Influence of Training Times and Sample Size on Results in Visual Try-On with GAN

Wenqing Jiang\textsuperscript{1,a,†}, Ruiyu Qi\textsuperscript{2,b,†} and Juntao Zheng\textsuperscript{3,c,†}

\textsuperscript{1}Computer Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, 519087, China
\textsuperscript{2}Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, PA, 18015, U.S.
\textsuperscript{3}Internet of things, Guangdong Ocean University, Zhanjiang, Guangdong, 524088, China

\textsuperscript{a}Corresponding author’s e-mail: an830026053mail@uic.edu.cn; \textsuperscript{b}ruq220@lehigh.edu; \textsuperscript{c}memphis50zzZ@163.com

\textsuperscript{†}These authors contributed equally.

Abstract. With online shopping becoming more and more trendy these days, Virtual Try-On Approach is proposed. It is a method of Generative Adversarial Network (GAN) composed of two models: Geometry Matching Module (GMM) and Try-On Module (TOM). It is proposed to solve the inconvenience of customers for not being able to see the fitting effect before receiving the goods. This paper is about finding the relationship between dataset and the effect of the experiment in the GAN experiment. After implementing the models and gaining the result images, we do some observations on the GMM one and try to analyze it and improve the effect. Comparison is made on different aspects: training times of the model and sample size of the training input. Similar trends are shown on the decrease of loss value when the training times or sample size rises. However, the visual effect is not necessarily improved all the time. Compared with the times it has been trained, the improvement from training times is far less than that from the sample size. Finally, we draw the conclusion that the effect is gradually improved as the training times or sample size increases at the beginning. Nevertheless, the improvement cannot be recognized evidently when the training times and sample size have grown large enough. Additionally, the sample size seems to be a more significant factor in the improvement of the visual effect than the training times.

1. Introduction
Nowadays, online shopping is becoming increasingly popular. While it is true that buying items online is quick in speed and convenient, customers often find it troublesome to be not sure about the actual look by only seeing the sale’s picture and return goods as a result. Thus, if customers can try on clothes without the effort of changing them physically, it will not only increase customer satisfaction, but also save shipping cost for retailers [1].

To solve this task, an image-based virtual try-on approach proposes a virtual try-on network (VITON) to synthesize the product image to the corresponding region of the clothed person [2]. The approach combines a U-net generator and thin plate spline (TPS) transform to transfer the patterns of clothes accurately. To improve the poor capability of preserving details in large geometric changes, a
Characteristic-Preserving image based Virtual Try-on Network (CP-VITON) is proposed to preserve cloth characteristics by a tailored convolutional neural network [3]. Similar to CP-VITON, Looking-Attractive Virtual Try-on Network (LA-VITON) is proposed to solve the limitation of VITON in generating fine and sharp images. It regularizes the transformation with grid interval consistency loss for more attractive looking results without distortions [4]. In the aim of reducing distortions and artifacts in the generated images, an image-based Virtual Try-On Module with Geometric Transformations (VITON-GT) is designed [5]. It is composed of a two-stage geometric transformation module and a transformation-guided try-on module. It learns an affine transformation to move the input clothes closer to the target body and increase the realism of generated images. Most recently Jandial et al. present a unified framework for robust image-based virtual try-on to improve the quality of generated try-on results [6]. It uses a multi-stage warping module with perceptual geometric matching loss to transform the try-on cloth to align with target model. It then uses a conditional segmentation mask generation module and a dueling triplet loss strategy to improve the texture transfer network.

Though there have been successful virtual try-on networks to synthesize the product cloth to the customer image, it is reported that some of these networks generates blurred arms in the final output when arms are crossed in front of clothes due to reconstruction loss [7]. For alleviating this problem, an adversarial loss is introduced from Generative Adversarial Networks. GAN, designed by Ian Goodfellow and his colleague, is a set of machine learning frameworks consisting of generator and discriminator [8]. The generator module generates new examples, and the discriminator tries to tell whether an example is real or fake [9]. GAN produces authentic looks, and it has been successfully used to generate high-resolution and realistic images because adversarial loss can incorporate perceptual features [10].

An image generator called Virtual Try-On GAN (VITON-GAN) is proposed to address the occlusion problem [5]. It consists of the GMM and the TOM with an additionally included adversarial loss in the TOM [3]. The proposed VITON-GAN model has been shown to have the ability to transfer the details of product clothes to the clothed person while generating a clear depiction of arms.

The performance of VITON-GAN model depends largely on the quality of dataset used to train the network. The balance of matching performance and training samples is also considered to be a crucial component of the judgement criteria of a model. Based on this, our work aims to investigate the relevance between the size of dataset and the matching effect of the generated results. The success of this finding will provide important information for choosing a suitable size of dataset to reach desired performance of the module in the future work. Our experiment evaluates the effect of sample size on the loss value of GMM module from 2,000 samples to 14,000 samples. The training epoch is also investigated with the loss value of GMM module. After finding the relationship between these three, the relevance between the loss value of GMM and the visual effect is explored.

Our experiment results find that as the training sample and training time increase, the loss value of GMM module decreases significantly in the beginning and then drops slowly. However, the matching effect of generated image only shows improvement in the beginning. As the sample size and training epoch become large enough, the small decrease of loss value does not make an equivalent improvement on the matching effect. Thus, the sample size and the matching effect are not strongly related and there could be other factors to be judged.

