Weakly Supervised High-Fidelity Clothing Model Generation Supplementary Materials

Ruili Feng[1], Cheng Ma[2,3], Chengji Shen[2,3], Xin Gao[3], Zhenjiang Liu[3],
Xiaobo Li[3], Kairi Ou[3], Deli Zhao[3], Zheng-Jun Zha[1*]
[1]University of Science and Technology of China, Anhui, China
[2]Zhejiang University, Hangzhou, China
[3]Alibaba Group, Hangzhou, China
ruilifengustc@gmail.com, {cheng.ma,chengji.shen}@zju.edu.cn,
{zimu.gx,stan.lzj,xiaobo.lixb}@alibaba-inc.com,
{mailokr,zhaodeli}@gmail.com, zhazj@ustc.edu.cn

Appendix

A Proof to Theorem 1

Theorem 1 *Assume that $\mathcal{W}+$ follows the multi-variable Gaussian distribution, then the output of the projector $\boldsymbol{P}$ will always fall in the high-density region of $\mathcal{W}+$, which is an n-dimensional ellipse $\mathcal{E}$ with axes $\boldsymbol{q}_1, ..., \boldsymbol{q}_n$, and axis lengths $\psi\sigma_1^{\frac{1}{2}}, ..., \psi\sigma_n^{\frac{1}{2}}$. Rigorously, let ω_{n-1} denote the volume of the $n-1$ dimensional unit ball, for a random sample $\boldsymbol{w}$ from $\mathcal{W}+$, the possibility of it outside $\mathcal{E}$ is*

$$\mathbb{P}(\boldsymbol{w} \notin \mathcal{E}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\psi}^{\infty} \omega_{n-1} r^{n-1} e^{-\frac{1}{2}r^2} dr, \tag{S1}$$

which drops to zero drastically as ψ grows larger; and for an arbitrary input $\boldsymbol{x}$, we have

$$\boldsymbol{P}(\boldsymbol{x}) \in \mathcal{E} = \{\boldsymbol{w} : (\boldsymbol{w}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}-\boldsymbol{\mu}) \leq \psi^2\}. \tag{S2}$$

Proof to this theorem yields three parts: proving that the output of the projector $\boldsymbol{P}$ always falls in $\mathcal{E}$ (Eq. (S2)), proving Eq. (S1), and demonstrating that it decreases to zero as ψ grows.

A.1 Proof to Eq. (S2)

Recall the computation of the Projection in Sec. 4.1. Given a rough alignment image $\boldsymbol{x}_a$, we first train an encoder $\boldsymbol{E}$ to get the strength code $\boldsymbol{s}$

$$\boldsymbol{s} = \boldsymbol{E}(\boldsymbol{x}_a). \tag{S3}$$

Then we truncate the strength code $\boldsymbol{s}$ to an ellipse centered at $\boldsymbol{\mu}$ with radius ψ:

$$\boldsymbol{w}_0 = Tr(\boldsymbol{q_1} s_1 \sqrt{\sigma_1} + ... \boldsymbol{q}_n s_n \sqrt{\sigma_n}) + \boldsymbol{\mu} = \boldsymbol{Q}\boldsymbol{\Lambda}^{\frac{1}{2}} Tr(\boldsymbol{s}) + \boldsymbol{\mu}, \tag{S4}$$

where Tr is a truncation operator with cutoff coefficient $\psi > 0$ such that

$$Tr(\boldsymbol{v}) = \begin{cases} \boldsymbol{v}, \|\boldsymbol{v}\|_2 < \psi, \\ \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\psi, \|\boldsymbol{v}\|_2 \geq \psi. \end{cases} \tag{S5}$$

For any $\boldsymbol{s}$, it is then easy to see that

$$(\boldsymbol{Q}^{-1}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{w}_0 - \boldsymbol{\mu}))^T (\boldsymbol{Q}^{-1}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{w}_0 - \boldsymbol{\mu})) = (\boldsymbol{w}-\boldsymbol{\mu})^T \boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T(\boldsymbol{w}-\boldsymbol{\mu}) = Tr(\boldsymbol{s})^T Tr(\boldsymbol{s}) \leq \psi^2. \tag{S6}$$

As $\boldsymbol{Q}$ is computed from PCA decomposition, we have

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I} \Leftrightarrow \boldsymbol{Q}^{-1} = \boldsymbol{Q}^T, \tag{S7}$$

$$\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T. \tag{S8}$$

Take Eq. (S7) & (S8) to (S6), we then have

$$(\boldsymbol{w}_0 - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}_0 - \boldsymbol{\mu}) \leq \psi^2, \tag{S9}$$

which verifies Eq. (S2).

