Supplementary Material for 3D-GCL

Zaiyu Huang[1*], **Hanhui Li**[1*], **Zhenyu Xie**[1,2]
Michael Kampffmeyer[3], **Qingling Cai**[1†], **Xiaodan Liang**[1,4†]
[1]Shenzhen Campus of Sun Yat-Sen University
[2]ByteDance, [3] UiT The Arctic University of Norway
[4]Peng Cheng Laboratory
{huangzy225, xiezhy6} @mail2.sysu.edu.cn
{lihh77, caiqingl} @mail.sysu.edu.cn
michael.c.kampffmeyer@uit.no, xdliang328@gmail.com

1 Introduction

In this supplementary material, we provide the architectural details of the proposed 3D-GCL model, additional experimental details, including training details, information about the human evaluation study that was conducted to evaluate the quality of warping and virtual try-on results, ablation result for different input settings and our mIoU metric. In addition, we present additional qualitative comparisons of the virtual try-on results and the ablation study and discuss the negative impact and limitations of our work. Lastly, we include the source code of 3D-GCL, which will be made publicly available upon acceptance.

2 Criterion for HardPose Subset Construction

We first employ the calculating mechanism for pose complexity in [11] to filter out easy samples, then manually pick out the pairs that contain visually "Hard" posture from the testing set. Pairs with large viewpoint and hand position contrast are added into the testing set. Specifically, we define "HardPose" as opposite to standard posture, which is described as face forward and hands down. Visual examples from our HardPose testing set as well as "easy" cases for comparison from the full testing set are illustrated in Fig. 2. Fig. 3 shows Virtual Try-on results of our proposed method, where the examples are ordered according to difficulty. Our method is able to handle different poses and body orientations, which further validates the superiority of our method under diverse cases.

3 Innovation and Difference with ZFlow [2] and FS-VTON [7]

While [2] also incorporates 3D priors during training, we argue that there exist intrinsic differences between [2] and our 3D-GCL, in terms of the intention and the derivation of the 3D prior. In [2], the 3D prior is introduced in the Segmentation-Assisted Dense Fusion (i.e., the try-on synthesis module) by taking the DensePose as input and reconstructing it in the output. This will facilitate the synthesis network to preserve structural and geometric integrity of the try-on results as mentioned in the original paper, but also means that the 3D prior in [2] is directly derived from the input DensePose. However, in our 3D-GCL, we innovatively employ the 3D prior to provide precise guidance to when learning the correspondence, which allows the warping module to preserve the garment texture even for challenging poses. Besides, the 3D prior of our 3D-GCL is derived from the 3D vertex correspondence between the SMPL model of the same person under various poses.

[*]Both authors contributed equally.
[†]Corresponding authors. Our code will be available at 3D-GCL.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

On the other hand, although [7] proposes a global flow estimation module for garment deformation, it does not explicitly model the global correspondence between the source garment feature and the target pose feature. Specifically, [7] utilizes the style vector to modulate the weights of the StyleGAN-based network, where the style vector is obtained by concatenating the 1-D garment vector and the 1-D person vector. However, such a 1-D global style vector just provides the flow estimation network with the global information of the garment and person, rather than the global correspondence between the source and target feature. Instead, our 3D-GCL explicitly models the global correspondence between the garment and the person features by calculating the correspondence matrix in the low-resolution block and uses it as initial state for the high-resolution flow estimating blocks.

4 Architecture and Implementation Details

Pose-Aware Feature Encoder. The Pose-aware Feature Encoder is introduce to obtain high-level feature representations of the given inputs. The two encoders are responsible for encoding the source and target representations, respectively, and share the same architecture apart from the first convolutional layer. Architecture details of the two Pose-aware Feature Encoders are listed in Tab.3.