The rest of this paper is organized as follows. Section 2 gives an overview of the GMM. Section 3 shows the experimental results and discussions. Section 4 summarizes this paper. And in Section 5, the prospects and challenges are discussed.

2. Method

GMM is an end-to-end neural network which is trained with pixel-wise L1 loss. It can coordinate the input clothes with the human shape, transmitting and distorting the given shape of the clothes to match the pose of human body. The Figure 1 below presents the detailed process of trying on.
A Geometric Matching Module consists of 4 parts: 1) Two neural networks used for abstracting advanced features from the clothes \( c \) and person \( p \). 2) A correlation Matching that unites the two features in the last step to form correlation layer for single tensor, as the input for regression network. 3) A regression network that is used to predicate the changing parameter \( \theta \). 4) TPS Warping that transforms the image into output with the formula

\[
c' = T_\theta(c)
\]

During the end-to-end connection of the network, the loss function is

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

3. Experiment Results and Analysis
The experiment focuses on the GMM model. The experiment datasets are given by the original designer. In the datasets, the designer offers cloth, cloth-mask, person, person-parse, and pose images. More than 35,000 pictures have been used to train the GMM model. The epoch number is set as 50, because in our experiment process it’s found that 50 times of training could decrease the loss value into a very stable number, which is approximate 9%.
Then the relationship between loss value and visual effect is explored. We select several matched groups to show this relationship. In the first matched group (Figure 3), we select two pictures from 2000, 5000, 8000, 11000, 14000 samples groups. The upper one (blue) is generated by the model which only be trained for one time and the 50-times-trained model generate the below one (pink).

![Figure 3. Images generated by GMM trained 1 and 50 times of five samples groups](image)

Then it is considered that how the training time influences the final visual effect. So, three matched groups are selected to see that when the samples number is constant how the visual effect would change while the training time is increasing.

In the following matched groups, we select the same picture from 2000, 8000, 14000 samples groups which are generated by the model trained for 1, 10, 20, 30, 40, 50 times (Figure 4).

![Figure 4. Three matched groups. A denotes the 2000 examples group, B denotes the 8000 examples group and C denotes the 14000 examples group.](image)

From figure 4, the 2000 examples group’s visual effect is not so well. The clothes are out of shape even after being trained for 50 times. Surprisingly, the 8000 examples and the 14000 examples show similar results. Both of their results could fit the human body, although it is not perfect.

In combination with three pictures, it is concluded that the visual effect has positive relation with the loss value. A lower loss value model could generate a better virtual image. In figure 4, while the two models’ loss value are similar, their generated images are similar as well. In conclusion, in GMM model, lower loss valued model could create higher quality images. But with the number of examples increasing, the loss value could only bring slightly changes and the visual effect is recognizable. However, the loss value is still an appropriate factor to the visual effect of the pictures generated from the model.
4. Conclusion
While the visual try-on model is applied to generate the fitting effect with the pictures of a person and the clothes, this work analyzes the output pictures with respect to the visual effects and the improvement made by training times and sample size. To conclude, in the beginning period, both factors make contribution to the improvement in result. However, when sample size and training times are big enough, the improvement becomes less visible.

5. Discussion
From the result in the experiment, the loss value reduces while sample size and training times increase. However, it seems like the decrease in loss does not necessarily means improvement in the result image in the later period, but only for the beginning. When the training times and the sample size become large enough, dropping loss value only brings slight changes that are sometimes even unrecognizable in the visual effect. From this result, it is doubtful whether loss value is an appropriate factor to judge the effect of the model, or whether there could be some more suitable factors to judge. On the other hand, with the loss value decreasing, it seems difficult to find the lowest point where the effect is the best so that no more effort is needed anymore. It is also a common problem in deep learning field, that is, although methods are given to realize some functions, finding the best effect is hard. Sometimes only a wide range where the best effect could lie is given, and it consumes too much to find where it is exactly.

Reference
[1] Yen, Y. Y., Narayanasamy, K., Lin, C. Y., Rasiah, D., & Ramasamy, S. (2017) Consumer’s perception towards real-time virtual fitting system. In: Proceedings of the 6th International Conference on Computing and Informatics. pp. 311-316.
[2] Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018) Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543-7552.
[3] Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018) Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 589-604.
[4] Jae Lee, H., Lee, R., Kang, M., Cho, M., & Park, G. (2019) LA-VITON: a network for looking-attractive virtual try-on. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.
[5] Honda, S. VITON-GAN: Virtual Try-on Image Generator Trained with Adversarial Loss.
[6] Fincato, M., Landi, F., Cornia, M., Cesari, F., & Cucchiara, R. (2020) VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In: International Conference on Pattern Recognition.
[7] Jandial, S., Chopra, A., Ayush, K., Hemani, M., Krishnamurthy, B., & Halwai, A. (2020) Sievenet: A unified framework for robust image-based virtual try-on. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2182-2190.
[8] Goodfellow, I., Pouget-Abadie, J., Mirza M., Xu B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) Generative Adversarial Networks. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672-2680.
[9] Xu, Y. (2019) Research on Face Super Resolution Reconstruction Based on DBGAN. In: 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). pp. 807-8075.
[10] Brock, A., Donahue, J., & Simonyan, K. (2018) Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: International Conference on Learning Representations.