§ Supplementary Materials §
Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network

Chieh-Yun Chen[1,2*] Yi-Chung Chen[1,3*] Hong-Han Shuai[2] Wen-Huang Cheng[3]
[1]Stylins.ai [2]National Yang Ming Chiao Tung University [3]National Taiwan University

A. Eliminating Error Accumulation

Fig. 1 shows that our proposed *Outfit Generator* eliminates error accumulation. In this case, only baselines with independent segmentation networks are compared. Also, ours* is an ablation study that is trained with a segmentation network and a try-on generator in two stages. The performance of all baselines is inferior due to the wrong labels in the segmentation results. In contrast, COTTON synthesizes high-quality try-on results since *Landmark-guided Transformation* provides accurate clothing region information and *Outfit Generator* is optimized globally. Comparing ours with ours*, since ours* obtains the wrong boundary for clothing labels, ours* results in a wrongly darker skirt shade. Overall, due to end-to-end global optimization, ours gets more accurate segmentation than ours* and consequently yields more visually convincing results.

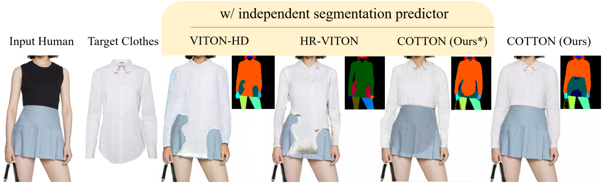

Figure 1. The visual comparison for independent segmentation network vs. end-to-end model.

B. Generalization of *CLP* and *CSN*

To alleviate the burden of annotation, we adopt a fine-tuning strategy using pre-trained backbones and data augmentation techniques. We also enforce strict clothing alignment, which rescales garment images to a uniform height and width. The accuracy of *CLP* and *CSN* trained on Pure Cotton is evaluated across datasets with each validation set consisting of 800 labeled images. PCK@0.2 [5] normalized by upper torso length and mIOU are picked as evaluation metrics. *CLP* demonstrates consistently high performance even in cross-dataset scenarios (from 99.95% to 97.44%). Though *CSN* experiences a slight performance drop (from 0.918 to 0.878) on the Dress Code dataset, this marginal drop does not significantly affect the overall quality of the final try-on results.

C. Ablation study of all designed components

Table 1 shows the ablation study on Pure Cotton dataset. The findings indicate that the proposed *Landmark-guided Transformation (LT)* plays a pivotal role in enhancing model performance. Replacing *LT* with alternative warping techniques leads to a substantial deterioration in performance. Additionally, the inclusion of the end-to-end *Outfit Generator (OG)* contributes to further improvements. However, compared to the other components, the *Clothing Elimination Policy (CEP)* shows a lower impact on model performance. This discrepancy arises as the metrics employed are not particularly sensitive to the preservation or lack thereof of crucial personal details.

Table 1. Ablation study of each component.

Warp	CEP	OG	SSIM↑	LPIPS↓	FID↓
TPS		✓	0.950	0.0460	18.33
flow		✓	0.953	0.0368	15.19
LT		✓	0.953	0.0349	10.47
LT	✓		0.955	0.0374	11.49
LT	✓	✓	**0.958**	**0.0315**	**10.17**

D. Ablation Study for Warping Methods

We compare our proposed clothing warping method with conventional learning-based warping networks, e.g., TPS [1], flow-based [6]. Fig. 2 shows that our approach has a significant advantage in handling bent-arm postures and maintaining the pattern distribution of clothing. Our crop-and-warp step (Fig. 3 in main paper) allows us to transform each piece of clothing independently, which results in better performance when dealing with complex warping. Table 2 presents a comparison of SSIM, LPIPS, and FID following the main paper setting. The results demonstrate that our *LT* outperforms TPS and appearance flow-based methods.

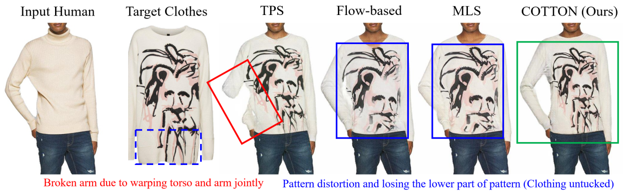
Input Human
Target Clothes
TPS
Flow-based
MLS
COTTON (Ours)
Broken arm due to warping torso and arm jointly
Pattern distortion and losing the lower part of pattern (Clothing untucked)

Figure 2. Visual comparison of different warping methods.

Table 2. Quantitative comparison of warped clothes on the test set.

Method	Dataset	SSIM↑	LPIPS↓	FID↓
TPS	Pure Cotton	0.9259	0.105	66.11
Flow-based		0.9366	0.082	48.21
Ours		**0.9387**	**0.078**	**46.26**
TPS	Dress Code	0.8624	0.173	62.20
Flow-based		0.9264	0.098	59.55
Ours		**0.9298**	**0.095**	**51.78**

E. Comparisons with Methods Equipped with Different Elimination Policies

The distinctions between our *CEP* and [3, 4] methods are twofold. Firstly, *CEP* preserves valuable human features. Secondly, *CEP* controls multi-layer clothing interaction for tucked or untucked cases. Our policy is the first to address these two challenges at once. In contrast, [3, 4] regenerates the limbs, resulting in the loss of valuable human features, as depicted in Fig. 3.

