FW-GAN: Flow-navigated Warping GAN for Video Virtual Try-on

Haoye Dong[1,2,*], Xiaodan Liang[3,*], Xiaohui Shen[4], Bowen Wu[1], Bing-Cheng Chen[1], Jian Yin[1,2,†]
[1]School of Data and Computer Science, Sun Yat-sen University
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
[3]School of Intelligent Systems Engineering, Sun Yat-sen University
[4]ByteDance AI Lab.
{donghy7@mail2,issjyin@mail,wubw6@mail2,chenbch9@mail2}.sysu.edu.cn
xdliang328@gmail.com, shenxiaohui@bytedance.com

Abstract

Beyond current image-based virtual try-on systems that have attracted increasing attention, we move a step forward to developing a video virtual try-on system that precisely transfers clothes onto the person and generates visually realistic videos conditioned on arbitrary poses. Besides the challenges in image-based virtual try-on (e.g., clothes fidelity, image synthesis), video virtual try-on further requires spatiotemporal consistency. Directly adopting existing image-based approaches often fails to generate coherent video with natural and realistic textures. In this work, we propose Flow-navigated Warping Generative Adversarial Network (FW-GAN), a novel framework that learns to synthesize the video of virtual try-on based on a person image, the desired clothes image, and a series of target poses. FW-GAN aims to synthesize the coherent and natural video while manipulating the pose and clothes. It consists of: (i) a flow-guided fusion module that warps the past frames to assist synthesis, which is also adopted in the discriminator to help enhance the coherence and quality of the synthesized video; (ii) a warping net that is designed to warp clothes image for the refinement of clothes textures; (iii) a parsing constraint loss that alleviates the problem caused by the misalignment of segmentation maps from images with different poses and various clothes. Experiments on our newly collected dataset show that FW-GAN can synthesize high-quality video of virtual try-on and significantly outperforms other methods both qualitatively and quantitatively.

1. Introduction

The emergence of image synthesis technique significantly advances the progress of the virtual try-on systems [16, 36], which are of great value to lots of applications, e.g., online shopping, movie making, and video editing. However, most of the try-on methods are based on single images, while the video-based virtual try-on problem has been largely unexplored. In this work, we make a first attempt to address this problem. Specifically, given a person image, the desired clothes, and a series of target poses, we synthesize a realistic-looking video that preserves the distinct appearance from both the person and clothes image. Some of the results are illustrated in Figure 1, showing that the proposed approach can generate high-quality virtual try-on videos with convincing details.

Most existing methods use encoder-decoder-like neural networks [16, 36] to synthesize virtual try-on images. They mainly focus on synthesizing the person image by replacing with other clothes, conditioned on a fixed pose, and thus fail to generate realistic videos due to the lack of ability of manipulating arbitrary poses and different clothes when virtual try-on is conducted in unconstrained scene. Besides 2D image synthesis, various 3D modeling techniques [22, 27, 30, 42] have been developed for virtual try-on. However, those methods focus on single images as well, and have not been extended to video generation. Moreover, it requires huge labor cost to collect the 3D annotation and massive computation to build the 3D model, which limits the performance of virtual try-on in the practical scenario.

Particularly, in a video sequence, person or clothes images often contain various visual appearance, viewpoints, and arbitrary human layouts due to different poses. It is impractical for current convolution-based generators to exploit entangled information without the aid of any external structured knowledge. Besides, different poses for a whole human body may result in heavy occlusions or dramatic appearance changes for some body parts. Furthermore, spatiotemporal consistency is critical to the visual quality of the synthesized video, which is not considered in the existing image-based synthesis methods.