A.2 Proof to Eq. (S1)

Let assume that $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu_w}, \boldsymbol{\Sigma_w})$. The Law of Large Numbers [3] then tells that, for a collection of *i.i.d* sampling $\{\boldsymbol{w}_i\}_{i=1}^N$ from the $\mathcal{W}+$ space, we have

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_i \rightarrow \boldsymbol{\mu_w}, \ as \ N \rightarrow \infty, \tag{S10}$$

$$\boldsymbol{\Sigma} = \frac{1}{N-1}(\boldsymbol{w}_1 - \boldsymbol{\mu}, ..., \boldsymbol{w}_n - \boldsymbol{\mu})^T (\boldsymbol{w}_1 - \boldsymbol{\mu}, ..., \boldsymbol{w}_n - \boldsymbol{\mu}) \rightarrow \boldsymbol{\Sigma_w}, \ as \ N \rightarrow \infty. \tag{S11}$$

Considering that we sample five million points to compute the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, N is large enough to let us assume that

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{w}}. \tag{S12}$$

Thus we have $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It is then easy to see that

$$\hat{\boldsymbol{w}} = \sqrt{\boldsymbol{\Sigma}^{-1}}^{T}(\boldsymbol{w} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathcal{O}, \mathcal{I}), \tag{S13}$$

where $\sqrt{\boldsymbol{\Sigma}^{-1}}\sqrt{\boldsymbol{\Sigma}^{-1}}^{T} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$, and $\mathcal{N}(\mathcal{O}, \mathcal{I})$ is the Standard Gaussian distribution. Note that

$$\boldsymbol{w} \notin \mathcal{E} \Leftrightarrow (\boldsymbol{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) > \psi^2 \Leftrightarrow \hat{\boldsymbol{w}}^T \hat{\boldsymbol{w}} > \psi^2 \Leftrightarrow \hat{\boldsymbol{w}} \notin B(0, \psi). \tag{S14}$$

Thus we have

$$\mathbb{P}(\boldsymbol{w} \notin \mathcal{E}) = \mathbb{P}(\hat{\boldsymbol{w}} \notin B(0, \psi)). \tag{S15}$$

As $\hat{\boldsymbol{w}}$ follows the Standard Gaussian, we know that the sum of squares of all its element follows the n-dimensional chi-square distribution [3, 10],

$$\hat{\boldsymbol{w}}^T \hat{\boldsymbol{w}} \sim \chi_n^2. \tag{S16}$$

Thus we have

$$\mathbb{P}(\boldsymbol{w} \notin \mathcal{E}) = \mathbb{P}(\hat{\boldsymbol{w}} \notin B(0, \psi)) = \mathbb{P}(\chi_n^2 > \psi^2). \tag{S17}$$

A.3 Proof to Eq. (S1) Decreasing to Zero

This is easy to see as it is well known that the tail of chi-square distribution falls to zero quickly. Here we quote the tail bound of chi-square distribution in [11] (Lemma 1 p1325), that

$$\mathbb{P}(\chi_n^2 > n + 2\sqrt{nt} + 2t) < e^{-t}, \forall t \in \mathbb{R}_+. \tag{S18}$$

When ψ is large enough, we further have

$$\mathbb{P}(\chi_n^2 > \psi^2) < e^{-\frac{\psi^2}{10}}. \tag{S19}$$

Both the above two estimations verify that the tail of chi-square distribution falls to zero drastically.