Garment Transfer Appearance Encoder. We modified the structure of the original Garment Transfer Appearance Encoder in [1] to match the setting of Virtual Try-on. Specifically, we add a face encoding branch to encode the identity information of the target model. Then, We replace the pose mask in the original decoder with obtained face feature maps, concatenating with previous appearance feature maps for each resolution. The architecture of the face encoding branch is provided in Tab.4

GACRM. In our GACRM, the SMPL Flow acts as the pseudo ground truth and is calculated according to the correspondence relationship determined by the predicted human mesh. We implement a parallel version of the SMPL Flow calculation based on source code of [8], so that the ground truth flow can be computed during Stage I training and does not have to be pre-computed. We also utilize coordinate mirroring in our SMPL Flow calculation similar to [1], which provides additional supervision signals to the commonly existing invisible areas in complex scenarios during the warping process. We adopt a feature pyramid of four layers (i.e., $L = 4$ and four GACRMs are used in our network) with spatial resolutions of $\{64 \times 64, 128 \times 128, 256 \times 256, 512 \times 512\}$, respectively.

5 Experiments Details

5.1 Experiment Setups

While we load the pretrained models directly for Pose-with-Style [1] and use the results provided by the authors of wFlow [4], we train all the other baseline models including PBAFN [5] and Dior [3] from scratch using our training data. To ensure a fair comparison, we additionally finetune the pretrained model of Pose-with-Style for 10 epochs for our experiments on the MPV dataset. Note that results of wFlow in our experiment are provided by the authors and we are therefore unable to finetune the pretrained model on our own setting.

5.2 Training Details

In this section, we elaborate the remaining experimental settings that were not covered in the main paper. In particular, we use an Adam optimizer with β_1=0.9, β_2=0.999 and set the learning rate to 0.0001. We train our network for around a day using 4 Tesla V100 GPUs. For hyperparameter settings, we first choose the parameters to ensure that different losses have the same order of magnitude. Then we further fine-tune the parameters by validating the performance of our model on a collected validation set. Finally, λ_r, λ^c_{l1}, λ^c_{perc} are set to 20, 5, 5, respectively. Besides, as mentioned in the paper, we follow the setting of [1] for our generator training, which means that λ_{adv}, $\lambda^g_{l_1}$, and λ^g_{perc} are set to 1, 1, 1, respectively.

5.3 Human Evaluation Details

We conduct human evaluations on DeepFashion, MPV and our collected HardPose subset. 40 volunteers are invited to fill in a questionnaire composed of 20 questions for each dataset and asked

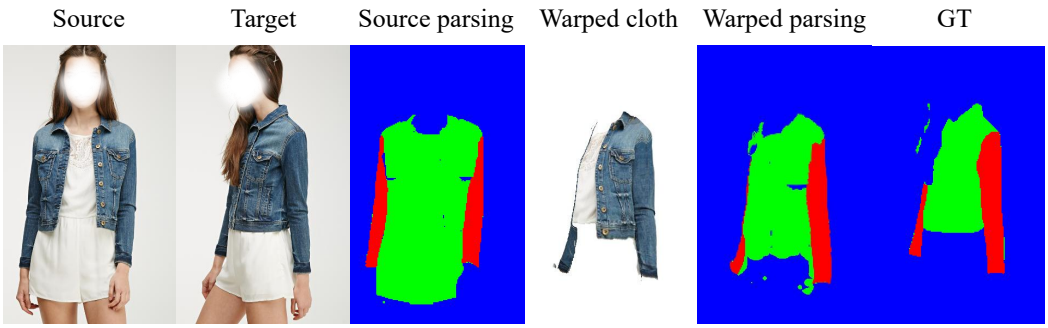

Figure 1: Visualization of an image pair and its corresponding source parsing, warped parsing and gt.