*equal contribution
†Corresponding author is Jian Yin

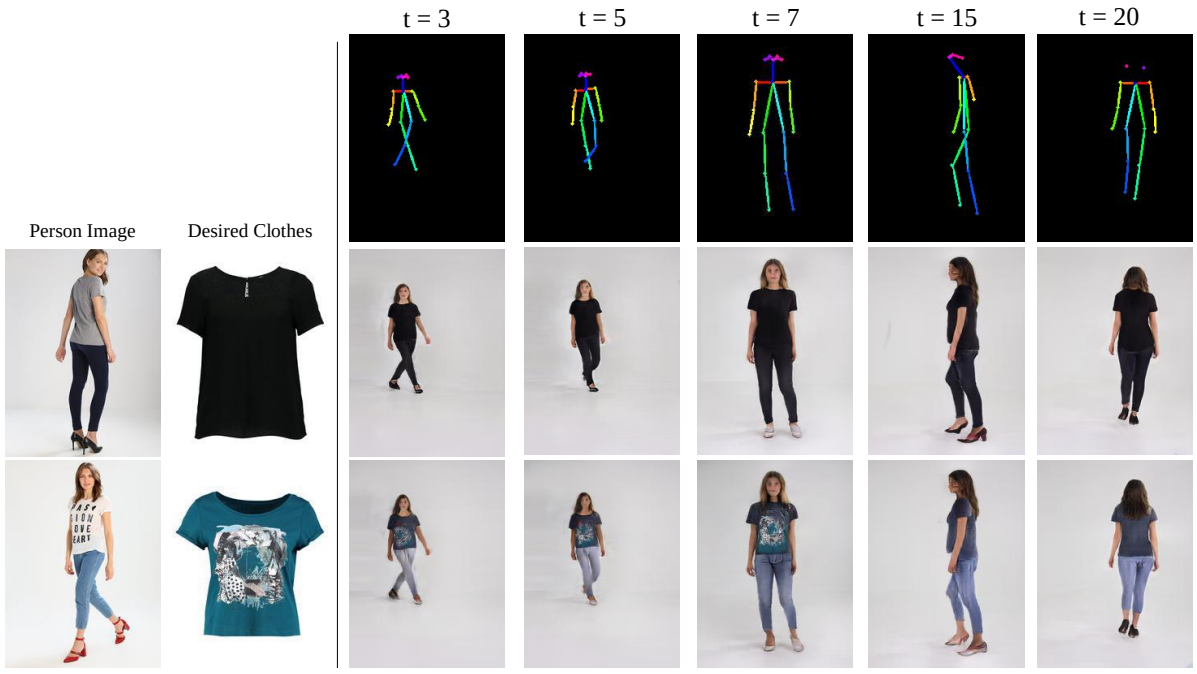

Figure 1. **Some results of our method.** Given a person image, the desired clothes, and the series of the target poses, our FW-GAN learns to automatically fit the desired clothes onto the person, restructure the pose of the person, and output the realistic video. Input images in the first column, the poses in the first row, the results of virtual try-on for each pose in the other columns.

To address the above mentioned challenges, we propose an FW-GAN to achieve the controllable video synthesis for virtual try-on, by manipulating both different poses and various clothes. FW-GAN consists of three main components: 1) a flow-navigated module that enforces the synthesized video to be spatiotemporal coherent and high-quality visual; 2) a warping net adapted to estimate the grid of transformation parameters that warps the desired clothes in order to fit the corresponding region of the person image; 3) a human parsing constraint loss that constrains body layouts to enforce consistency from a global view. In particular, the optical flow [3] plays a critical role in the proposed FW-GAN for making the generated videos coherent, which warps the pixel of the preceding frames to the new frames, and is also used as the conditioned input of the flow-embedding discriminator, resulting in more photo-realistic frames and spatiotemporal smoothing videos. Besides, to preserve the details of the desired clothes, a weight mask is leveraged to adaptively select the pixel values from the warped desired clothes or synthesized clothes.

We conduct extensive experiments on our newly collected dataset, including quantitative comparison, ablation study, and human perceptual study on the Amazon Mechanical Turk platform. The proposed FW-GAN substantially outperforms all existing methods on synthesizing virtual try-on video with arbitrary poses both qualitatively and quantitatively. The main contributions of our work include:

- To generate high-quality synthesized video of virtual try-on under a sequence of poses, a person image, and the desired clothes, we propose an FW-GAN to incorporate the optical flow with warping net for warping the frames and clothes images, respectively, which can preserve the details in global and local views.
- A flow-embedding discriminator is proposed that incorporate an effective flow input to the discriminator to improve the spatiotemporal smoothing.
- We employ a parsing constraint loss function as one form of structural constraints to explicitly encourage the model to synthesize results under difference poses and various clothes to produce coherent part configurations with the input image.

2. Related Work

Image synthesis. Generative adversarial networks (GANs) [12] has recently achieved impressive results on image synthesis. To capture the image distribution, GANs is capable of generating fake images which are indistinguishable from the real images. Conditional generative adversarial networks (cGANs) [26] can generate samples with desired attributes by appending condition on the inputs of both the generator and discriminator, and showed promising results on image-to-image translation[20, 15, 14, 9, 10, 8]. For person image generation, Lassner *et al*. [23] proposed