B Numerical Measurements of Each DGP Components

Table S1. Numerical metrics of the Projection, Semantic Search, and Pattern Search steps of the DGP method on the CMI dataset, at the resolution of 512×512. $\downarrow$ indicates lower is better.

Metrics	Projection	Semantic Search	Pattern Search	DGP (final output)
FID$\downarrow$	54.2	52.9	58.9	58.9
SWD$\downarrow$	56.7	45.7	46.5	46.5

C Rough Alignment

The rough alignment step uses a collection of model key points together with a collection of clothes key points to compute the alignment. The collection of model key points include points of neck, shoulder, hip, elbow, and wrist. Among them, key points of elbow and wrist are inherited from the corresponding annotations in the COCO [12] dataset for human poses; key points of neck, shoulder, and hip are acquired from an API interface from an anonymous online AI platform. Fig. S1 and Tab. S2 demonstrate more details of those key points. We feed images from the COCO dataset to the anonymous API interface to obtain the labels for neck, shoulder, and hip, and merge them with the original labels in COCO to obtain the ground truth key point annotations for training. An HRNet [13] model is trained on those images to predict these key points. The HRNet model reaches 67.2 AP on the test set after convergence.

Table S2. Definition of model key points.

Index	Name	Index	Name	Index	Name	Index	Name
1	left neck	5	left wrist	9	right knee	13	right elbow
2	left collarbone	6	left hip	10	right thigh	14	right shoulder
3	left shoulder	7	left thigh	11	right hip	15	right collarbone
4	left elbow	8	left knee	12	right wrist	16	right neck

Table S3. Definition of clothing key points.

Category	Index	Name	Index	Name	Index	Name	Index	Name
Sling	1	left collarbone	2	left hip	3	right hip	4	right collarbone
Undershirt	1	left collarbone	2	left hip	3	right hip	4	right collarbone
Short sleeve top	1	left shoulder	2	left hip	3	right hip	4	right shoulder
Long sleeve top	1	left neck	2	left hip	3	right hip	4	right neck
Long sleeve outwear	1	left neck	2	left hip	3	right hip	4	right neck
Windbreaker	1	left neck	2	left hip	3	right hip	4	right neck

Figure S1. Definition of model key points and clothing key points with six basic categories.

The collection of clothes key points are inherited from the DeepFashion2 [6] dataset. Again, an HRNet model is trained on the DeepFashion2 dataset, with 56 AP on the test set after convergence. Key points such as neckline, shoulder, and bottom corner of the clothing are selected to perform the perspective transformation, the else are used to guide the As Rigid As Possible (ARAP) [1, 8] transformation. Fig. S1 and Tab. S3 illustrate this process.

After getting those key points, the rough alignment can be split into two phases: 1) a perspective transformation, followed by 2) an ARAP transformation. We use the OpenCV function *cv2.getPerspectiveTransform* to compute the transformation matrix $\boldsymbol{A}$ that aligns the clothing key points at the left and right boundaries of the neckline and the bottom corner to the model key points at the left and right boundaries of the neck and the hip, correspondingly. Another OpenCV function *cv2.WarpPerspective* is used to carry out the alignment of clothing and model image under transformation $\boldsymbol{A}$, and generate the result of the whole perspective transformation phase. Fig. S2 (a) illustrates this process. For different types of clothes, the key points we used admit subtle differences, Tab. S4 reports those differences. Specifically, we omit the ARAP transformation for sleeveless and short-sleeve tops, thus the result of the perspective transformation is the final result of the rough alignment. For the long-sleeve type, we further use ARAP to align limbs, as shown in Fig. S2 (b). ARAP offers an interface to compute an energy minimized isotropy deformation given the offsets of couples of control points. Here we use the four skeleton points of the elbows and wrists of the clothing as the control points, and the four skeleton points of the elbows and wrists of the model image mentioned above as their destinations, correspondingly. Key points of the neck and hip, which have been aligned in the previous perspective transformation phase, are also added to the pool of control points as the immovable control points, which will be fixed during the ARAP deformation. Then the ARAP will roughly align the sleeves of the clothing to the arms of the model, and keep the body part of the clothing fixed. We take this as the final output of the rough alignment for long-sleeve

categories such as jackets and sweaters.

