§ Supplementary Materials §

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

Appendix A - Detailed Objective Functions

Co-attention mask network *(CMN)*

The overall objective function for training the single-frame *CMN* is defined as:

$$\mathcal{L}_{CMN} = \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{Patch}\mathcal{L}_{Patch}. \quad (1)$$

For capturing the global structural information, we adopt $\mathcal{L}_{L_1}$ and $\mathcal{L}_{BCE}$:

$$\mathcal{L}_{L_1} = \sum_W \sum_H \left\| M_r - \hat{M}_r \right\|_1 + \sum_W \sum_H \left\| M_c - \hat{M}_c \right\|_1, \quad (2)$$

$$\begin{aligned}\mathcal{L}_{BCE} = &-\hat{M}_r \log(M_r) + (1-\hat{M}_r)\log(1-M_r) \\ &- \hat{M}_c \log(M_c) + (1-\hat{M}_c)\log(1-M_c).\end{aligned} \quad (3)$$

For capturing the regional information (e.g., excluding the human limbs cross in front of the clothes and considering the spaghetti strap), the patch loss is conducted to focus on small patches within the clothing region. We randomly select three 64x64 patches within the clothing region to learn the regional information based on the L_1 distance loss.

$$\begin{aligned}\mathcal{L}_{Patch} = &\sum_{i=1}^{3}\sum_W\sum_H \left\| (M_r)_i - (\hat{M}_r)_i \right\|_1 \\ &+ \sum_{j=1}^{3}\sum_W\sum_H \left\| (M_c)_j - (\hat{M}_c)_j \right\|_1,\end{aligned} \quad (4)$$

where *i* and *j* represent the index of the selected patches. For the 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}.

Human and clothing feature remapping

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

$$\mathcal{L}_{tryon} = \mathcal{L}_{spatial} + \mathcal{L}_{temporal}. \quad (5)$$

The spatial loss ($\mathcal{L}_{spatial}$) can be categorized into two parts:

$$\begin{aligned}\mathcal{L}_{spatial} &= \mathcal{L}^s_{human} + \mathcal{L}^s_{clothes} \\ &= \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_b\mathcal{L}_{body} \\ &+ \lambda_c\mathcal{L}_{content} + \lambda_s\mathcal{L}_{style} + \lambda_p\mathcal{L}_{patch}.\end{aligned} \quad (6)$$

In the following, we introduce every loss function. $\mathcal{L}_{GAN}$ helps the generator G_h to learn the real image distribution.

$$\mathcal{L}_{GAN} = \mathbb{E}[logD_s(h^t)] + \mathbb{E}[log(1 - D_s(h_g^t))], \quad (7)$$

where D_s is the spatial discriminator. $\mathcal{L}_{L_1}$ ensures the synthesis quality at the pixel level.

$$\mathcal{L}_{L_1} = \sum_W \sum_H \left\| h_g^t - h^t \right\|_1. \quad (8)$$

We design the body part loss ($\mathcal{L}_{body}$) to learn the structural information of the semantic segmentation since we do not take them as inputs to accelerate the inference time, i.e.,

$$\begin{aligned}\mathcal{L}_{body} = &\sum_W\sum_H \left\| (h_g^t - h^t) \otimes \hat{M}^t_{limb} \right\|_1 \\ &+ \sum_W\sum_H \left\| (h_g^t - h^t) \otimes \hat{M}^t_{bg} \right\|_1,\end{aligned} \quad (9)$$

where $\hat{M}^t_{limb}$ and $\hat{M}^t_{bg}$ represent the human limb mask and the background mask in frame t, respectively. The content loss and the style loss focus on the salient regions: head and clothes to imitate the feature distribution of the groundtruth