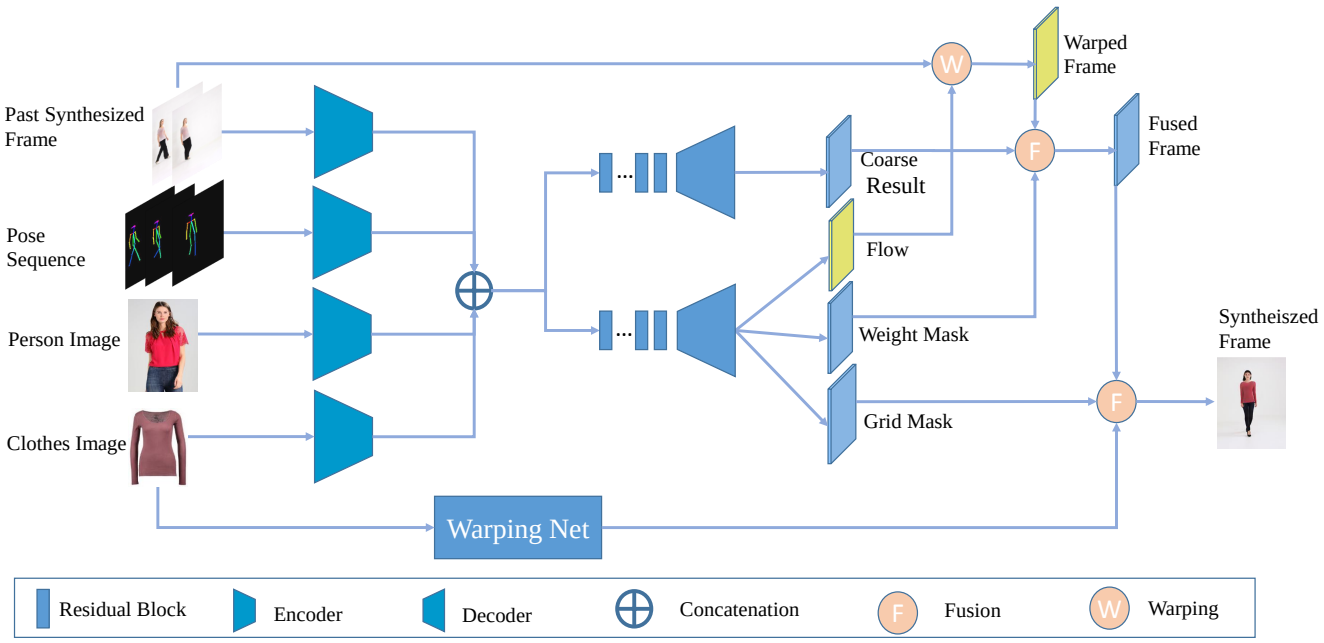

Figure 2. **The framework of the proposed FW-GAN.** FW-GAN consists of four encoder and two decoder with residual blocks. FW-GAN first to predict the flow, then warps the last past synthesized frame. We use weight mask and grid mask to polish the results.

a generative model of people in clothing for the full body. They first learned to generate human parsing maps and then learned a model to translate the resulting segments to realistic images, but the fashion attributes are not controllable in this method. Zhao *et al*. [43] proposed an image generation model to generate multi-view cloth images from only a single view input. [25, 10, 33, 8] synthesized person images conditioned on arbitrary poses.

Virtual try-on. Most previous works on virtual try-on were based on computer graphics. Guan *et al*. [13] designed a framework for synthesizing clothes on 3D bodies, with ignoring the shape and pose. Anna *et al*. [1] proposed a method for dynamically tracking and retexturing cloth for real-time visualization in a virtual mirror environment. Sekine *et al*. [30] developed a virtual fitting method for adjusting 2D clothing images to users by modeling their 3D body shapes from single images. Pons-Moll *et al*. [27] addressed the problem of capturing multiple garments on fully dressed people in motion by using a multi-part 3D model of clothed bodies. Yang *et al*. [41] proposed an approach for computing a realistic 3D model of a human body from a single photograph. There are also a few works based on image-based generative models which aim to synthesize perceptually correct images from real 2D images. Jetchev and Bergmann [21] introduced a conditional analogy GAN to swap fashion items. However, during inference, they needed the pair images of the original item on the person and the target item, which might not be easy to acquire. VITON [16] used a coarse-to-fine framework to replace the original fashion item on the person with the desired item, and enhance the fidelity of the synthesized image with a refinement network. [36, 9] addressed a similar problem, but it also aimed to preserve clothing characteristic by learning a thin-plate spline transformation with a geometric matching module.

Video synthesis. Extensive studies have been conducted on video synthesis. Video inpainting [40], video matting and blending [2, 7] and video super-resolution [31, 32] were proposed for addressing specific problems. Chan *et al*. [6] proposed a method for transferring dance movement from a source video of a person dancing into a target if acquiring a video lasting for a few minutes in which the target subject performs standard moves. Their method was based on pix2pixHD [38] and a state of the art pose detector OpenPose [4, 34, 39]. vid2vid [37] addressed the problem of video-to-video synthesis based on GANs coupled with a spatiotemporal adversarial objective. The video technique has huge application potential, but the virtual try-on for generating video is less explored.

3. FW-GAN

3.1. Problem Formulation

Given a pose sequence, a person image, and a clothes image, we aim to generate a photo-realistic video in which the person wears the desired clothes, and the person's movement is the same as the pose sequence. In order to generate the correspondent result with arbitrary inputs, we can train a generative model with a training dataset. Formally, let I_p,

I_c, and P_i represent the person image input, the clothes image input, and the i-th frame of pose sequence respectively. And we denote the pose sequence input by $S = \{P_i\}_{i=1}^{N}$ and the video output by $V = \{R_i\}_{i=1}^{N}$ where N is the frame number of the pose sequence and R_i is the i-th frame of the output. Our goal is to learn the mapping $(I_p, I_c, S) \rightarrow V$. Our training dataset is $\{V_t^i, I_c^i, I_p^i\}_{i=1}^{n}$, where V_t^i, I_c^i, I_p^i are the i-th training video, clothing image input, and person image input respectively, and n is the number of samples.

