FashionMirror: Co-attention Feature-remapping Virtual Try-on with Sequential Template Poses

Chieh-Yun Chen Ling Lo Pin-Jui Huang Hong-Han Shuai Wen-Huang Cheng
National Chiao Tung University
Hsinchu, Taiwan
{cychen.ee09g,lynn97.ee08g,i309505013.eic09g,hhshuai,whcheng}@nctu.edu.tw

Abstract

Virtual try-on tasks have drawn increased attention. Prior arts focus on tackling this task via warping clothes and fusing the information at the pixel level with the help of semantic segmentation. However, conducting semantic segmentation is time-consuming and easily causes error accumulation over time. Besides, warping the information at the pixel level instead of the feature level limits the performance (e.g., unable to generate different views) and is unstable since it directly demonstrates the results even with a misalignment. In contrast, fusing information at the feature level can be further refined by the convolution to obtain the final results. Based on these assumptions, we propose a co-attention feature-remapping framework, namely FashionMirror, that generates the try-on results according to the driven-pose sequence in two stages. In the first stage, we consider the source human image and the target try-on clothes to predict the removed mask and the try-on clothing mask, which replaces the pre-processed semantic segmentation and reduces the inference time. In the second stage, we first remove the clothes on the source human via the removed mask and warp the clothing features conditioning on the try-on clothing mask to fit the next frame human. Meanwhile, we predict the optical flows from the consecutive 2D poses and warp the source human to the next frame at the feature level. Then, we enhance the clothing features and source human features in every frame to generate realistic try-on results with spatio-temporal smoothness. Both qualitative and quantitative results show that FashionMirror outperforms the state-of-the-art virtual try-on approaches.

1. Introduction

In this paper, we envisage a new shopping scenario. Imagine that we stand in front of a fashion mirror inside a shopping mall. The fashion mirror shows the real-time virtual try-on results of the selected clothes. Therefore, we can exhibit arbitrary poses to guide the synthesized try-on result in the fashion mirror for viewing how suitable the garments are in multi-aspects as demonstrated in Fig. 1. To achieve this goal, one intuitive way is to apply the single-pose virtual try-on methods (e.g., [41]) for the first frame and apply the sequential pose transformation (e.g., [31]) for the following frames. However, since the clothing information only depends on the first frame, errors may accumulate in sequential generation. Another possible solution is to use the multi-pose virtual try-on methods (e.g., [19]) in a frame-by-frame manner. Nevertheless, the results of consecutive frames may be inconsistent (i.e., flickering artifacts) since they are generated independently.

Figure 1. **Virtual try-on results in sequential poses.**

To consider the sequential information for video-based virtual try-on, FWGAN [10] proposes a flow-navigated warping GAN, which (i) warps the clothing image and refines the clothing texture at the pixel level, (ii) warps the previous frame via the optical flow generated by [11], and (iii) conducts the pre-processed semantic segmentation to remove the clothes from the source human image. However, the performance is limited by the generated optical flow and warping the clothes at the pixel level prohibits

the network from generating new contents. For example, the clothing contents in the side view cannot be obtained from the front view. Meanwhile, fusing the refined clothes at the pixel level easily generates unstable results (e.g., the critical occlusion problem). For example, when the try-on model conducts the fusion between humans with limbs in front of the torso and warped clothes at the pixel level, the clothes veil the limbs as the green box shown in Fig. 4. Most previous works [38, 14, 41, 28, 9, 45, 19, 10] require the pre-processed semantic segmentation, which is time-consuming, and the quality of the segmentation highly affects the follow-up try-on results. To reduce the time cost, [22] proposes a parsing-free virtual try-on method. However, it cannot transfer the users' pose to obtain the information from different views. This is important for the real-world try-on scenario since users usually try on the clothes and exhibit different poses for evaluating whether the garment is suitable.

To address these issues, we propose a co-attention feature-remapping try-on framework, namely FashionMirror. Given a source human image, a target try-on clothing image, and a guiding pose sequence, the goal is to synthesize the try-on results in sequential poses according to the guiding pose sequence. The proposed FashionMirror framework consists of two stages: (I) parsing-free co-attention mask prediction and (II) human and clothing feature remapping. In stage (I), instead of using the semantic body parts, FashionMirror directly leverages a co-attention mechanism to learn the relation between consecutive human frames and the target try-on clothes for finding the regions related to the try-on clothes. Based on the co-attended results, FashionMirror predicts i) the removed mask, representing the clothing region that should be removed from the source human, and ii) the target try-on clothing mask, representing the region that the target clothing should fit. In stage (II), the goal is to synthesize the target try-on results based on the removed and target try-on clothing masks. Since remapping the visual information of target try-on clothes at the pixel level suffers from different-view and unstable issues as mentioned earlier, we warp the human and clothing information at the feature level for achieving the realistic try-on results. Specifically, the skeleton flow extraction network learns the feature-level optical flows among consecutive frames. By warping the current-frame human features with the extracted feature flows, we conduct the human sequence generation to transfer the current-frame human to the next pose. Furthermore, we enhance the source human feature and the target clothing feature within every frame for improving detailed information.