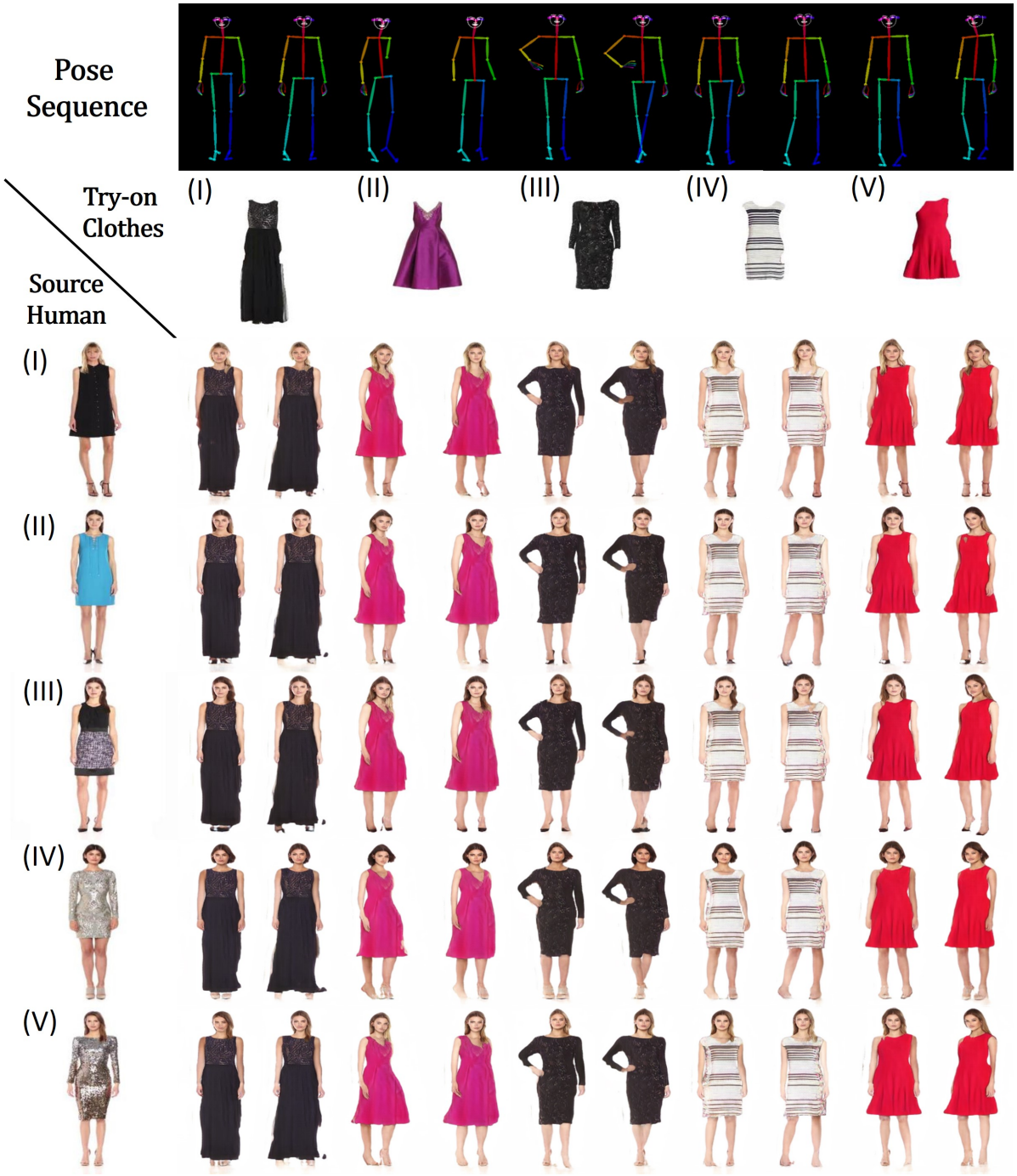

Figure 1. **Qualitative results.** FashionMirror performs robustly in changing the clothing tightness, e.g., from tight to loose: the source human (V) tries on the try-on clothes (V), and from loose to tight: the source human (I) tries on the try-on clothes (III).

for generating realistic details. The content loss is calculated by the pre-trained VGG19 [5] and the style loss is calculated by the gram matrix $\mathcal{G}$ [2].

$$\begin{aligned}\mathcal{L}_{content} = &\sum_i \left\|\phi_i(h_g^t) - \phi_i(h^t)\right\|_1 \\ &+ \sum_i \left\|\phi_i(h_g^t \otimes M_c^t) - \phi_i(h^t \otimes \hat{M}_c^t)\right\|_1 \\ &+ \sum_i \left\|\phi_i(h_g^t \otimes \hat{M}_{head}^t) - \phi_i(h^t \otimes \hat{M}_{head}^t)\right\|_1,\end{aligned} \tag{10}$$

where ϕ_i represents the feature map obtained from the i^{th} layer. $\hat{M}^t_{head}$ denotes the head mask in frame t.