3.2. Pose Embedding

A pose of a person in an image is composed of 2D skeletons with M joints $P = (\mathbf{l}_1, ..., \mathbf{l}_M)$, where $\mathbf{l}_i = (x_i, y_i)$ is the coordinate of the i-th joint in the image. As interpreted in [28], the coordinate $\mathbf{l}_i$ can be regarded as a random variable and has a probability density map $\mathbf{p}_i$ formed by:

$$\mathbf{p}_i[x, y] = P(\mathbf{l}_i = (x, y)) \;\; \forall (x, y) \in \mathcal{U} \tag{1}$$

where $\mathcal{U}$ is the coordinate space of the input image. Then the pose P is equivalent to a concatenation of all probability density maps $\mathbf{p} = (\mathbf{p}_1, ..., \mathbf{p}_M)$.

3.3. Network Architecture

3.3.1 Generator

We propose a residual-like generator to incorporate the optical flow with warping net for exploiting temporal information, a personal appearance and clothes information simultaneously. Formally, our generator is based on a conditional GAN framework which aims to capture the conditional probability distribution. We denote the generator by G. Let I_p represents the random variable of the person image input. I_c is the random variable of the clothes image input, and $S = \{P_i\}_{i=1}^{N}$ is the pose sequence. Then we have the pose embedding $\mathbf{p}$ of the pose sequence S. Let $V^{'} = \{R_i\}_{i=1}^{N}$ represents the output of G. Moreover, let V represents the ground truth of the video. The generator G is equivalent to a conditional distribution so that we can compute the probability of $V^{'}$ with $G(V^{'}|\mathbf{p}, I_c, I_p)$. We optimize G by solving the standard minimax optimization problem. Formally, the objective function is defined by:

$$\begin{aligned} \min_G \max_D \mathcal{L}_{\text{gan}} &= \mathbb{E}_{V \sim p_{\text{data}}(V)}[\log D(V)] \\ &+ \mathbb{E}_{\mathbf{J} \sim p(\mathbf{J})}[\log(1 - D(G(V^{'}|\mathbf{p}, I_c, I_p)], \end{aligned} \tag{2}$$

where $\mathbf{J} = (\mathbf{p}, I_c, I_p)$ and D is the discriminator.

As shown in Figure 2, in the generator, every input has a correspondent encoder to extract feature maps. Then we concatenate and input these feature maps into two separate networks which are both composed of several residual blocks. The outputs of residual networks are fed to decoders which will generate optical flow and photo-realistic images.

3.3.2 Discriminator

Several works [38, 20, 24] show that using multiple discriminators could lighten the model collapse problem in GAN training. At the meantime, our task requires both visual quality of each frame and temporal consistency. Based on the above observation, we design two discriminators: frame discriminator, and flow-embedding discriminator.

Frame Discriminator is responsible for the visual quality of each frame. In other words, it ensures that each generated frame looks like real video frames. Frame discriminator takes four inputs, pose sequence $S = \{P_i\}_{i=1}^{N}$, person appearance image I_p, cloth image I_c, generated frame v. Tuple (S, I_p, I_c) could be thought of conditional input of frame discriminator. This discriminator should output 1 for a true pair $((S, I_p, I_c), v)$ and 0 for fake pair $((S, I_p, I_c), \widetilde{v})$.

Flow-embedding Discriminator is responsible for temporal consistency between neighboring frames. We think consecutive generated frames should have temporal dynamics of consecutive real frames with the same optical flow. Just like frame discriminator, flow-embedding discriminator also takes conditional input, optical flow. We denote O as $K - 1$ optical flow for the K consecutive frames. This discriminator should output 1 for a true pair(O, v) and 0 for fake pair $(O, \widetilde{v})$. During experiments, we find those discriminators well on video try-on. It makes the person and the clothing move more smoothly on the generated video.

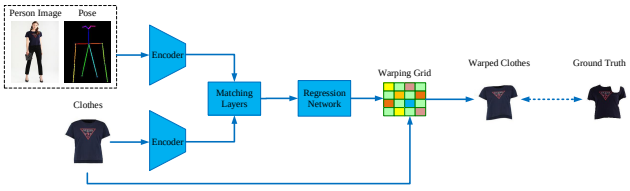

Figure 3. **The framework of Warping Net.** We first input the person image, target pose, and desired clothes into the encoder to extract the feature maps, respectively. Then, the matching layers are trained to compute the relation among of the feature maps. Followed by the matching layers is a regression network that outputs the warping grid of transformation mappings. Finally, we use this warping grid to warp the desired clothes.