Table S4. Mapping rules of the perspective transformation .

Clothing type	Points mapping rules (clothing-model)
Sling	1-2, 2-6, 3-11, 4-15
Undershirt	1-2, 2-6, 3-11, 4-15
Short sleeve top	1-3, 2-6, 3-11, 4-14
Long sleeve top	1-1, 2-6, 3-11, 4-16
Long sleeve outwear	1-1, 2-6, 3-11, 4-16
Windbreaker	1-1, 2-6, 3-11, 4-16

Figure S2. Illustration of (a) the perspective transformation and (b) the ARAP, the index numbers of the corresponding key points on the model have been marked.

D Attribute Classifier

The optimization process of inversion on human face images often uses identity loss to help in recovering detailed information of the face. Similarly, the training loss for the Encoder includes the attribute similarity loss $\mathcal{L}_{attr}$, which is captured by a pretrained clothing attribute classifier $\boldsymbol{R}$ trained on the FashionAI dataset [15]. We select 7 category dimensions (neck, collar, lapel, neckline, sleeves length, skirt length, top length), 49 tags in total, from the FashionAI dataset, and train an attribute classification model based on the ResNet50 architecture. The model is trained on 4 Tesla V100 GPUS, with an mAP score of 0.95 and an accuracy score of 0.84 after convergence. The final convolution layer is taken as the feature space to compute the similarity between the generated image and the rough alignment image.

Figure S3. Dynamic spatial weight $\boldsymbol{W}$ in Semantic Search.

E Dynamic Spatial Weight

Fig. S3 illustrates the isosurface of the Dynamic Spatial Weight $\boldsymbol{W}$ in Semantic Search and Pattern Search, which is computed according to Eq. (18) of the paper context. Deeper red color denotes the higher weight, and the black background

denotes zero weight. The distance function $d((i,j),\partial I)$ for an arbitrary pixel point in position (i,j) is computed based on the erosion operation (*cv2.erode*) in OpenCV Library. We recurrently conduct erosion operation to the region I with 3×3 kernel size. In the k-th step, if the spatial position (i,j) gets eroded, we set $d((i,j),\partial I) = k$. For spatial position (i,j) at the boundary of the region I or outside the region I, we set $d((i,j),\partial I) = 0$.

F The E-Shop Fashion (ESF) Dataset

The E-Shop Fashion (ESF) dataset contains 180,000 clothing model images collected from an anonymous e-commercial website under legal circumstances. To ensure the purity of the dataset, we filter out images with complex backgrounds and models holding bags or other items. Each image contains only one full-body, frontal standing pose model. A detection model is used to get the bounding box of the model, and are center aligned with resize and padding operations. The images are all cropped to the region between jaw and thigh, and resized to the resolution of 512 × 512. Fig. S4 gives the category names and the corresponding quantities, and Fig. S5 shows some examples of the ESF dataset. The whole dataset contains 180,000 images, which is further split into 170,000 training samples and 10,000 testing samples. This dataset will be open-sourced later.

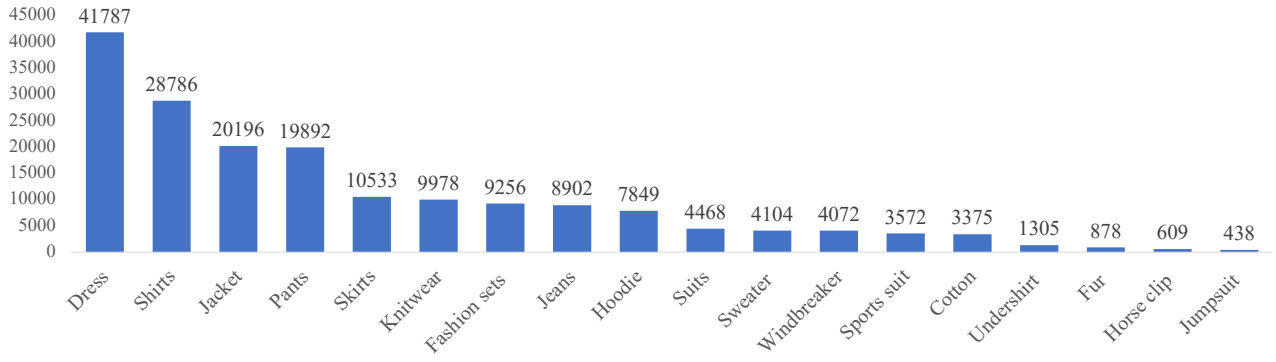

Figure S4. Number of each category on the ESF dataset. The total number is 180,000.

