Virtual Try-On Considering Temporal Consistency for Videoconferencing

Daiki Shimizu[1] and Keiji Yanai[1]

The University of Electro-Communications, Tokyo, Japan
shimizu-d@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. Virtual fitting, in which a person's image is changed to an arbitrary clothing image, is expected to be applied to shopping sites and videoconferencing. In real-time virtual fitting, image-based methods using a knowledge distillation technique can generate high-quality fitting images by inputting only the image of arbitrary clothing and a person without requiring the additional data like pose information. However, there are few studies that perform fast virtual fitting from arbitrary clothing images stably with real person images for situations such as videoconferencing considering temporal consistency. Therefore, the purpose of this demo is to perform robust virtual fitting with temporal consistency for videoconferencing. First, we created a virtual fitting system and verified how effective the existing fast image fitting method is for webcam video. The results showed that the existing methods do not adjust the dataset and do not consider temporal consistency, and thus are unstable for input images similar to videoconferencing. Therefore, we propose to train a model that adjusts the dataset to be similar to a videoconference and to add temporal consistency loss. Qualitative evaluation of the proposed model confirms that the model exhibits less flicker than the baseline. Figure 1 shows an example usage of our try-on system which is running on Zoom.

Keywords: virtual try-on · image translation · temporal consistency · videoconferencing

1 Introduction

With the spread of social networking services, the use of filters to transform a person's appearance into a desired one has become a popular practice. In recent years, there has been a growing demand for real-time appearance transformation in videoconferencing as well, especially in virtual fitting for clothing transformation.

Conventional virtual fitting in videoconferencing is generally performed by positioning garments using poses and body meshes based on 3D data of the garment created by the developer. However, because garments are non-rigid, simple positioning often results in wrinkles and other discomfort, and is computationally expensive in order to improve quality. In addition, users without expertise

must choose from a limited set of clothes created by the developer, making it difficult for them to freely change clothes.

On the other hand, image-based virtual fitting methods using deep learning have been studied to convert a person image from an arbitrary clothing image into a fitting image. In the past, it was difficult to collect data because it required images of the person after the fitting. However, by removing the clothing regions of the person using a region decomposition model and using a person representation that is not limited by clothing, it is now possible to learn virtual fitting from only the clothing and the image data of the person wearing it without the try-on image. However, since the pose and region information of the person must also be inferred, the generation accuracy depends on the accuracy of these inferences, and the inference speed is reduced.

However, fast image-based virtual fitting methods [5,3] has emerged that does not require pose information and can generate fitting images from only the clothing and person images.

In recent years, there has been active research on video transformation, starting with Few-shot vid2vid [8]. Among them, several models [2,6,9,10] for virtual fitting of videos have appeared. However, even these models require pose information, making real-time generation difficult.

The purpose of this research is to apply image-based virtual fitting to perform stable virtual fitting from arbitrary clothing in actual video images, assuming videoconferencing. In this demo, we first created a virtual fitting system using a web camera, and verified how effective it is in actual person videos using existing fast image-based virtual fitting methods. We also improved the video virtual fitting dataset and proposed learning of a model with an additional loss of temporal consistency that reduces the difference from the previous frame, and qualitatively verified the results.

2 Related Work

2.1 Virtual Try-On

The development of image-based virtual fitting has been remarkable, and various methods have been devised. In recent years, many models [4,7,3] have emerged that use deep learning to generate fitting images from pose and segmentation in addition to person and clothing images.

VITON [4], in particular, consists of two stages, clothing deformation and synthesis, and has become a mainstream method today. These image-based virtual fitting models require pose and body region information in addition to clothing and person images, making them dependent on the accuracy of pose or segmentation and requiring additional inference time. Therefore, the fast image virtual fitting model WUTON, which uses knowledge distillation to infer directly from only the clothing and person images, and its improved version PF-AFN [3], which generates high quality images, have been introduced. In this demo, we adopt PF-AFN as a base method, which is a fast model and does not require guides such as poses.

PF-AFN [3] enabled direct clothing and person generation without the need for person masks by using the generated results of a parser-based teacher model with person pose and segmentation as input for the student model. These innovations have made it possible to perform virtual fitting generation at high speed and high quality. In this demo, we adopt PF-AFN as baseline and perform learning that considers temporal consistency.

