Supplementary – DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cross-modal Semantic Alignment

Xujie Zhang[1*] Binbin Yang[2*] Michael C. Kampffmeyer[4] Wenqing Zhang[1]
Shiyue Zhang[1] Guansong Lu[3] Liang Lin[2] Hang Xu[3] Xiaodan Liang[1,5†]
[1]Shenzhen Campus of Sun Yat-Sen University [2]Sun Yat-Sen University
[3]Huawei Noah's Ark Lab [4]UiT The Arctic University of Norway [5]MBZUAI
{zhangxj59,yangbb3,zhangwq76,zhangshy223}@mail2.sysu.edu.cn,michael.c.kampffmeyer@uit.no
luguansong@huawei.com,linliang@ieee.org,chromexbjxh@gmail.com,xdliang328@gmail.com

In this supplementary material, we provide additional details of the implementation (Sec. 1) and the human evaluation (Sec. 2). Further, we include additional qualitative results of our method and illustrate its use in another novel application of 2D try-on as well as 3D try-on (Sec. 3). Given a sentence and a picture of a person, we generate an image of a person wearing the clothes that match the description. This can further improve and significantly impact the fashion design process.

1. Additional Implementation Details

While most of the implementation details have been described in the paper, we provide the remaining details here.

Dynamic thresholding for garment manipulation. This section provides further details on the dynamic thresholding p mentioned in Section 3.4. Given an attention map $M_i^j \in \mathbb{R}^{H\times W}$, we collect and sort its elements before selecting the 75^{th} percentile pixel value as the threshold p, following the approach described in [3]. We then obtain a binary mask B_i^j by thresholding M_i^j with p:

$$B_i^j = \mathbb{I}[M_i^j \geq p], \tag{1}$$

where $\mathbb{I}$ is the indicator function.

Training details. For the segmentor, we use PointRend [1] as our segmentation network, to obtain segmentation results of the noisy garments. During the segmentor's training, we randomly select a timestep t between 0 to 1000 and add the corresponding noise to the training image.

As for the pretrained models, we loaded the parameters of stable diffusion [2] and then fine-tuned them on our data. The text encoder and image encoder architecture follows stable diffusion [2].

*Equal contribution. †Corresponding author.

2. Human Evaluation Details

For the human evaluation, we designed three separate questionnaires that evaluate different aspects of our DiffCloth method and the baseline methods. Presented with results produced by all methods, the first questionnaire asks participants to indicate the synthesis result that had the least amount of garment part leakage. In the second questionnaire participants are asked to select the result with least attribute confusion. Finally, in the third questionnaire, we asked which method best fit the manipulation prompt without changing any other part. Each questionnaire consisted of 100 tasks, and the presentation order of the different methods was randomized. Before the start of the human evaluation, we invited five volunteers to complete the questionnaires in a thorough manner to test the required time for completion. During the evaluation, we invited 100 random volunteers for each questionnaire, and they were instructed to spend at least 6 seconds to complete each task in the questionnaire.

3. Additional Results and Application

Visual Results for garment synthesis: Fig. 1 displays additional generation results by DiffCloth on the CM-Fashion dataset.

Visual Results for garment manipulation: Fig. 2 displays additional manipulation results by DiffCloth on the CM-Fashion dataset.

Application to Virtual Try-On: Virtual try-on aims to transfer a given garment onto a person and is often necessary in fashion design to receive feedback on designs. To address this, we developed a simple application that generates 2D and 3D try-on results for a given person and a textual description of a garment. Specifically, our application utilizes DiffCloth to generate a matching garment and then leverages ACGPN [4] and M3Dvton [5] to provide 2D and 3D generation results, respectively. Figure 3 and Figure 4 illustrate the try-on results generated by DiffCloth on the CM-Fashion dataset.

4. Potential Societal Impact

From a positive perspective, producing a garment that fits the description can have numerous advantages for the clothing design industry. It can lead to lower labor and material costs, while also providing designers with greater creative inspiration. However, it is important to implement measures that can prevent the misuse of this technology, such as copyright infringement and unethical practices like hiding trademarks during training.