$$\begin{aligned}\mathcal{L}_{style} = &\sum_i \left\|\mathcal{G}(\phi_i(h^t_g)) - \mathcal{G}(\phi_i(h^t))\right\|_1 \\ &+ \sum_i \left\|\mathcal{G}(\phi_i(h^t_g \otimes M^t_c)) - \mathcal{G}(\phi_i(h^t \otimes \hat{M}^t_c))\right\|_1 \\ &+ \sum_i \left\|\mathcal{G}(\phi_i(h^t_g \otimes \hat{M}^t_{head})) - \mathcal{G}(\phi_i(h^t \otimes \hat{M}^t_{head}))\right\|_1 .\end{aligned} \tag{11}$$

In order to focus on the detailed patch from the clothing region, we apply the same design as Eq. (4) to extract the local clothing patches from $h^t_g \otimes M^t_c$.

$$\begin{aligned}\mathcal{L}_{patch} = &\sum_i \left\|\phi_i(h^t_g \otimes M^t_c)_k - \phi_i(h^t \otimes \hat{M}^t_c)_k\right\|_1 \\ &+ \sum_j \left\|\mathcal{G}(\phi_j(h^t_g \otimes M^t_c)_k) - \mathcal{G}(\phi_j(h^t \otimes \hat{M}^t_c)_k)\right\|_1 ,\end{aligned} \tag{12}$$

where $k = ((x1, y1), (x2, y2))$ represents the box coordinate of the picked patch. **The temporal loss ($\mathcal{L}_{temporal}$)** can be divided into three parts:

$$\begin{aligned}\mathcal{L}_{temporal} &= \mathcal{L}_{flow} + \mathcal{L}^t_{human} + \mathcal{L}^t_{clothes} \\ &= \lambda_{corr}\mathcal{L}_{corr} + \lambda_{GAN_t}\mathcal{L}_{GAN_t} + \lambda_{p_t}\mathcal{L}_{patch_t}.\end{aligned} \tag{13}$$

In the following, we first introduce $\mathcal{L}_{GAN_t}$ and $\mathcal{L}_{patch_t}$.

$$\begin{aligned}\mathcal{L}_{GAN_t} = &\mathbb{E}[logD_t(\{h^t\}^n_{t=1})] \\ &+ \mathbb{E}[log(1 - D_t(\{h^t_g\}^n_{t=1}))],\end{aligned} \tag{14}$$

where D_t is the temporal discriminator and n represents a size of the video subsequence. For retaining the smooth variation of clothing wrinkles, we design a temporal patch loss:

$$\begin{aligned}\mathcal{L}_{patch_t} = &\mathbb{E}[logD_t(\{(h^t \otimes M^t_c)_k\}^n_{t=1})] \\ &+ \mathbb{E}[log(1 - D_t(\{(h^t_g \otimes M^t_c)_k\}^n_{t=1}))],\end{aligned} \tag{15}$$

where $k = ((x1, y1), (x2, y2))$ represents the box coordinate of the picked patch.

Appendix B - Additional Qualitative Results

Given five source humans and five try-on clothes with one pose sequence, Fig. 1 demonstrates the try-on results for different source humans and try-on clothes in one pose sequence. The results show the try-on results are stable to the clothing categories. For example, source human (I) with no sleeve shift dress can try on the long sleeve bodycon dress (try-on clothing III). In this case, the lower part of the shift dress is loose, but the lower part of the bodycon dress needs to tightly fit the human body shape and demonstrate the hourglass body. The try-on results indeed demonstrate this difference. Besides, source human (V) with long sleeve bodycon dress can try on the try-on clothes (II) or (V), which are no sleeve A-line dresses. The lower part of the bodycon dress tightly fits the source human body, but the lower part of the A-line dress widens from the waist to the hem. Furthermore, our try-on model can distinguish how wide the A-line dress is. The bottom of the try-on clothes (II) is wider than the try-on clothes (V) and the try-on results well demonstrate this characteristic, especially the try-on results of the source human (V).

Besides the clothing characteristic of tightness, it is worth discussing the clothing texture. For try-on clothes (I), the upper part of the dress contains unique texture and the dress is designed to slim down the waistline. Both two characteristics of the dress are well shown in the first two columns of the try-on results. For try-on clothes (IV), it contains horizontal lines in haphazard distribution, and the try-on results follow the distribution without over distortion.

Besides the clothing texture preservation, the human characteristic and human posture are essential for the virtual try-on results. For example, the source human (IV) has short and dark hair, and the source human (I) has long and blonde hair. Their try-on results in row (I) and (IV) preserve the human characteristics and demonstrate the difference. Take the middle columns in Fig. 1 as examples for unique human posture. The try-on results achieve unique poses with arm akimbo and feet cross, which is a unique pose for fashion catwalks.