We evaluate the efficacy of the proposed FashionMirror with several state-of-the-art methods on both subjective and objective experiments. The results manifest that FashionMirror outperforms state-of-the-art methods qualitatively and quantitatively. The contributions are summarized as follows:

- We propose a co-attention feature-remapping virtual try-on framework, namely FashionMirror, to envisage a new shopping scenario via synthesizing the realistic try-on results in sequential poses with spatio-temporal smoothness.
- The proposed parsing-free co-attention mask mechanism replaces the commonly-used semantic segmentation for virtual try-on and reduces the inference time by 42.84%.

2. Related Work

2.1. Pose-guided video generation

Various researches are proposed for the pose-guided video generation [39, 6, 43, 46, 31, 40, 42]. For instance, Wang *et al.* [39] first proposed a video-to-video synthesis approach to tackle the incoherent problem caused by directly applying the image-based method to synthesize the video sequence. Chan *et al.* [6] focused on the motion transfer task and considered 2D pose skeleton extracted by OpenPose [4] as the intermediate representation to simplify the model. Zablotskaia *et al.* [43] further adopted 3D pose representation extracted by DensePose [1] to preserve the detailed appearance from the source human. Ren *et al.* [31] first cleaned the 2D pose representation to obtain more smooth representation sequence, and proposed a global-flow local-attention network to reassemble the input features. However, directly applying pose-guided video generation to the first try-on frame may lose detailed clothing information.

2.2. Virtual try-on

Virtual try-on methods can be categorized into two groups, 3D-based methods [30, 13, 2, 24, 27, 36, 29] and 2D-based methods [15, 38, 18, 19, 45, 9, 14, 41, 28, 22, 8]. Since this paper aims at 2D-based methods due to the requirement of the fast inference speed, we only review the 2D-based methods here.

Specifically, VITON [15] proposed a coarse-to-fine image-based network to synthesize try-on results with proper geometric alignment. CPVTON [38] further improved VITON by replacing the hand-crafted shape-context matching with the learnable thin-plate spline transformation to preserve the details of clothing characteristics. However, the network tends to be restricted from generating new contents (different views) and may be unstable when trying on the clothes via warping and fusing them with the human image at the pixel level. ACGPN [41] further manipulated the semantic segmentation to learn the body parts information for dealing with the occlusion problem. Nevertheless,

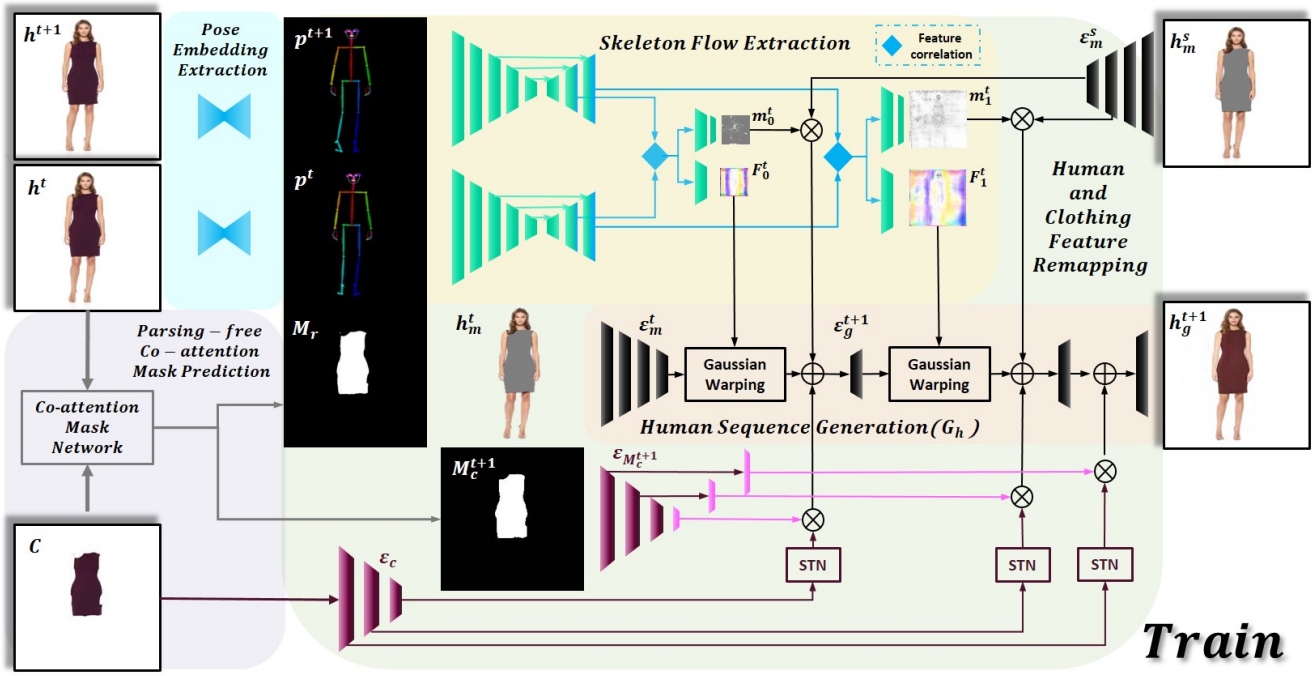

Figure 2. **Training overview**, which consists of two stages. Stage I (*Parsing-free co-attention mask prediction*) predicts the removed mask M_r and the try-on clothing mask M_c for providing the information of the target try-on clothing shape corresponding to the source human body shape. Stage II (*Human and clothing feature remapping*) extracts the pose embedding $\{p^t\}_{t=1}^N$ from the guiding pose sequence $\{h^t\}_{t=1}^N$ in N frames via the *Pose Embedding Extraction*, and learns the feature flows between the consecutive frames with the *Skeleton Flow Extraction*. Then, *Human Sequence Generation* remaps the features to generate the try-on results $\{h_g^t\}_{t=1}^N$.