3.3.3 Warping Net

As shown in Figure 3, the Warping Net consists of two encoders, matching layers, and a regression network. Let Ck denote a Convolution layer with kernel size of 4, a stride 2, and k filters. Let Rk dnotes a Convolution layer with kernel size of 3, a stride 1, and k filters, followed by BatchNorm2d Normalization and ReLU activation function. Let Lk denotes a Linear function output k dimension. For the Matching layers, we directly use the correlation map computation from the GEOCNN [29]. Therefore, encoder contains: C64, C128, C256, C512, R512, R512. Regression network consists of: C512, C256, C128, C64, L32.

3.4. Learning Objective Functions

In this paper, the objective function of FW-GAN is a weighted sum of several different losses. We will introduce them in details in the following sections.

Perceptual Loss. To obtain the high level and various features, we extract two different features from pre-trained VGG network and discriminators of our adversarial network, following [38, 16]. Then, we combine them to denote perceptual loss of this work.

$$\begin{aligned}\mathcal{L}_{\text{perceptual}} = &\sum_{i=0}^{N} \lambda_i \|\phi_i(\hat{I}) - \phi_i(Y)\|_1 \\ &+ \sum_{k=0}^{M}\sum_{j=0}^{M} \lambda_k \lambda_j \|\varphi_{(k,j)}(\hat{I}) - \varphi_{(k,j)}(Y)\|_1,\end{aligned} \tag{3}$$

where $\phi_i(\hat{I})$ describe the i-th feature map of the synthesized image $\hat{I}$ within VGG network, while λ_i controls the weight of them. Similarly, $\varphi_{(k,j)}(\hat{I})$ is the j-th layer feature map in the k-th discriminator of the synthesized image $\hat{I}$, while λ_j denotes the weight of j-th layer and λ_k describe the weight of k-th discriminator. N denotes the number of VGG layers, we set $N = 5$. M denotes the number of discriminator's layers, we set $M = 3$.

Parsing Constraint Loss. However, the above objectives do not consider the local information from sub-parts. To further improve the quality of the generated image, we propose a novel parsing consistent loss to make the part configuration of the generated image and those of ground truth coherent. Let ψ is a human parser. We require the parsing results of the synthesized image and the ground truth image should be the same. In this paper, we apply a light network [20] to train human parser. Especially, we denote the parsing result of the ground truth image as $F = \psi(Y) \in \mathbb{R}^{n\times n\times c}$, where n is the height/width of the image and c is the number of the semantic labels. The output for the synthesized image is defined as $P = \psi(\hat{I})$. For each pixel, the parsing results should be the same, e.g., the predicted parsing labels $F(h, w) \in \mathbb{R}^c$ for the pixel index (h, w) is equal to the $P(h, w)$. Since the softmax loss is a widely used method in deep CNNs that quantifies the dissimilarity between the two probabilities. Thus, we define the parsing consistent loss as

$$\mathcal{L}_{\text{pcl}} = -\sum_{w=0}^{W}\sum_{h=0}^{H}\sum_{l=0}^{C} F(h, w, l) \log P(h, w, l), \tag{4}$$

where the C denotes the number of parsing labels, H denotes the width of image, W denotes the width of image.

3.4.1 Overall Objective Function

Besides, we directly adopt a flow loss as $\mathcal{L}_{\text{flow}}$ from FlowNet [11]. We take the L1 loss from the pix2pix [19] as our grid loss $\mathcal{L}_{\text{grid}}$ to constrain the generator to learn more pixel from the warped clothes. Let $\mathcal{L}_{\text{gan}}$ denotes the loss of generator in this paper. In summary, FW-GAN objective describes a weighted sum of all the losses as the Eq. (5) shown.

$$\begin{aligned}\mathcal{L}_{\text{syn}} = &\alpha_1 \mathcal{L}_{\text{gan}} + \alpha_2 \mathcal{L}_{\text{perceptual}} + \alpha_3 \mathcal{L}_{\text{pcl}} \\ &+ \alpha_4 \mathcal{L}_{\text{flow}} + \alpha_5 \mathcal{L}_{\text{grid}},\end{aligned} \tag{5}$$

where hyper-parameters $\alpha_i, (i = 1, 2, 3, 4, 5)$ control the weight of each loss.

4. Experiments

In this section, we first introduce the implementation details of the proposed FW-GAN. Then we describe the evaluation metrics for evaluating the quality of the generated video. Next, we introduce the baseline method and our collected dataset. Finally, we make a visual comparison with the method of baseline and ablation study and analyze the quantitative and qualitative results.

4.1. Implementation Details

In training, the generator and the discriminators are updated alternatively with a mini-batch size of 4 through the stochastic gradient solver, i.e., Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$). We alternate between 1 steps of optimizing the generator and 1 step of optimizing the discriminators. The initial learning rate is 0.0002. The implementations are based on the Pytorch platform on four Titan XP GPU. After 30 epochs, high-quality results can be obtained. During testing, only the generator is activated, and it takes about 50ms for generating one image.