In summary, the above cases show that our novel designed model, FashionMirror, can synthesize realistic try-on results in different poses. Specifically, FashionMirror prevents the try-on results from being affected by the clothes on the source human, demonstrates the clothing tightness, preserves the clothes texture, and achieves unique poses. For the try-on results in sequential poses, please refer to the video examples on https://github.com/FashionMirror/FashionMirror.

Appendix C - Additional Experiments

Runtime comparison of try-on model

Table 1. Inference time comparison. (sec/frame)

Method	Pose Estimation	Semantic Segmentation† or Co-attention Mask‡	Virtual Try-on	Total
CPVTON+GFLA	0.1168	0.3469†	0.1918	0.6555
ACGPN+GFLA			0.2336	0.6973
FashionOn (G_r)			0.0588	0.5225
VTNCAP			0.2325	0.6962
FWGAN			0.0980	0.5617
Ours		0.1983‡	0.0624	0.3775

We randomly sample 40,000 input sets to report the average inference time of baselines on one NVIDIA 2080-Ti GPU. Table 1 shows that our model is the most efficient and

outperforms all baselines, e.g., FashionMirror outperforms FWGAN in terms of runtime by 32.8%.

Visual quality comparison between co-attention mask and semantic segmentation

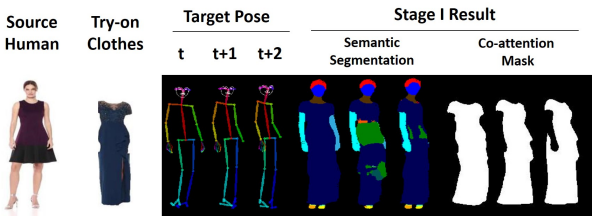

Figure 2. Co-attention mask vs. semantic segmentation.

We use [3] to generate semantic segmentation and use the first sub-network of [1] to predict the target semantic segmentation. Fig. 2 shows that the semantic segmentation for the clothing regions are inconsistent, including blue (dress) and green (skirt) even within the three consecutive frames. In contrast, the proposed co-attention mask focuses on the clothing region, and thus improves the performance.

Ablation study of the clothing patch loss (L_{Patch})

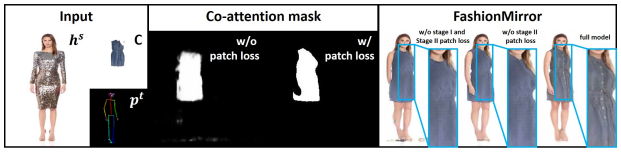

Figure 3. The ablation study of the clothing patch loss.

Fig. 3 shows the ablation study of the clothing patch loss (L_{Patch}) for stage I *(co-attention mask network)* and stage II *(human and clothing feature remapping)*. L_{Patch} makes stage I predict masks with more robust contour and helps stage II generates detailed information, e.g., the buttons.

Try on another clothing type

Figure 4. Try-on results with another clothing type.

Since the FashionVideo dataset [6] only contains one clothing type, i.e., dress, and there is no other fashion video dataset, we train our model on the image-based fashion dataset [4] by duplicating images as a video. Fig. 4 reports the try-on results with T-shirts, which demonstrates that FashionMirror can try on different kinds of clothes within different races if try-on videos with other styles are available.

Appendix D - Limitation

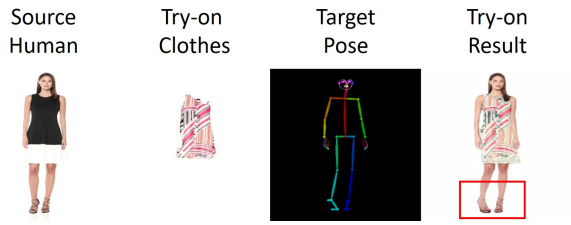

Figure 5. Failure case caused by the limited network concentration.

The high heels on the source human cannot always be well-preserved in the try-on results. As shown in Fig. 5, this failure is caused by the limited network concentration. The objective functions guide the network to focus on the global human information and the clothing region to synthesize the essential part correctly. As such, the region of the high heels compared to the global human is too small to be well-preserved since it only contributes a small amount to the objective function. Hence, it could be better to apply attention or a regularization term on the high heel region.

References

[1] 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. 4

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[3] 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. 4

[4] 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 27th ACM International Conference on Multimedia (ACM MM)*, 2019. 4

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[6] 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. 4