these approaches are limited since they can only use for fixed poses. [19, 45, 9] incorporated pose transformation into the virtual try-on task. For instance, VTNCAP [45] improved the architecture of [38] to achieve pose transformation. FashionOn [19] designed a semantic-guided image-based network to generate the realistic virtual try-on results with arbitrary poses. TF-TIS [8] further extended FashionOn to synthesize the suitable poses based on the user-specified clothes. However, conducting the pre-processing semantic segmentation is time-consuming and easily causes error accumulation [22]. To consider the temporal information, FWGAN [10] proposed a flow-navigated warping GAN to tackle the video-based virtual try-on based on the pose-guided video generation [39] and the clothes warping module within the virtual try-on [38].

3. Proposed Method

Given a source human image h^s, a target try-on clothing image C, and a sequence of guiding pose images $\{h^t\}_{t=1}^N$ in N frames, the goal is to generate the sequential try-on results according to the guiding pose sequence. We propose a co-attention feature-remapping framework, called FashionMirror. Fig. 2 shows the architecture of FashionMirror, which consists of two stages: (I) parsing-free co-attention mask prediction and (II) human and clothing feature remapping.

3.1. Parsing-free co-attention mask prediction

The information of human body parts is essential for virtual try-on. It helps the model know where the target try-on clothes should fit the source human image and capture the dependency between bodies and clothes, e.g., folded arms occlude the clothes on the chest area. Most state-of-the-art virtual try-on models [38, 14, 41, 28, 9, 45, 19, 10] rely on the semantic segmentation, which contains multiple channels with each channel representing one body part, to provide the information of the human body parts. The semantic segmentation helps guide the learning of virtual try-on to generate the results with clear spatial information. Nevertheless, pre-processing the semantic segmentation information is time-consuming and prone to cause the accumulated error [22]. Inspired by [20, 26], we propose a co-attention mask network to learn the clothing-related masks, which retains the helpful semantic information but alleviates the efficiency issue.[1]

Co-attention Mask Network (*CMN*). Given the source human image h^s and the target try-on clothing image C, *CMN* jointly learns to predict the removed mask M_r and the try-

[1]The running time improvement is reported in Sec. 4.3.

on clothing mask M_c, i.e.,

$$M_r, M_c = CMN(h^s, C), \tag{1}$$

where $M_r, M_c \in [0,1]^{1\times W\times H}$ represent the clothing region that should be removed on h^s and the target try-on clothing shape corresponding to the source human body shape, respectively. *CMN* is devised to equip two mechanisms: single-frame and multi-frame mechanism, as demonstrated in Fig. 3. Specifically, the single-frame mechanism is for the first two frames. After processing the first two frames, *CMN* considers the previous two frames and the current-frame human image to predict the next-frame try-on clothing mask M_c^{t+1}. Following the architecture of [20, 26], *CMN* adopts a siamese ResNet [16] with first 5 layers followed by an atrous spatial pyramid pooling (ASPP) module [7] to extract both features ε_h and ε_c from h^s and C. Then, it calculates the similarity matrix $S = {\varepsilon_h}^T W \varepsilon_c$ at the feature level for finding the co-attended features between the clothing features on the human image and the clothing image, where W is a learnable weight matrix. S is further normalized column-wise and multiplies the clothing feature matrix ε_c to get the clothing co-attention feature $\varepsilon_{cA} = \varepsilon_c softmax(S)$. Similarly, we multiply the human feature matrix ε_h to the normalized S to get the human co-attention feature $\varepsilon_{hA} = \varepsilon_h softmax(S)$. As the background varies in different images, one 1×1 convolution is performed on ε_{cA} (or ε_{hA}) to learn how much attention should be put on C (or h^s) for adapting the variation. Next, ε'_{cA} (or ε'_{hA}) is concatenated with ε_h (or ε_c) as the input of the prediction layers to generate M_r and M_c.

To calculate the difference of both M_r and M_c to their corresponding groundtruth $\hat{M}_r$ and $\hat{M}_c$, which are the clothing channel obtained from the state-of-the-art semantic segmentation [12], we use the L_1 distance loss ($\mathcal{L}_{L_1}$), the binary cross-entropy loss ($\mathcal{L}_{BCE}$), and the clothing patch loss ($\mathcal{L}_{Patch}$). The overall objective function[2] for training the single-frame *CMN* is derived as follows.

$$\mathcal{L}_{CMN} = \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{Patch}\mathcal{L}_{Patch}, \tag{2}$$

For multi-frame *CMN*, the overall loss function is only based on M_c^{t+1} since it only predicts one output M_c^{t+1}.

3.2. Human and clothing feature remapping

After predicting the removed clothing region from h^s and the target try-on clothing shape, we introduce the feature remapping mechanism for trying on the source human with the guiding pose sequence. In this stage, the *Pose Embedding Extraction* first simplifies the guiding pose sequence into the skeleton representation. Afterward, the *Skeleton Flow Extraction* extracts the feature-level optical flows and the selection masks based on the skeleton representation of the consecutive frames. Furthermore, the *Human Sequence Generation* warps the current-frame human to the next pose. Simultaneously, enhancing the robust features based on the clothing features and the source human features makes the try-on results more realistic.