4.2. Dataset

We constructed a new video dataset appropriate to Video Virtual Try-on, named **VVT**. We first collected 791 videos of fashion model catwalk, which backgrounds are mostly white colour, ensuring us to focus on the task of virtual try-on and providing convincing evaluations for our models. Moreover, then we removed the noisy frames without pose results or parsing results. The frame number of each video mainly lies in the range between 250 and 300. We split the videos into a training set and a testing set with 661 videos and 130 videos respectively. The total frame numbers of the training set and the testing set are 159170 and 30931 respectively. We also crawled 791 person images and 791 clothes images and made every video associated with a person image and a clothes image. We also ensured that every person image is different from the person in the associated video and every clothes image is different from the clothes in the associated person image. Therefore, a sample in the dataset is composed of a video, a person image and a clothes image.

Figure 4. **Visual compare with the baseline method and the ablation methods on the VVT dataset.** First three columns start from left are inputs to our task. They are person image, desired clothes and target pose respectively. The last three columns are generated frame from different methods. The images of the last column are generated from our proposed algorithm. It looks better than the other two algorithms.

4.3. Evaluation metrics

Fréchet Inception Distance(FID) [18] is a metric for evaluating image synthesis quality. It uses the inception model [35] as a feature extractor after removing the last few layers of the network, and extracts feature vectors from real images and synthesized images respectively. Then it computes the mean μ and covariance matrix Σ for the feature vectors from the real images. It also computes the same statistics $\widetilde{\mu}$ and $\widetilde{\Sigma}$ for the feature vectors from the synthesized images. Then the FID is calculated as $\|\mu - \widetilde{\mu}\|^2 + \mathrm{Tr}(\Sigma + \widetilde{\Sigma} - 2\sqrt{\Sigma\widetilde{\Sigma}})$. Because this paper focuses on the video synthesis problem, we deploy a variant of FID following vid2vid [37], which is more suitable for evaluating video synthesis quality than the original FID. We use I3D [5] and 3D-ResNeXt-101 [17] as our pre-trained video recognition CNNs. In detail, we take 10 frames as a video clip, and exploit the output of the last average pooling layer in the network as our feature vector.

4.4. Baselines

CP-VTON [36] stands for Characteristic-Preserving Virtual Try-On Network proposed by Wang *et al*. [36]. Compared with VITON [16], they mainly deal with key characteristics of clothes. It is obvious that CP-VTON [36] indeed generate cloth with much more key characteristics. On our experiments, we retrain CP-VTON and VITON [16] on the VVT dataset. When testing, we adapt them to our task which means we input pose heatmap of each frame rather than fixed pose heatmap. During the experiment, we find that it generates almost the same image no matter what pose heatmap we input. Then, we take a glance at the dataset used for training CP-VTON and VITON and found that most images of that dataset are in almost the same pose.

4.5. Qualitative Results

Figure 6 and Figure 4 show some qualitative results on VVT dataset. The results show that the flow module and grid module play an important role in synthesizing realistic-look video. Without the grid, module leads to synthesize blurred and low-resolution video, and the pattern on the clothing is lost. Without flow-embedding discriminator network (w/o) fails to obtain spatiotemporal smoothing. Figure 6 demonstrates that given a person image, a clothes image, and a target pose image, FW-GAN is capable of synthesizing our desired image result in which the desired person

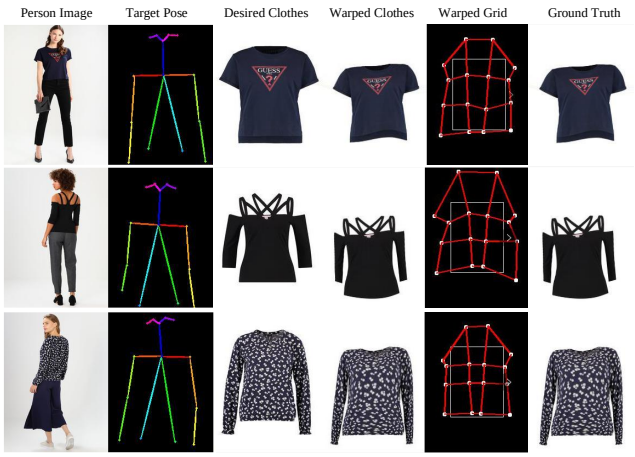

Figure 5. Some results of the Warping Net, which are shown in the 4th column. The warped grid are in the 5th column. The Warping Net predict the transformation mapping parameters to warp the clothes which at a similar level of realism as the ground truth.

