LA-VITON: A Network for Looking-Attractive Virtual Try-On

Hyug Jae Lee*, Rokkyu Lee*, Minseok Kang*, Myounghoon Cho, Gunhan Park
NHN Corp.
hyugjae.lee@nhn.com,rokkyu.lee@nhn.com,minseok.kang@nhn.com,mhcho77@nhn.com,gunhan.park@nhn.com
*makred authors contributed equally to this work

Abstract

In this paper, we propose an image-based virtual try-on network, LA-VITON, which allows the generation of high fidelity try-on images that preserves both the overall appearance and the characteristics of clothing items. The proposed network consists of two modules: Geometric Matching Module (GMM) and Try-On Module (TOM). To warp in-shop clothing item to the desired image of a person with high accuracy in GMM, grid interval consistency loss and an occlusion handling technique are proposed. Grid interval consistency loss regularizes transformation to prevent distortion of patterns in clothes and an occlusion handling technique encourages proper warping despite target bodies are covered by hair or arms. The following TOM synthesizes the final try-on image of the target person seamlessly with the warped clothes from GMM. Extensive experiments on fashion datasets show that the proposed method outperforms the state-of-the-art methods.

Figure 1. Results of the proposed LA-VITON. The try-on images are synthesized seamlessly and realistically with given in-shop clothing images and target person-images.

1. Introduction

As image-to-image networks and Generative Adversarial Networks (GANs) have become more prevalent, research on virtual try-on systems is also increasing.

In this paper, we propose an image-based virtual try-on network, LA-VITON, which shows significant improvement over the previous methods [3, 7] in producing attractive looking results without damages or distortions. LA-VITON consists of two components, a Geometric Matching Module (GMM) and a Try-On Module (TOM), similar to previous works [3, 7].

Since GAN-based methods [4, 9] show their limitations in generating fine and sharp images, adoption of an operation, such as GMM, is necessary to deliver information directly from the source to the target. Regularization in the form of a loss function and a scheme of occlusion handling are introduced for more accurate matching performance. The final help of a Spectral-Normalization GAN (SNGAN) discriminator [5] lets the GMM of LA-VITON overcome limitations of intensity based loss functions. The coarse layers of TOM composites the warped clothes from GMM and an intermediate person-image. The following refinement layers remove the artifacts and enhance the quality of the generated image from the coarse network.

Virtual try-on images synthesized through the proposed LA-VITON are presented in Figure 1. The final try-on images preserve the original details and characteristics of the clothing items without seams.

The main contributions of this paper are as follows: (1) regularizing transformation with grid interval consistency loss and (2) hiring occlusion-handling technique for robust geometric matching. Extensive experimental results show that our method surpasses the previous state-of-the-art methods by a wide margin.