[2]The details are shown in Appendix A.

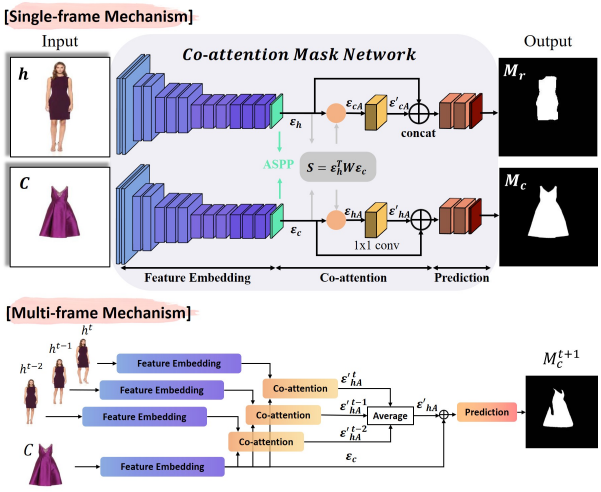

Figure 3. The architecture of the Co-attention Mask Network.

Pose Embedding Extraction. The sequence of guiding pose images $\{h^t\}_{t=1}^N$ is simplified into the 3-channel RGB skeleton images, obtained by the state-of-the-art pose estimation function $PE(.)$ (e.g., *OpenPose* [3]).

$$\{p^t\}_{t=1}^N = PE(\{h^t\}_{t=1}^N), \tag{3}$$

where p represents the skeleton image. Instead of using the commonly-used 18-channel joint coordinates to represent only body part, our pose representation is up to 137 points, i.e., 25 points for the human body, 2×21 points for hands, 70 points for face, as p^t shown in Fig. 2.

Skeleton Flow Extraction (*SFE*). After deriving the skeleton sequence, we design *SFE* to extract the feature-level optical flow (denoted by F) and the selection mask (denoted by m) between the two consecutive frames t and $t+1$.

$$F^t, m^t = SFE(p^t, p^{t+1}), \tag{4}$$

where F^t and m^t can be selected by the attentive layers as shown in Fig. 2. F^t helps the current-frame human h^t warp to h^{t+1}, and $m^t \in [0,1]$ represents whether the next-frame information h^{t+1} should be obtained from h^t or h^s. For extracting F^t and m^t, *SFE* first extracts the features from p^t and p^{t+1} separately via 5 encoded layers. Then, it decodes and combines the encoded features. Finally, it pays attention to the attentive layers and calculates the correlation between p^t and p^{t+1} at the feature level to predict the optical flow and the selection mask.

While optical flow provides a good supervision to guide the regional spatial transformation, it is challenging to learn

the optical flow due to the poor alignment using commonly-used bilinear sampling with large spatial differences. To tackle this problem, we set the model to learn the optical flow of two frames with a slightly spatial transform, i.e., between consecutive frames with 0 - 2 random skip frames in 30 fps. Besides, following the previous works [33, 32], we conduct the Gaussian sampling instead of the bilinear sampling to warp the features and apply the sample correctness loss to guide *SFE* to learn efficiently. Instead of using the optical flows generated by the state-of-the-art methods [11, 21] as the groundtruth, the sample correctness loss $\mathcal{L}_{corr}$ is conducted via the cosine similarity $\mu(*)$ with the pre-trained VGG19 [35]. Equipped with the sample correctness loss, *SFE* learns the data-driven flows specifically for human pose transfer and virtual try-on (as the visualization shown in Fig. 5) and prevents being limited by the performance of [11, 21].

$$\mathcal{L}_{corr}(h^t, F^t, h^{t+1}) = \frac{1}{N}\sum_{l\in\Omega} exp(-\frac{\mu(\varepsilon^t_{l,F}, \varepsilon^{t+1}_l)}{\mu^{max}_l}), \quad (5)$$

where $\varepsilon^t_{l,F}$ represents the feature extracted from h^t via the selected attentive layers of VGG19 at location $l = (x, y)$ and warped with F^t. ε^{t+1}_l is the feature extracted from h^{t+1} at location l. μ^{max}_l is the normalization term helping to constraint the flow variation:

$$\mu^{max}_l = \max_{l'\in\Omega} \mu(\varepsilon^t_{l'}, \varepsilon^{t+1}_l), \quad (6)$$

where Ω is the coordinate set consisting of all N coordinates within the feature map.