Fréchet Inception Distance	I3D	ResNeXt-101
CP-VTON [36]	32.35	159.50
VITON [16]	30.05	129.74
FW-GAN (w/o grid + flow + parsing)	**6.57**	14.01
FW-GAN (w/o grid + flow)	7.37	17.47
FW-GAN (w/o grid + parsing)	7.47	15.88
FW-GAN (w/o grid)	7.04	15.31
FW-GAN (w/o parsing)	7.30	19.34
FW-GAN (w/o DT)	7.45	20.78
FW-GAN (w/o flow)	6.98	**13.17**
FW-GAN (Ours)	7.052	23.94

Table 1. Comparison with previous methods on the VVT dataset.

is wearing the desired clothes with the desired pose. Figure 7 shows some failure results of our method caused by uncommon styles of clothing. Some results of Warping Net are shown in the Figure 5. We can observe that the proposed warping net can achieve promising performance.

4.6. Quantitative results

We used our learned models and the baseline to synthesize 3000 video clips in the validation set. Every video clip was composed of 10 continuous frames. Then we deployed I3D and 3D-ResNeXt-101 to extract spatial-temporal feature vectors from the synthesized video clips and the real video clips and computed the FID based on these feature vectors. Table 1 reports the FID of our approach and the baseline, demonstrating that our method significantly outperforms the baseline. It also shows the detailed ablation studies conducted on our model. Although the ablation results in Table 1 do not demonstrate remarkable improvement, we think that this is because FID uses deep convolution layers to extract feature maps and will lose some information important for evaluating video synthesis quality. As shown in Table 1, the FID scores of VITON [16] in the last row that indicates the proposed FA-GAN can generate more spatio-temporal smoothing videos, compared with other methods. The lower number indicates the better performance. In particular, w/o flow denotes FW-GAN without optical flow. w/o parsing denotes FW-GAN without the parsing constrain loss. w/o grid deonotes FW-GAN without warping network. w/o DT denotes FW-GAN without the flow-embedding discriminator. w/o (grid + flow + parsing) denotes FW-GAN without warping network, optical flow, and the parsing contraint loss. w/o (grid + flow) denotes FW-GAN without warping network, and optical flow. w/o (grid + parsing) denotes FW-GAN without warping network, and the parsing contraint loss.

5. Human Perceptual Study

To achieve the fair visual comparison, we deploy a user study in the Amazon Mechanical Turk (AMT) platform. AMT is a platform that operates a marketplace for work that requires human intelligence. We carefully design a subjective A/B test similar to Wang *et al*. [37]. Different from them, we let the image GIFs represents the video. We show the images for workers that contain person image, the desired clothes image, and the target pose GIFs, followed by two shuffled options GIFs. All the images and the GIFs are the sizes of 256×192. There are about 100 workers, and about 1000 assignments in the AMT study. The assignments are shown for workers in unlimited time. The workers are asked for picked one option which captures the pose sequence, the desired clothes, and the appearance of the person well. The results are shown in the Table 2, which reports the FW-GAN outperform the other methods and achieve the highest human preference scores.

	Human Preference Score (limited time)	Human Preference Score (unlimited time)
FW-GAN (ours) / CP-VTON [36]	0.5940 / 0.4060	0.889 / 0.111
FW-GAN (ours) / VITON [16]	0.5721 / 0.4279	0.893 / 0.107

Table 2. Human perceptual study with others on the VVT dataset.

5.1. Ablation Study

We conduct an ablation study to explore the effects of the important component of FW-GAN. The results are reported in Table. 1. Our model without grid module, flow module and parsing constraint loss got the best FID score in I3D, and the model without flow module achieved the best FID score in ResNeXt-101. Although our full model didn't obtained the best FID score, Figure 4 demonstrates that our full model is capable to synthesize more photo-realistic images with clearer and more complete clothing patterns. On the other hand, FID uses the output of last pooling layer

