors are occurred and the warping results from it may not fully match the mask.

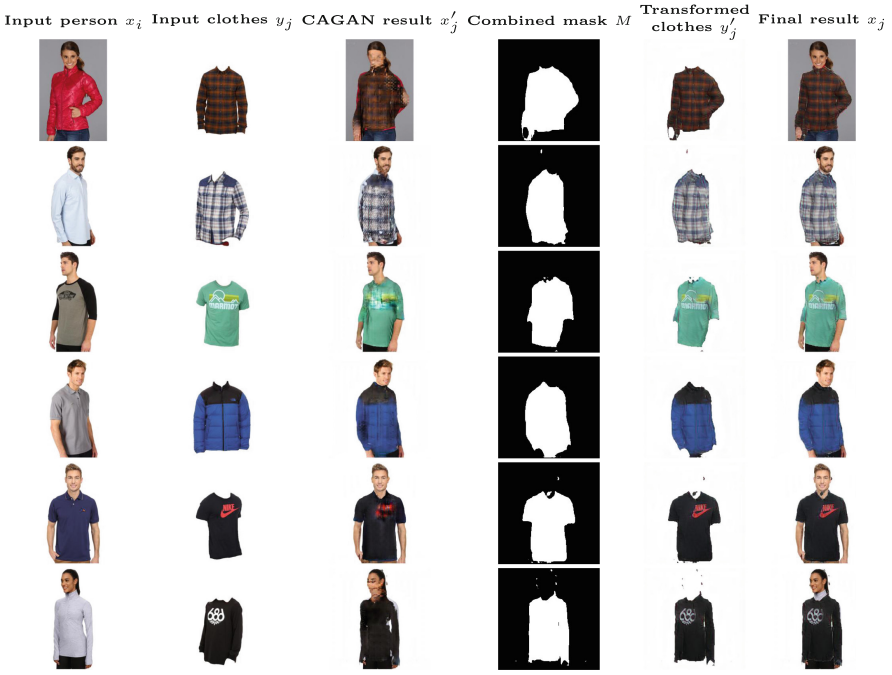


Fig. 4. Results of our proposed method

On the other hand, VITON [3] aims to warp clothes in the front-view. Masks of other viewing angles besides front-view will result in error of transforming the correct shape and details of the clothes by VITON [3].

Meanwhile, since our proposed method is based on CAGAN [4], comparison with CAGAN [4] is provided which showed that ours gives more details on the clothes than CAGAN [4] does. Comparison results are showed in Fig. 5.

4.2 User Study

29 volunteers were participated in our user study. CAGAN [4] is regarded as our baseline. 499 clothes is picked from our dataset for testing.

Two different versions of questionnaire are made. Each questionnaire contains 60 pairs images which were randomly sampled from the testing dataset. The questionnaires are showing two generated results, one from CAGAN [4] and one from ours, and asking respondents which one do they prefer.

Statistical results are presented in Table 1. 88.3%, 92.5% and 90.3% of the votes are in favor of our results in questionnaire A, questionnaire B and on average respectively. That is, in about 90% cases the respondents prefer our results over the results from CAGAN [4].

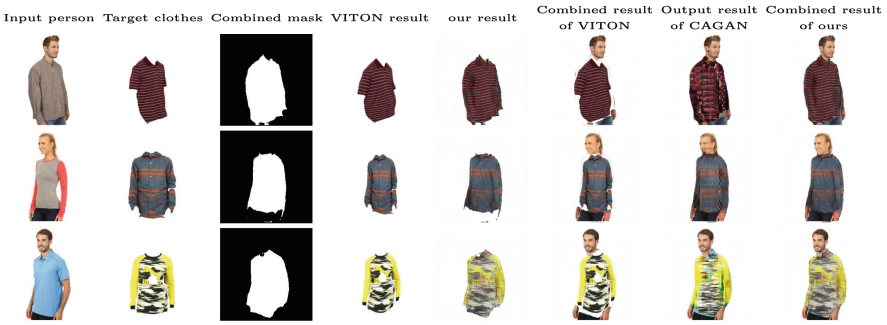


Fig. 5. Comparison with VITON [3] and CAGAN [4]

Table 1. Statistical results of user study

	Q_A	Q_B	Mean
Samples	15	14	
Our votes (%)	88.3	92.5	90.3

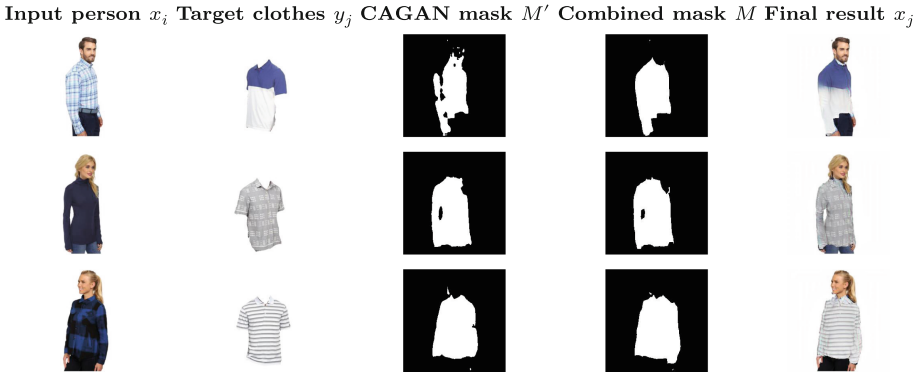


Fig. 6. Failure cases

5 Conclusion

5.1 Conclusion

We proposed an image-based virtual try-on system, which is able to change a clothes on a person image to another with multiple view directions, while preserving the details on clothes. Our system is consists of four steps:

i Use CAGAN to get a preliminary result and a mask.
ii Modify the mask with segmentation of input person and output person from CAGAN.
iii Transform the target clothes to the modified mask with Transform Net.

iv Combine the transformed clothes with the preliminary result from CAGAN.

In the user study, in about 90% cases, our results are preferred over CAGAN [4].

5.2 Discussion

The masks generated from CAGAN [4] are not completely correct. Since the places they represented have changed, if the input person wears a long-sleeves shirt while the target clothes is in short-sleeves, the mask will wrongly transformed with the arms part which does not belongs to the clothes part and resulting in the following failure examples shown in Fig. 6. To solve this issue, our future work aims to get a correct mask.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
2. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
3. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: an image-based virtual try-on network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
4. Jetchev, N., Bergmann, U.: The conditional analogy GAN: swapping fashion articles on people images. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, October 2017
5. Kuan-Hsien, L., Ting-Yen, C., Chu-Song, C.: MVC: a dataset for view-invariant clothing retrieval and attribute prediction. In: ACM International Conference on Multimedia Retrieval, ICMR (2016)
6. Mirza, M., Osindero, S.: Conditional generative adversarial nets. CoRR, abs/1411.1784 (2014)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)