Human Sequence Generation *(HSG)*. The *HSG* network finalizes the try-on results in sequential poses, denoted by G_h. By taking the inputs of the source human h^s, the target try-on clothing image C, the pose embedding of the sequence $\{p^t\}^N_{t=1}$, and the help of *CMN*, the goal of G_h is to generate the try-on sequence $\{h^t_g\}^N_{t=1}$.

$$\{h^t_g\}^N_{t=1} = G_h(h^s, C, \{p^t\}^N_{t=1}, CMN(\{h^{t-1}\}^N_{t=1}, C)), \quad (7)$$

where h^0 is equal to h^s. Specifically, *HSG* first masks the clothes from the source human $h^s_m = h^s \otimes (1 - M_r)$, where $\otimes$ represents pixel-wise multiplication, for preventing the source clothes from perturbing the try-on process. Then, M_c provides the structural information about the clothing shape of C fitting h^s. Simultaneously, we extract the features of h^s_m, h^t_m[3], C, and M^{t+1}_c for preparing the following warping process. Let $(\varepsilon^s_m)_i$, $(\varepsilon^t_m)_i$, $(\varepsilon_c)_i$, and $(\varepsilon_{M^{t+1}_c})_i$ denote the features of h^s_m, h^t_m, C, and M^{t+1}_c in the i^{th} attentive layer. Here, we integrate all the information at the feature level to make the network be able to generate new contents (different views) and can apply convolution operations to refine the results.

$$\begin{aligned}(\varepsilon^{t+1}_g)_i = &\, GaussianWarping((\varepsilon^t_m)_i, (F^t)_i) \\ &+ \lambda_{h^s}(\varepsilon^s_m)_i \otimes (m^t)_i \\ &+ \lambda_{c_i}(STN(\varepsilon_c)_i) \otimes (\varepsilon_{M^{t+1}_c})_i,\end{aligned} \quad (8)$$

where $(\varepsilon^{t+1}_g)_i$ is the i^{th} layer feature of the generated try-on result in $t+1$ frame. λ_{h^s} and λ_{c_i} are the hyperparameters for controlling the balance between ε^t_m, ε^s_m, and ε_c. λ_{c_i} increases as the receptive field decreases since it contains more details.

The overall objective function consists of both spatial and temporal loss.

$$\mathcal{L}_{tryon} = \mathcal{L}_{spatial} + \mathcal{L}_{temporal}. \quad (9)$$

The spatial loss guides the learning to generate high-quality try-on results and the temporal loss teaches the smooth clothing details variation (e.g., smooth variations in wrinkle structure). $\mathcal{L}_{spatial}$ can be categorized into two parts and $\mathcal{L}_{temporal}$ can be categorized into three parts:

$$\mathcal{L}_{spatial} = \mathcal{L}^s_{human} + \mathcal{L}^s_{clothes}, \quad (10)$$

$$\mathcal{L}_{temporal} = \mathcal{L}_{flow} + \mathcal{L}^t_{human} + \mathcal{L}^t_{clothes}, \quad (11)$$

where $\mathcal{L}_{human}$ and $\mathcal{L}_{clothes}$ deal with the human and clothing information, respectively. $\mathcal{L}_{flow}$ is the weighted version of sample correctness loss $\mathcal{L}_{corr}$. Due to the space constraint, please refer to Appendix A for more details.

4. Experiment

This section presents the details of the experimental setup, i.e., dataset, implementation details, state-of-the-art baselines, and evaluation metrics. Afterward, qualitative and quantitative analyses are conducted with the state-of-the-art methods. For the video examples, please refer to https://github.com/FashionMirror/FashionMirror.

4.1. Experimental Setup

Dataset. To train and evaluate the sequential virtual try-on, a dataset containing a sequential-pose video and a clothing image related to the human in the video is required. However, there is no public dataset[4] consisting of the related clothing image. We design the novel training process to simplify the dataset requirement, which only needs a sequential poses video that can be accessed from the video

[3]For training, we masked h^t in every frame for preventing the clothing information from perturbing the training since C and the clothes on h^t are the same one.

[4]It is worth noting that the dataset published by FWGAN [10] on the competition website is only partial, i.e., only contains the first frame human, the related clothes, and the pose representation, without the human frames related to the pose representation.