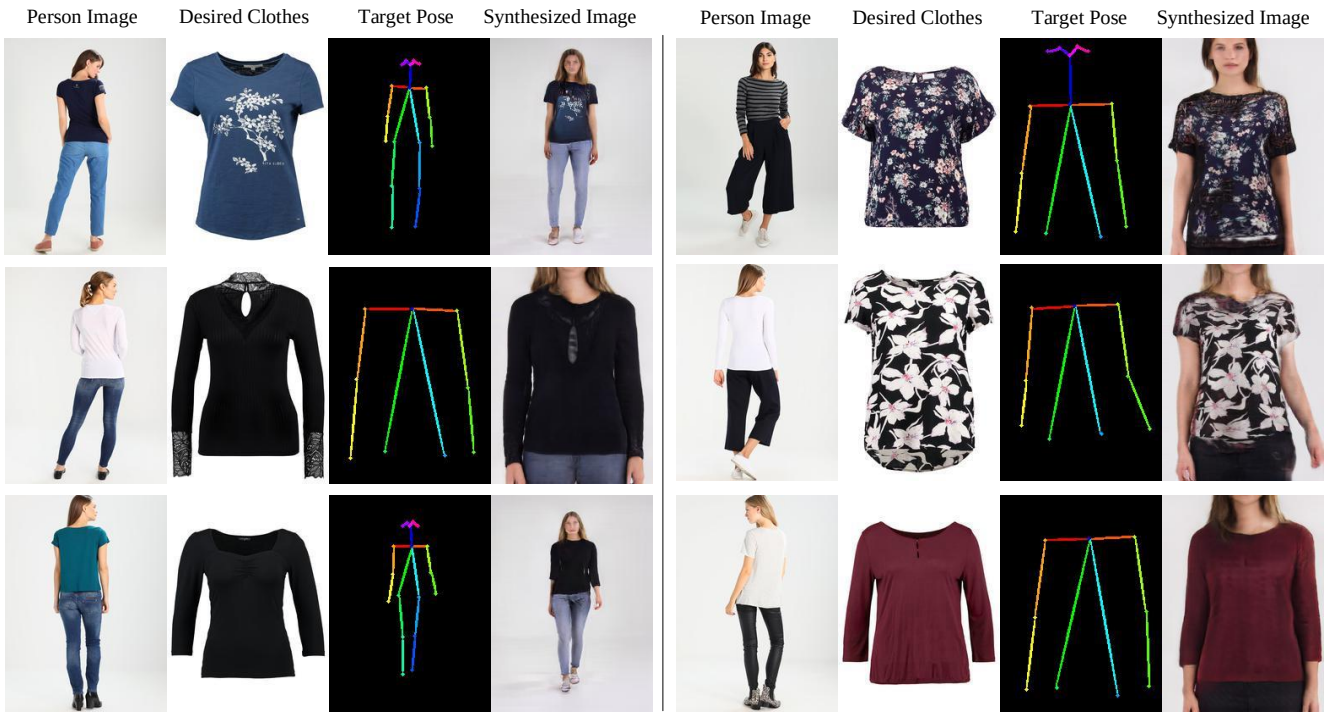

Figure 6. Some results of FW-GAN on the VVT dataset.

as the feature vector, which loses some information of the original image input, and the FID scores among our ablation models differ not much.

Figure 7. Some failure results conduct on the VVT dataset, which were caused by uncommon clothes.

6. Conclusion

We propose Flow-navigated Warping Generative Adversarial Network (FW-GAN) for video virtual try on, which generates novel person video in arbitrary poses and various clothes. To achieve good virtual try-on quality, our FW-GAN mainly contains three components: 1) FW-GAN incorporate the optical flow and geometric matching for warping the frames and clothes image, respectively, which can preserve the details in global and local views, 2) a flow-embedding discriminator is proposed that incorporate an effective flow input to the discriminator to improve the spatiotemporal smoothing, and 3) a parsing constraint loss function as one form of structural constraints to explicitly encourage the model to synthesize results under difference poses and various clothes to produce coherent part configurations with the input image. Our experimental results demonstrate that the proposed FW-GAN significantly outperforms other state-of-the-art approaches on synthesizing video of virtual try-on by manipulating pose and clothes.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U1611264, 61472453, U1401256, U1501252, U1711261, U1711262, 61602530, 61836012, 61622214), the National High Level Talents Special Support Plan (Ten Thousand Talents Program), the Natural Science Foundation of Guangdong Province under Grant No. 2017A030312006, and the Key R&D Program of Guangdong Province (2018B010107005).

References

[1] Hilsmann Anna et al. Tracking and retexturing cloth for real-time virtual clothing applications. *dated May*, 4:12, 2009.

[2] Xue Bai, Jue Wang, David Simons, and Guillermo Sapiro. Video snapcut: robust video object cutout using localized classifiers. In *ACM Transactions on Graphics (ToG)*, page 70. ACM, 2009.

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36. Springer, 2004.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, July 2017.

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018.

[7] Tao Chen, Jun-Yan Zhu, Ariel Shamir, and Shi-Min Hu. Motion-aware gradient domain video composition. *IEEE Trans. Image Processing*, 22(7):2532–2544, 2013.

[8] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, pages 474–484, 2018.

[9] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *ICCV*, 2019.

[10] Haoye Dong, Xiaodan Liang, Chenxing Zhou, Hanjiang Lai, Jia Zhu, and Jian Yin. Part-preserving pose manipulation for person image synthesis. In *ICME*, pages 1234–1239, 2019.

[11] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision*, 2015.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35–1, 2012.

[14] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019.

[15] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry S. Davis. Compatible and diverse fashion image inpainting. In *ICCV*, 2019.

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018.

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 18–22, 2018.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[21] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. *ICCVW*, 2(6):8, 2017.

[22] Zorah Laehner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2018.

[23] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 6, 2017.

[24] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, 2017.

[25] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[27] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017.

[28] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018.

[29] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, volume 2, 2017.

[30] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014.

[31] Eli Shechtman, Yaron Caspi, and Michal Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):531–545, 2005.

[32] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[33] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR 2018-Computer Vision and Pattern Recognition*, 2018.

[34] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[36] Bochao Wang, Huabin Zhang, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018.

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[40] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time video completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[41] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016.

[42] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, volume 2, page 3, 2017.

[43] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In *Proceedings of the 26th ACM international conference on Multimedia*, 2018.