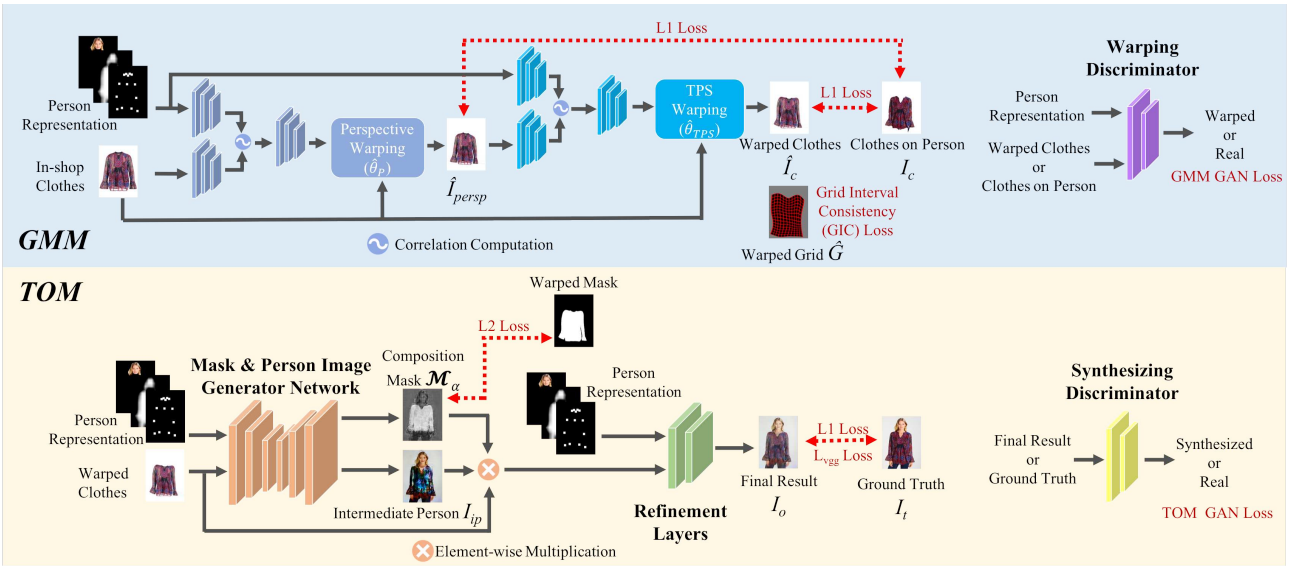

Figure 2. Architecture of the proposed LA-VITON.

2. Related works

Rocco et al. [6] proposed a convolutional neural network architecture for geometric matching which can be trained end-to-end. They mimicked the standard steps of feature extraction, matching and inlier detection within their architecture. Their architecture is composed of two stages, affine transformation and TPS transformation. In our work, we adopt the idea of two-stage transformation consisting of perspective transformation and TPS transformation.

Among the previous virtual try-on studies, the most closely related works are VITON [3] and CP-VTON [7]. As input, VITON uses a person representation, which is composed of the person's head, pose heatmap [1], and body-shape mask [2]. With the coarse-to-fine strategy using a two-stage architecture, the desired clothing item is seamlessly transferred onto the corresponding region of the target person-image. Meanwhile, CP-VTON proposed end-to-end trainable GMM that showed better warping performance for a desired clothing item.

3. Virtual Try-on Network

We want to put in-shop clothes on a target person in an image virtually without seams, degradation, or distortion of the prints and patterns. As illustrated in Figure 2, the final result image is synthesized with the given in-shop clothes and the person representation. The in-shop clothes are warped and roughly aligned through GMM. The proposed scheme to learn matching and warping of clothes result in non-degraded and non-distorted warped results. TOM generates the intermediate person-image and builds a mask for composition of the warped in-shop clothing image and the generated intermediate person-image.

3.1. Geometric Matching Module

One of main purposes of GMM is to align clothing while preserving the characteristics [3, 7]. We adopt end-to-end trainable geometric matching using the bottom-up strategy proposed by Rocco et al. [6]. A restriction on the matching grid, an occlusion handling scheme and GAN loss are employed to enhance the performance of GMM.

In GMM, the perspective and the TPS transformation are trained together in every step. The perspective transformation has an additional train path for back-propagating L1 loss $\mathcal{L}_{persp}$ between the warped clothes $\hat{I}_{pesrp}$ and the clothes on person I_c.

Grid Interval Consistency: TPS transformation generally shows good performance but its high flexibility often causes distortion of patterns and prints. Even when the coarse-to-fine strategy is applied, TPS with high degree of freedom causes undesirable warping results. Clothes are deformable objects, but deformation on a human torso is very limited. Therefore, geometric matching needs to be restricted to maintain the shape and appearance of clothing. We introduce grid interval consistency (GIC) loss which retains the characteristics of clothes after warping. GIC loss, $\mathcal{L}_{gic}$ in Equation 1, is based on the distance $DT(a, b)$ between neighbors, a and b. The absolute difference is used for the distance metric in this paper. $\hat{G}_x(x, y)$, $\hat{G}_y(x, y)$, H_G, and W_G are the x and y coordinates of the grid to be mapped and height and width of the grid, respectively.

In addition to patterns and prints, shapes are also preserved after warping by maintaining the consistency of the intervals. As shown in Figure 3, L1 loss only often does

Figure 3. The input grid and clothing images are warped stably with the proposed GIC loss, whereas a swirling pattern is presented in the warped grid without GIC loss.