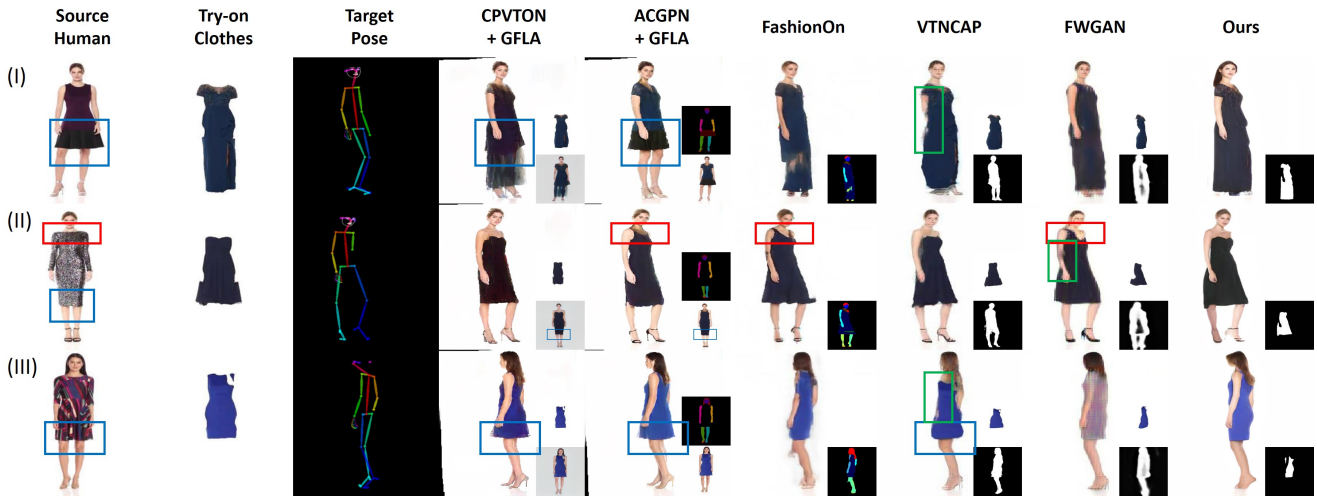

Figure 4. **The visual comparison within 5 baselines.** The most left three columns are input sets. We show the try-on results with auxiliary results positioned in the right-bottom side. The auxiliary results for each baseline are [CPVTON + GFLA]: the warped clothes and the try-on result before adopting GFLA, [ACGPN + GFLA]: the predicted semantic segmentation and the try-on result before adopting GFLA, [FashionOn]: the predicted semantic segmentation, [VTNCAP]: the warped clothes and the predicted target body shape mask, [FWGAN]: the warped clothes and the grid mask, and [Ours]: the try-on clothing mask M_c. (Zoom in for getting clear information.) *Please refer to* `https://raw.githubusercontent.com/FashionMirror/FashionMirror/main/Try-on%20results/visual_comparison.gif` *for having the temporal information in a browser.*

generation task [43, 6, 42] via replacing the clothing image with the clothes retrieved from the human frontal frame. However, only the FashionVideo dataset [43] has a connection to Fashion and contains high variations of humans. Therefore, we conduct the FashionVideo dataset to train and evaluate the proposed FashionMirror. There are 500 videos (191,684 frames) for training and 100 videos (38,838 frames) for testing with the resolution of 256 $\times$ 256.

Implementation Details. We train our network in stages. The co-attention mask network is first trained. Then, we train the whole model end-to-end with batch size of 4 on two NVIDIA 2080-ti GPUs. Every iteration generates 6 frames and the consecutive frames randomly skip 0 - 2 frames in the source video to make the model learned the variation. We apply the Adam [25] optimizer with learning rate of 0.0001.

Baselines. We compare the proposed *FashionMirror* with five baselines, including three types of virtual try-on works, trained on the FashionVideo dataset [43]. (I) Single-pose image-based virtual try-on: CPVTON [38] and ACGPN [41], (II) Multi-pose image-based virtual try-on: FashionOn [19] and VTNCAP [45], and (III) Video-based virtual try-on: FWGAN [10]. For fairness, instead of directly comparing the single-pose image-based works with video-based work, we first deploy the single-pose image-based virtual try-on work and then feed the try-on result to the video-based human generation work GFLA [31] for transferring the poses, denoted by "+ GFLA".[5]

[5]It is worth noting that we do not compare the proposed approach with [14, 9, 23] since they do not release their codes.

Metrics. We evaluate the quality in terms of both image-based evaluation metrics and video-based evaluation metrics. For image results, we conduct i) Inception Score (IS) [34], measuring the image quality and diversity, ii) Structural Similarity (SSIM) [47], measuring the similarity between the reconstruction results and the groundtruth, and iii) Learned Perceptual Image Patch Similarity (LPIPS) [44], measuring the perceptual similarity between the reconstruction results and the groundtruth. For video results, we conduct Video Fréchet Inception Distance (VFID) [17], which is used to measure visual quality and temporal consistency. We conduct two pre-trained video recognition CNN backbones: I3D [5] and 3D-ResNet-18 [37] to extract both temporal and spatial features.

4.2. Qualitative Results

Fig. 4 shows the visual comparison of the proposed method with 5 baselines in 3 challenging cases: (I) human with no sleeve A-line dress tries on short sleeve one-piece dress, (II) human with long sleeve bodycon dress tries on no sleeve off-shoulder A-line dress, and (III) human with long sleeve shift dress tries on no sleeve bodycon dress. The results of different try-on models are summarized as follows.

The FashionVideo dataset is more complicated as compared with the commonly-used try-on datasets, e.g., VITON [15], since the latter only contains the close-fitting clothes. Hence, in the FashionVideo dataset, it is challenging for the try-on methods to change the clothes between loose and tight (as the blue box shown in Fig. 4). CPVTON, ACGPN, and VTNCAP fail to tackle this challenge.

As shown in case (I), CPVTON, VTNCAP, and FWGAN perform well on warping the clothes to fit the body shape (as shown in the auxiliary result). However, when CPVTON tries on the warped clothes by fusing the source human image and the warped clothes via a composition mask, the source A-line dress worsens the result. This is due to fusing clothes and human information at the pixel level. Besides, in case (III), the warped garment for FWGAN is misaligned. FWGAN further refines the try-on results at the pixel level via the grid mask leads to unstable results. In contrast, FashionMirror remaps the clothing information at the feature level and avoids the artifacts.

On the other hand, the semantic segmentation guided works (e.g., ACGPN and FashionOn) heavily rely on the semantic segmentation. In case (I), ACGPN predicts the bottom of the source human image as a skirt via the semantic segmentation. However, to try on the one-piece dress, ACGPN is not aware that it needs to take off the skirt and thus leads to failure. Meanwhile, FashionOn gets a broken dress due to the broken semantic segmentation in case (I). While in case (II), both ACGPN and FashionOn mistakenly synthesize the open-shoulder top (as the red box shown in Fig. 4) due to the corresponding semantic segmentation. In contrast, FashionMirror leverages the co-attention mask network for efficiently predicting the removed mask and the try-on clothing mask. For more qualitative results, please refer to Appendix B.