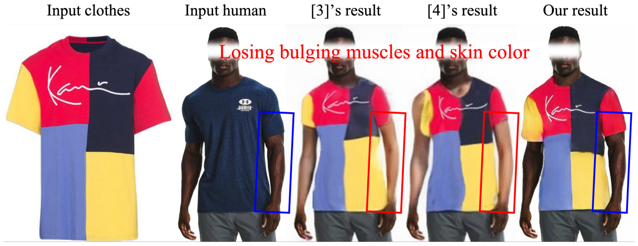

Figure 3. [3, 4] loses bulging muscles and skin color.

F. Results with complex try-on samples

F.1. Non-Frontal and Non-standing Poses

This experiments provides more complex try-on samples. The utilized pose prediction method [2] accurately predicts landmarks even for partially occluded regions, as shown in the blue box in Fig. 4, enabling sleeve transformations for non-frontal postures. In severe occlusion cases, as in the middle case, the transformation is not applied to the occluded region due to the absence of a visible sleeve. Non-standing poses are processed similarly, as shown in Fig. 4.

F.2. How to handle dress?

Dress can be handled in a similar manner to upper clothing with the lower clothing on the target human removed. Fig. 4 showcases the try-on result for dress.

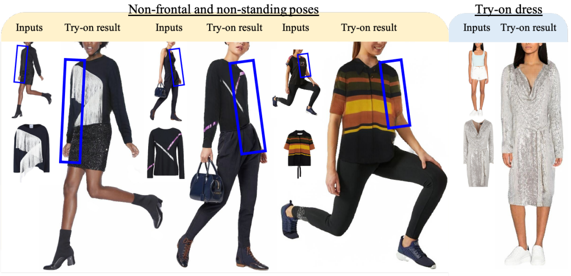

Figure 4. [Left] Our results with non-frontal and non-standing poses. [Right] Our try-on result of dresses.

G. Purified Rules for Self-collected Dataset

Specifically, as Fig. 5 shown, we strictly set several rules to filter out unsuitable data: i) the upper clothes on human image can't be multi-layer since the corresponding clothing image would only be one of the layers, ii) the clothes' sleeves on human image can't be rolled up since it would affect model to learn wrong sleeve lengths, iii) the human postures must present in frontal view with full body since the input frontal clothing image does not contain the side or back view information, etc.

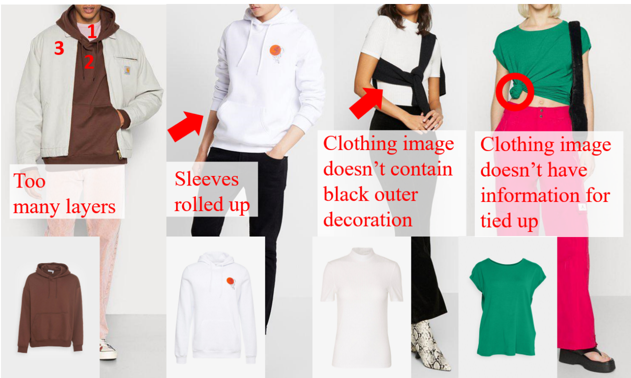
3
1
2
Too many layers
Sleeves rolled up
Clothing image doesn't contain black outer decoration
Clothing image doesn't have information for tied up

Figure 5. Noisy data samples for upper clothes.

H. Limitation

Due to the limitation of segmenting special accessories on human images, our try-on model is limited to preserve special accessories, e.g., straps and necklaces, as shown in Fig. 6. That is because the segmentation network is hard to identify whether it is a design on clothes or an individual accessory. Therefore, the segmentation model is led by the data-driven technique and may mistakenly segment the straps and necklaces as clothing designs.

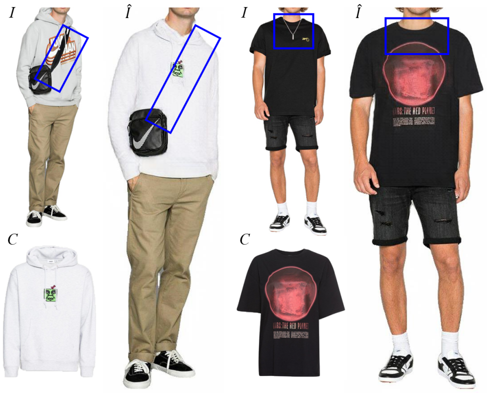

Figure 6. Our model is limited to preserve special accessories, e.g., straps and necklaces.

References

[1] Serge J. Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002. 1

[2] 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. 2

[3] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[4] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016. 1

[6] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, 2016. 1