Method	mIoU ↑	LPIPS ↓	FID ↓
Image Input	73.04	0.1876	10.74
IUV Input	**75.76**	**0.1725**	**10.58**

Table 1: Ablation study of different input settings on DeepFashion [9].

to pick out the most visually compelling synthesis result among the 5 provided pictures that were produced by our approach and the four baseline methods (presented in random order). Specifically, for the warped garment evaluation, we provide the ground-truth warped garments in the questionnaire and ask the participants to pick out the best result considering the similarity to the given image, while for the evaluation of the final try-on results, we provide the source and target model images together with different try-on synthesis results and ask the volunteers to choose the most realistic results considering preservation of garment texture details and target person identity. Fig.8 and Fig.9 illustrates the interface of these two questionnaires on the DeepFashion dataset.

5.4 Ablation Study for Different Input Settings

We conduct additional experiments on the DeepFashion dataset to validate the effectiveness of taking IUV maps as input in our pipeline. Concretely, we replace the IUV input with sparse keypoint maps in our Pose-Aware Feature Encoder and take the source image as input for the source encoding branch. Tab.1 demonstrates the comparison between these two settings on the DeepFashion datset.

5.5 mIoU Metric

In the experiments, we adopt the mIoU metric to evaluate the semantic correctness of our learned flow. It is worth noting that currently public available human parsers are only able to distinguish different garments e.g. coats, pants, dress as well as skin. Thus it is impossible to accurately access warping mistakes where one part of a garment occludes a different part of itself. One example of this that typically occurs for difficult poses is that part of the sleeve occludes the rest of the upper-body garment. To address this problem and facilitate more accurate evaluation, we pretrain a fine-grained human parser to split the garments image into three semantic regions: background, body, and sleeve. This allows us to calculate the mIoU between the warped semantic maps and the corresponding ground truth. Fig.1 illustrates the warped semantic maps and the corresponding ground truth for an example image.

6 Analysis of Potential Negative Social Impacts and Limitations

Potential negative social impact The proposed method might be applied to malicious image manipulations, as most current synthesis methods. Nevertheless, such negative impact could be alleviated via forensics analysis and other manipulation detection methods.

Limitations As our primary target is to learn 3D-aware global correspondences for warping, we do not devise an inpainting module to generate unseen garment textures. Thus a try-on generator that has a certain capability of inpainting is required in our method. Visual results of some failure cases

Figure 2: Visual comparison between "easy" cases (left) in the full testing set and "hard" cases (right) in our HardPose subset.

Figure 3: Synthesized images produced by our model on the testing set of MPV, where examples are ordered according to difficulty (easy to hard, left to right).

are illustrated in Fig.4. Additionally, our model is unable to handle parsing errors and wrong pose estimation. These errors could potentially be addressed by using knowledge distillation methods to exclude non-image inputs. Fusing temporal information from a video is also a interesting future direction to deal with these problems.

7 Additional Results

7.1 Effects of Filtering Out Samples without Densepose/IUV Maps

We filter out the cases where no human is detected at all by DensePose [6] and SMPL [10] as these can be considered as outliers when regarding the global distribution of the dataset. Totally 345 of the original 101,967 pairs in the training set and 6 of the original 8,570 pairs in the testing set for Deepfashion are excluded from the testing set. To ensure that this setup has a negligible effect on the overall performance and does not lead to an unfair experimental setup, we report the result of our method and the closest competitor [1] on the full testing set of Deepfashion in Tab. 2.

Figure 4: Failure cases of our 3D-GCL network.

Method	Full			Filtered		
	mIoU ↑	LPIPS ↓	FID ↓	mIoU ↑	LPIPS ↓	FID ↓
Pose with style[1]	62.49	0.1998	16.35	62.74	0.1997	16.35
Ours	75.91	0.1725	10.58	75.76	0.1725	10.58

Table 2: Comparison results of our method and [1] on the full and the filtered testing set of Deepfashion.

7.2 More Examples for Virtual Try-on

Fig.5 and Fig.6 show additional visual comparison results for our proposed 3D-GCL network and the baseline methods on the DeepFashion and MPV datasets.

7.3 More Examples for Ablation Study

Fig.7 displays additional examples for the ablation study on the DeepFashion and MPV datasets.