Ablation Study. To verify the essential components of *FashionMirror*, Fig. 5 visualizes the results, feature-level optical flow, and selection mask of the following four models: **(I) FashionMirror**, **(II) FashionMirror (w/o h^s boost)**, which eliminates the source human feature enhancement for every frame and only relies on the input source human in the first frame, **(III) FashionMirror (w/o λ_{c_i})**, which fuses the clothing features without the layer weighting, and **(IV) FashionMirror (with multi-flow)**, which applies flow prediction network to replace the STN from the clothing feature enhancement and conducts flow prediction network to source human feature enhancement to help source human feature approach similar distribution as h^{t+1}.

Fig. 5 demonstrates that the feature-level optical flow of our full model concentrates on the human region of h_g^t and synthesizes more detailed results, especially the facial region. When FashionMirror eliminates h^s, the optical flow loses the focus to the specific human region and puts the effort on the whole image. While FashionMirror conducts multiple flows, the flow of h^t is perturbed by the other two flows and becomes partially strip, making the try-on result unstable. As for the selection mask, it is highly different within the four models. The selection mask distribution of our full model with and without λ_{c_i} are similar. However, the former is more confident than the latter, manifesting that the selection mask with λ_{c_i} is more confident about where

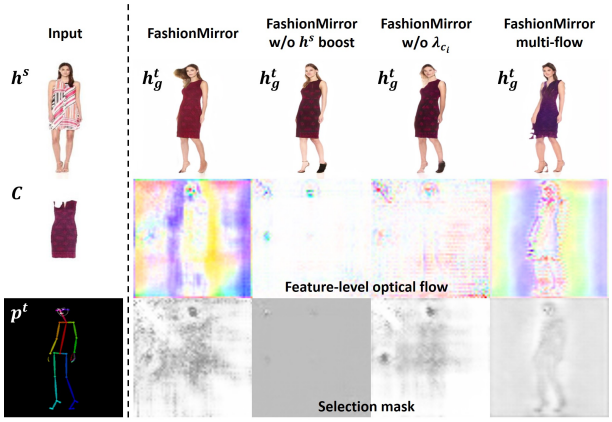

Figure 5. The visualization of the ablation study.

the feature should be extracted and synthesizes more realistic results. The selection mask of multi-flow FashionMirror contains the evident human region and remaps the human features with the average contribution from h^{t-1} and h^s. However, h^t must contain more features from h^{t-1} and less features from h^s. The average selection mask makes the result of multi-flow FashionMirror far from realistic. The selection mask of FashionMirror w/o h^s does not demonstrate any difference since the model without h^s does not need to select where the features should be taken.

4.3. Quantitative Results

To evaluate the reconstruction results[6], we randomly synthesize 2000 video clips from the test dataset while each video clip contains 20 frames. Table 1 compares the proposed FashionMirror with baselines in terms of image-based and video-based evaluation metrics, i.e., IS, SSIM, LPIPS and VFID. First, the difference of score distribution between image-based and video-based evaluation metrics is interesting. While FashionOn gets better or competitive scores than that of ACGPN + GFLA in image-based metrics, FashionOn gets worse scores in video-based metrics since the temporal coherence of FashionOn is worse. FashionMirror outperforms all the other baselines and the ablation models within both image-based and video-based evaluation metrics, demonstrating the highest visual quality and temporal coherence. It is worth discussing that FashionOn gets the highest IS score, but it does not synthesize the highest visual quality as shown in Fig. 4 and in the user study. Since IS evaluates the image quality based on ImageNet, it cannot well measure the details for try-on datasets [9].

User Study. We conduct a user study with 120 volunteers to evaluate visual quality. We randomly sample 13 input sets containing a source human, try-on clothes, and a pose

[6]Each reconstruction result is synthesized by a source human (masked the clothing region), a target garment (the same one on the source human), and a target pose sequence. Therefore, there are the corresponding groundtruths for the reconstruction results.

Table 1. Quantitative comparison on the test dataset with image-based metrics and video-based metrics.

Method		CPVTON +GFLA	ACGPN +GFLA	VTNCAP	FashionOn (G_r)	FWGAN	$Ours^f$	$Ours^{h^s}$	$Ours^{\lambda_{c_i}}$	$Ours$
IS↑		1.355 ± 0.009	2.219 ± 0.026	2.134 ± 0.030	**2.388 ± 0.040**	2.290 ± 0.028	2.290 ± 0.022	2.177 ± 0.027	2.187 ± 0.038	2.234 ±0.035
SSIM↑		0.815	0.864	0.877	0.889	0.907	0.887	0.906	0.919	**0.923**
LPIPS↓		0.228	0.109	0.119	0.111	0.074	0.092	0.073	0.060	**0.057**
VFID↓	I3D	5.499	4.809	10.182	9.622	7.961	5.226	4.593	3.141	**3.097**
	3D-ResNet	5.615	1.543	7.496	3.802	2.902	1.551	1.690	1.206	**1.033**

NOTE: $Ours^f$ denotes ours with multi-flow, $Ours^{h^s}$ denotes ours w/o h^s boost, and $Ours^{\lambda_{c_i}}$ denotes ours w/o λ_{c_i}.