In recent years, research on video generation has developed with the advent of Few-shot vid2vid [8], which enables conversion of faces, poses, etc. in videos. Starting with this, a video-based virtual fitting method using Optical Flow as a guide was proposed, and others [6,9,10] have appeared. FW-GAN [2] was the pioneering work in video-based virtual fitting and also provided the only video virtual fitting dataset, VVT. the network consists of a module that takes as input the previous frame's generated image and pose sequence, the previous generated frame, the person image, and the clothing image, and generates a deformation network for the clothing image and a flow for warping the previous generated frames. ShineOn [6] added DensePose as an input and introduced self-attention calculated from DensePose and person images. They also investigated the activation function bottom-up. However, this method also requires guides such as Optical Flow and DensePose, making real-time inference difficult.

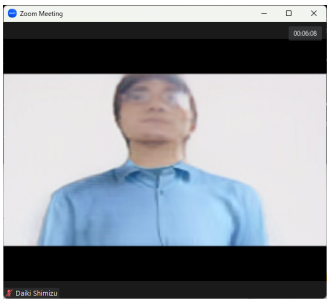

Fig. 1. Our try-on system running on Zoom

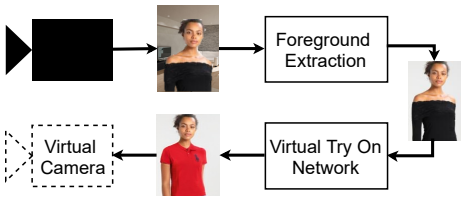

Fig. 2. Virtual Try-On System

3 Proposed Method

3.1 Virtual fitting system using a webcam

We developed a real-time virtual try-on system using a webcam and trained a virtual try-on model on the VITON [4] dataset. The virtual try-on system is shown in Fig.2. First, in order to apply the virtual try-on model to videoconferencing, it is necessary to remove the background from the webcam video. The foreground is extracted using the pre-trained segmentation model input to the virtual try-on model. The resulting fitting image is input to a virtual camera,

and the user selects the virtual camera to display the fitting image in the actual videoconferencing application.

In our investigation, we observed flickering in PF-AFN, as shown in Fig.4. To solve this problem, we propose a learning method and Temporal PF-AFN. We also use a video-based virtual fitting dataset, VVT [2], as the training dataset. By using a video dataset, we can learn frames that are close in time continuously, and can generate a stable PF-AFN even when there is a slight moving. The model is called Temporal PF-AFN, in which the composite result of the previous frame is added to PF-AFN as an input.

3.2 Learning considering temporal consistency

Our network structure flowchart is shown in Fig.3. Our model takes as input a person image with the background removed, a fitting image generated in the previous frame, and a reference clothing image. The output is a deformed clothing image, its mask, and an image containing the person and background. During training, a loss function is added to the existing loss function so that the results generated in the previous frame and the current frame are close. Temporal try-on loss function $\mathcal{L}_t$ is as follows:

$$\mathcal{L}_t(S_t, S_{t-1}) = \lambda_t(\lambda_{p_1}\mathcal{L}_p(C_t, C_{t-1}) + \lambda_c\mathcal{L}_{L1}(C_t, C_{t-1}) + \lambda_{p_2}\mathcal{L}_p(I_t, I_{t-1}) + \lambda_i\mathcal{L}_{L1}(I_t, I_{t-1}) + \lambda_m\mathcal{L}_{L1}(M_t, M_{t-1})) \quad (1)$$

where C, I, and M represent the generated clothing image, the try-on image, and the mask, respectively, and t represents the frame number. λs are hyperparameters to weight the loss functions. The total loss consists of two loss functions. The first is the VGG Loss, which is visually close, and the second is the L1 Loss, which is pixel close. Each loss is calculated between the two frames of the generated clothing image and the final fitting image, and then added together. The mask M of the generated clothing image is also calculated using L1 Loss between the two frames and then added together.