Figure 4. Occluded clothes on person cause unreasonable transformation and distortion. With our occlusion handling method, the network conducts accurate geometric matching.

not work with mono-colored or repetitive pattern printed clothes. A swirling pattern is presented in the warped grid. However, the input grid is warped stably with GIC loss.

$$\mathcal{L}_{gic}(\hat{G}_x, \hat{G}_y) = \sum_{y}^{H_G} \sum_{x}^{W_G} \Big(\Big| DT(\hat{G}_x(x,y), \hat{G}_x(x+1,y)) - DT(\hat{G}_x(x,y), \hat{G}_x(x-1,y)) \Big| + \Big| DT(\hat{G}_y(x,y), \hat{G}_y(x,y+1)) - DT(\hat{G}_y(x,y), \hat{G}_y(x,y-1)) \Big| \Big) \tag{1}$$

Occlusion Handling: In real life environment, clothes on person are easily occluded by hair and arms. Since we use clothes region extracted using LIP (Look-Into-Person) [2] as ground truth, the network tries to fit input in-shop clothes into the extracted observable clothes region. As a result, the clothes are deformed unreasonably as in Figure 4. We address the problem by excluding the occluded regions from $\mathcal{L}_{warp}$ calculation. This solution encourages the network to be trained for more precise and reasonable transforming parameter estimation. Thus, the network conducts accurate geometric transformation regardless of occlusion by hair and arms.

We, further, employ the notion of GANs to improve the geometric matching performance. An inherent limitation of intensity difference based losses is that they have difficulty distinguishing similarly colored foregrounds and backgrounds.

The total loss function $\mathcal{L}_{gmm}$ for training GMM is defined as Equation 2 where $\mathcal{L}_{warp}$ and $\mathcal{L}_{ggan}$ are the pixel-wise L1 loss and GAN loss. λ_{warp}, λ_{gic}, and λ_{ggan} are weights for the corresponding loss functions.

$$\mathcal{L}_{gmm} = \lambda_{warp}\mathcal{L}_{warp} + \lambda_{gic}\mathcal{L}_{gic} + \lambda_{ggan}\mathcal{L}_{ggan} \tag{2}$$

The proposed geometric matching method delivers precisely aligned clothes to TOM with the characteristics preserved. It is crucial to guarantee realistic try-on results.

3.2. Try-on Module

The purpose of TOM is to synthesize the final try-on image for the input person-image. To preserve the original details of in-shop clothing items, the warped clothing should be exploited as much as possible. If the warped clothing is directly copied onto the target person-image, unnatural seams will appear. A seamless image is obtained by blending the warped clothing from GMM and the intermediate person-image with a composition mask.

Refinement layers following the generator further improve the quality of the blended image. The refinement layers employs dilated convolutions [8] to maintain a high resolution feature map, preserving the details of the image. To improve the quality of images generated with TOM, SNGAN [5] is employed for training.

4. Experiments

We use the paired dataset collected by Han et al. [3]. This dataset contains 14,221 of training and 2,032 of testing pairs.In Figure 5, the alignment results of the proposed GMM with those of the previous alignment methods are compared. The proposed method not only maintains characteristics of patterns and prints but also aligns in-shop clothes correctly.

For quantitative comparison study, 523 comparisons of CP-VTON and LA-VITON are presented to 10 unique workers, who are asked to pick the results that are more preferable and realistic. The results of the study are presented in Table 1. Try-on results synthesize by LA-VITON were chosen by 78.78% of the workers. In other words, the proposed method was the favorite.

In Figure 6, qualitative comparison with VITON and CP-VTON is presented. Most of the results from VITON and CP-VTON hardly seem to have the same patterns and prints for in-shop clothes. However, LA-VITON preserves the characteristics of in-shop clothes in the results.

Figure 5. Comparison of alignment performance. The proposed GMM is compared with SCMM and CP-VTON's.

Figure 6. Comparison of LA-VITON with VITON and CP-VTON. The proposed LA-VITON synthesizes undistorted and realistic image with details.

5. Conclusion

In this paper, we introduced an image based virtual try-on network, LA-VITON, which produces high fidelity output images while preserving the patterns and prints of the original clothing item. Using the proposed network, we could produce attractive results. Two stages of GMM, perspective transformation and TPS transformation, are combined and trained as a single network. The proposed GIC loss and occlusion-handling technique enables more accurate warping. We also implement a warping discriminator to compensate for the limits of intensity difference based loss functions.

The results of extensive experiments demonstrate that the proposed method outperforms the state-of-the-art methods by a wide margin.

Table 1. Comparison Study of Preference.

	CP-VTON	LA-VITON
Preference	21.22%	78.78%

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1302–1310, 2017. 2

[2] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. 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*, 2017. 2, 3

[3] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. VITON: An image-based virtual try-on network. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3

[4] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 1

[5] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 1, 3

[6] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[7] B. Wang, H. Zheng, X. Liang, Y. Chen, and L. Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision*, pages 589–604, 2018. 1, 2

[8] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2015. 3

[9] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017. 1