Table 2. Results of user study.

Method	Same Ctype	Different Ctype	Average
CPVTON + GFLA	5.28%	9.50%	8.53%
ACGPN + GFLA	21.67%	23.92%	23.40%
FashionOn (G_r)	4.17%	5.83%	5.45%
VTNCAP	1.39%	2.50%	2.24%
FWGAN	13.33%	5.92%	7.63%
FashionMirror (Ours)	**54.17%**	**52.33%**	**52.76%**

sequence of the length between 20 to 150 frames from the test dataset. Volunteers are first shown with 6 videos at a time (synthesized by the five baselines and FashionMirror with the same inputs) and then requested to choose the most realistic try-on result in the desired pose sequence. The baselines can be categorized into three types: (I) Single-pose image-based virtual try-on: CPVTON and ACGPN, (II) Multi-pose image-based virtual try-on: FashionOn and VTNCAP, and (III) Video-based virtual try-on: FWGAN.

Table 2 summarizes the results. FashionMirror gets 52.76% of votes (823 votes) while five baselines in total get 47.24% of votes (737 votes), which verifies that FashionMirror outperforms the baselines and synthesizes realistic try-on results with spatio-temporal smoothness. The results manifest that the type (I) virtual try-on work cooperating with the video-based pose transform work is better to tackle the video-based virtual try-on task than the type (II) virtual try-on work because the type (II) work does not contain the coherent information. To further analyze the results, we separate the results according to the clothing type correlation between source human and the target try-on clothes, i.e., the same clothing type (Same Ctype) and different clothing types (Different Ctype). The result shows that FashionMirror outperforms all the baselines regardless of the clothing type correlation. FWGAN outperforms CPVTON + GFLA for the same clothing type try-on but is worse than CPVTON + GFLA for different clothing types. This is because tackling the different clothing types is more challenging. Though the clothing warping mechanism of CPVTON and FWGAN are both warping and fusing the clothing information at the pixel level, FWGAN would lead to more unstable results since it operates every frame.

Runtime. Since the proposed *CMN* (Sec. 3.1) aims to replace the commonly-used pre-processed semantic segmentation [12], which is time-consuming and easily causes error accumulation, we report the running time comparison to verify the efficacy of our co-attention mask network. Specifically, we randomly sample 40,000 input sets to report the average runtime for the co-attention mask network and the semantic segmentation with one NVIDIA 2080-ti GPU. The co-attention mask network costs 0.1983 sec on average, and the semantic segmentation costs 0.3469 sec. The runtime ratio of the co-attention mask prediction to the semantic segmentation is 57.16%, which shows that the novel co-attention mask network efficiently speeds up the virtual try-on process. For the runtime comparison within the whole try-on model, please refer to Appendix C.

5. Conclusion

This paper proposes a co-attention feature-remapping framework, namely FashionMirror, to synthesize the realistic virtual try-on results in sequential poses. We design a co-attention mask mechanism to maintain the advantage of semantic segmentation for virtual try-on (providing the region information of clothes), and reduce the inference time by 42.84%. Afterward, FashionMirror extracts the feature flows from consecutive frames to warp the current-frame human feature to the next pose with the source human and clothing feature enhancement within every frame to achieve realistic results. Experiments manifest that FashionMirror surpasses the state-of-the-art virtual try-on works both qualitatively and quantitatively. In the future, we plan to go towards the real-world shopping scenario and tackle the high-resolution virtual try-on task.

6. Acknowledgement

This work is supported in part by the Ministry of Science and Technology (MOST) of Taiwan under the grants MOST-109-2218-E-002-015, MOST-109-2221-E-009-114-MY3, MOST-109-2223-E-009-002-MY3, MOST-110-2218-E-A49-018, and MOST-110-2634-F-007-015. We are grateful to the National Center for High-performance Computing for computer time and facilities.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[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 (CVPR)*, 2017.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[8] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Template-free try-on image synthesis via semantic-guided optimization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[10] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. FW-GAN: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[11] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[12] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *European Conference on Computer Vision (ECCV)*, 2018.

[13] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. GarNet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[14] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. ClothFlow: a flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] 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 *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2017.

[18] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, and Wen-Huang Cheng. Fit-me: Image-based virtual try-on with arbitrary poses. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[19] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27^{th} ACM International Conference on Multimedia (ACM MM)*, 2019.

[20] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. 2019.

[21] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision (ECCV)*, 2020.

[23] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. SieveNet: a unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[24] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2020.

[25] Diederick P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[26] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese

networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[31] Yurui Ren, Ge Li, Shan Liu, and Thomas H. Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing (TIP)*, 2020.

[32] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. StructureFlow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. 2016.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[36] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*, 2020.

[37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *European Conference on Computer Vision (ECCV)*, 2018.

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2018.

[40] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. GAC-GAN: A general method for appearance-controllable human video motion transfer. *IEEE Transactions on Multimedia (TMM)*, 2020.

[41] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating $\leftrightarrow$ preserving image content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[42] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. TransMoMo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. DwNet: Dense warp-based network for pose-guided human video generation. In *British Machine Vision Conference (BMVC)*, 2019.

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27^{th} ACM International Conference on Multimedia (ACM MM)*, 2019.

[46] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[47] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 2004.