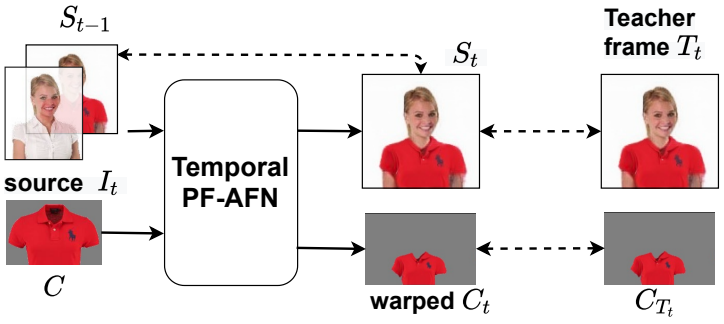

Fig. 3. Temporal PF-AFN

4 Experiments

4.1 Implementation Details

First, we describe the implementation of a virtual fitting system in videoconferencing. To apply the system, the background must be excluded from the webcam video. The pre-trained segmentation model is extracted from the foreground using DeepLab V3+ [1] and input to the virtual fitting model. Next, we describes the training method with Temporal PF-AFN. We used VVT as a training dataset. The first 5 epochs were trained to generate the initial frame. A single image and a concatenated zero-filled image were used as input. After that, we trained 35 epochs by adding temporal losses to the input frames and previously generated frames. We employed Adam optimizer for training. The learning rate was set to 5×10^{-5}, $\beta_1 = 0.5$, $\beta_2 = 0.999$. The parameters of the proposed method were trained as $\lambda_{p_1} = 0.2$, $\lambda_i = 1$, $\lambda_{p_2} = 0.2$, $\lambda_j = 1$, $\lambda_m = 2$. λ_t was trained as 0.01 in the Warping Module, and 1.0 in the Generation Module of PF-AFN. Other parameters were set following to PF-AFN.

4.2 Experimental Results

In the upper two rows of Fig.4, the method is better at generating rotations around 8 sec, and there is less flickering. However, there is a quick sitting motion between 14 and 16 seconds, which is not generated well by the proposed method because the pixel remains. In addition, because the initial frame was not generated well, the image was not corrected immediately and slowly changed to the correct fitting image.In the movie in the lower row, both methods were able to generate images between 13 and 19 seconds, but the baseline method caused flickering.The results of each movie showed that the scene was generated more smoothly with less flickering compared to PF-AFN, which processed each frame.

Fig. 4. Comparison between Temporal PF-AFN and PF-AFN. The center two columns represent Temporal PF-AFN results and the right two columns of Temporal PF-AFN results. (Click to play)

5 Conclusion

In this demo, we demonstrate a virtual fitting system for videoconferencing employing an improved PF-AFN. From the experimental results, PF-AFN showed an uncomfortable flickering. Therefore, we proposed a temporal PF-AFN that considers temporal consistency in order to suppress the flickering. Qualitative evaluation confirmed that the fluctuation in the frontal plane could be suppressed. However, it was found that there is a problem that pixels remain in quick movements. For the future, we plan to further improve the stability of the generation process. The current method uses a foreground extractor to remove the background, which is dependent on the accuracy of foreground extraction and reduces the overall processing speed. We will make it possible to generate images of people with actual backgrounds in real time by providing them into the model as an input without a foreground extractor.

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. of European Conference on Computer Vision (2018)
2. Dong, H., Liang, X., Shen, X., Wu, B., Chen, B.C., Yin, J.: FW-GAN: Flow-navigated warping gan for video virtual try-on. In: Proc. of IEEE International Conference on Computer Vision (2019)
3. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 8485–8493 (2021)
4. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: An image-based virtual try-on network. In: Proc. of IEEE Computer Vision and Pattern Recognition (2018)
5. Issenhuth, T., Mary, J., Calauzènes, C.: End-to-end learning of geometric deformations of feature maps for virtual try-on. arXiv:1906.01347 (2019)
6. Kuppa, G., Jong, A., Liu, V., Liu, Z., Moh, T.: ShineOn: Illuminating design choices for practical video-based virtual clothing try-on. In: Proc. of IEEE Winter Conference on Applications of Computer Vision Workshops. pp. 191–200 (2021)
7. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: CP-VTON+: Clothing shape and texture preserving image-based virtual try-on. In: Proc. of IEEE Computer Vision and Pattern Recognition (2020)
8. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Proc. of Advances in Neural Information Processing Systems (2019)
9. Wei, D., Xu, X., Shen, H., Huang, K.: C2F-FWN: Coarse-to-fine flow warping network for spatial-temporal consistent motion transfer. Proc. of the AAAI Conference on Artificial Intelligence **35**(4), 2852–2860 (May 2021)
10. Zhong, X., Wu, Z., Tan, T., Lin, G., Wu, Q.: MV-TON: Memory-based video virtual try-on network. In: Proc. of ACM International Conference Multimedia. p. 908–916 (2021)