References

[1] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9796–9805. Computer Vision Foundation / IEEE, 2020.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[4] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, pages 7850–7859, 2020.

[5] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *ICCV*, 2021.

Figure 1. Additional results of garment synthesis on the CM-fashion dataset.

Figure 2. Additional results of garment manipulation on the CM-fashion dataset.

Figure 3. 2D try-on results for the garments generated by DiffCloth.

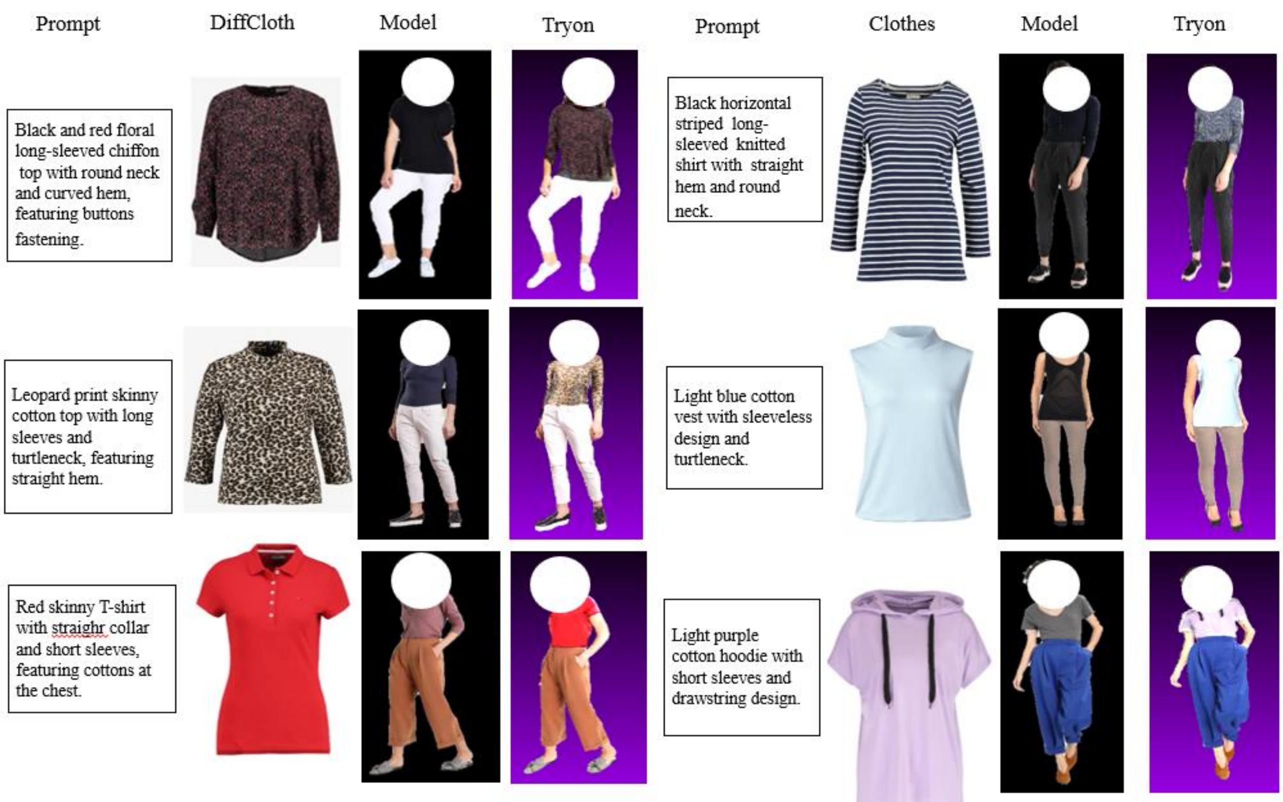

Figure 4. 3D try-on results for the garments generated by DiffCloth.