7.4 Source Code and HardPose Dataset

The source code including the HardPose data split can be found in the zipped supplementary file.

Table 3: The architecture details of the Pose-Aware Feature Encoder.

Pose-Aware Feature Encoder		
Layer	**Type**	**Output Size**
Input	Input	(512,512,-)
Conv	Conv2dLayer 3×3, InstanceNorm	(512,512,32)
Enc1	Conv2dLayer 3×3, InstanceNorm	(512,512,64)
	ResBlock 3×3, PReLU	(512,512,64)
Enc2	Conv2dLayer 3×3, down=2, InstanceNorm	(256,256,128)
	ResBlock 3×3, PReLU	(256,256,128)
Enc3	Conv2dLayer 3×3, down=2, InstanceNorm	(128,128,128)
	ResBlock 3×3, PReLU	(128,128,128)
Enc4	Conv2dLayer 3×3, down=2, InstanceNorm	(64,64,128)
	ResBlock 3×3, PReLU	(64,64,128)
Enc5	Conv2dLayer 3×3, down=2,InstanceNorm	(32,32,128)
	ResBlock 3×3, PReLU	(32,32,128)

Table 4: The architecture details of the Face Encoding Branch.

Face Encoding Branch		
Layer	**Type**	**Output Size**
Input	Input	(512,512,3)
Conv	Conv2dLayer 3×3, LeakyReLU	(512,512,64)
Res1	ResBlock 3×3, down=2, LeakyReLU	(256,256,64)
Res2	ResBlock 3×3, down=2, LeakyReLU	(128,128,64)
Res3	ResBlock 3×3, down=2, LeakyReLU	(64,64,64)
Res4	ResBlock 3×3, down=2, LeakyReLU	(32,32,64)

Figure 5: Additional virtual try-on comparison results between our proposed 3D-GCL network and baseline methods on the DeepFashion dataset.

Figure 6: Additional virtual try-on comparison results between our proposed 3D-GCL network and baseline methods on the MPV dataset.

Figure 7: Additional virtual try-on results under the different settings of the ablation study.

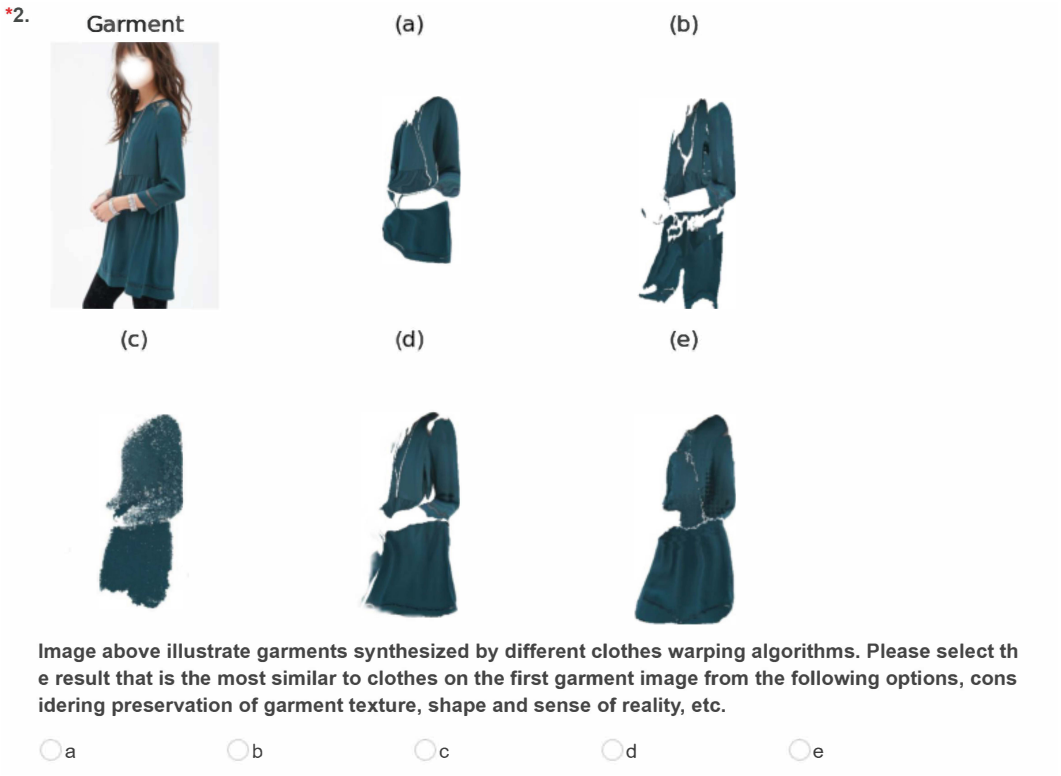

Figure 8: Screenshot of the questionnaire used to evaluate garment warping on the DeepFashion dataset.

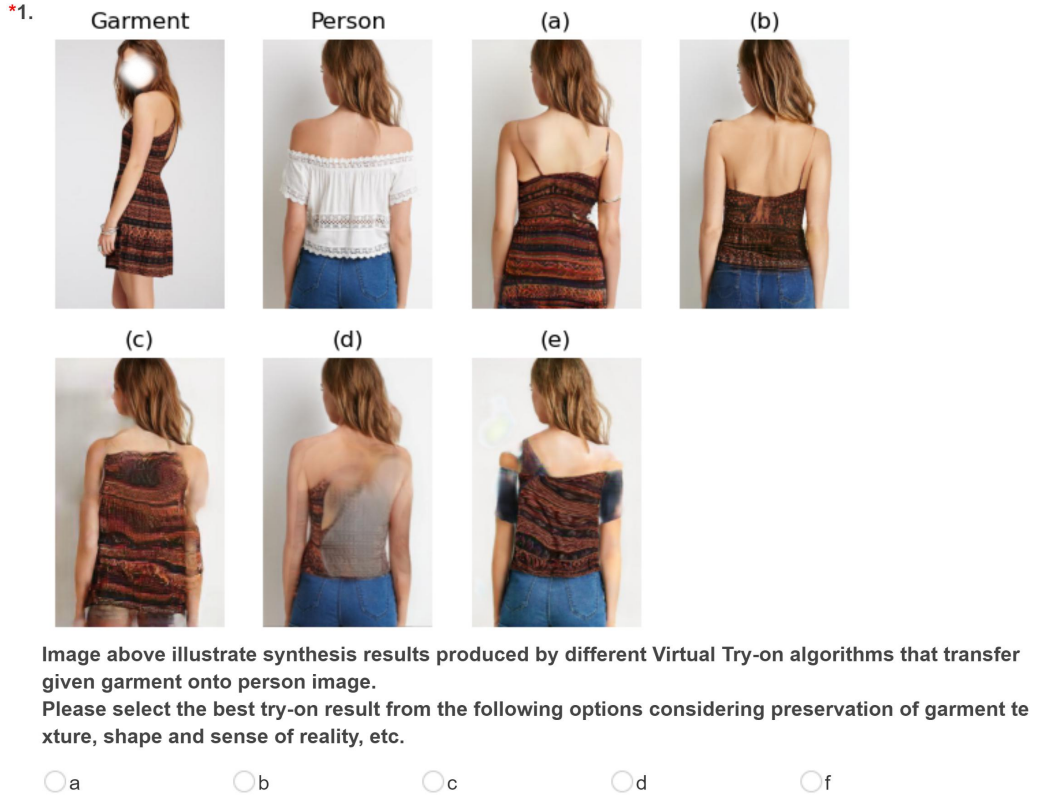

Figure 9: Screenshot of the questionnaire used to evaluate the final try-on results on the DeepFashion dataset.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics*, 40(6):1–11, 2021.

[2] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[3] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[4] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K. Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2021.

[6] 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*, pages 7297–7306, 2018.

[7] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, 2022.

[8] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[9] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.