G The Commercial Model Image (CMI) Dataset

The Commercial Model Image (CMI) dataset contains 2,348 images taken by 200 models on underwear, including different genders, ages, body shapes, and poses. All model images are taken in professional studios, and all models signed the confidentiality and authorization statement file. Fig. S7 shows some models of the CMI dataset. Another part of the CMI dataset is 1,881 clothing images with clean backgrounds from an anonymous e-commercial website under legal circumstances, which evenly contains 16 categories of top clothing, including 10 categories of women's tops and 6 categories of men's tops. Each category contains 120 images, excepting woman's leather jacket that only contains 81 images. Fig. S8 shows examples of these 16 categories.

H License of the CMI Dataset

Since the CMI dataset contains 2,348 images taken by 200 models on underwear, authorization and privacy issues must be considered. Each model is asked to sign a confidentiality and authorization statement file. The contents of the file include authorization, confidentiality and privacy statement. This file named CONFIDENTIALITY AND AUTHORIZATION STATEMENT.pdf can be found in the zip archive, please read this file for more details. In addition, to protect the models' privacy, the face areas are blurred for all the images. We cropped the face part during the pre-processing step to ensure that all figures in the paper and the supplementary materials do not contain faces.

I User Study

We conduct user studies on our CMI benchmark and MPV [4] dataset by recruiting 50 volunteers. The proposed DGP method is compared with three state-of-the-art supervised methods, VITON-HD [2], PF-AFN [5, 7] and ACGPN [14]. For

Figure S5. Examples of the ESF dataset.

Figure S6. Fake images generated by StyleGAN2. The final FID is 2.16.

each model image of the CMI dataset, we randomly pick up a garment image from the 1,881 clothing images of CMI. It yields a testing set of 2,348 model and clothing pairs. For the MPV dataset, We pick 1,476 image pairs of person and clothing to construct the testing set. Each sample of the testing set contains six images, i.e., a target clothing, a reference person image, together with the try-on results of three compared methods and our DGP method. The order of the four try-on results is randomly shuffled. All 50 volunteers are required to mark all samples of the test set, and are asked to answer the following questions: 1) which method generates the clearest pattern; 2) which method generates the most realistic wearing; 3) and which

Figure S7. Examples of model images on the CMI dataset.

Figure S8. Examples of 16 categories of clothing images on the CMI dataset.

method generates the best overall effect. For each sample, the method with the highest number of votes will be the winner.

J Hyper-parameter Table

Please refer to Tab. S5 for hyper-parameter selection of training and optimization objectives.

K Numerical Results on 128 and 512 Resolutions

Please refer to Tab. S6 for numerical metrics of DGP, ACGPN, PF-AFN, and VITON-HD on CMI and MPV datasets.

Table S5. Hyper-parameter selection of training and optimization objectives.

Components	Loss	Learning Rate	Terminating Condition
StyleGAN $\boldsymbol{G}$	The same as [9]	$2.5e-3$	1,250,000 iterations
StyleGAN $\boldsymbol{D}$	The same as [9]	$2.5e-3$	1,250,000 iterations
Projector $\boldsymbol{P}$	$\lambda_p = 1.0, \lambda_f = 5e-5, \lambda_{attr} = 5e-5, \lambda_{adv} = 0.1, \psi = 6.$	$2e-5$	562,500 iterations
Semantic Search	$\eta_p = 1.0, \eta_f = 5e-5, \eta_{attr} = 5e-5, \eta_{adv} = 1.0.$	$1e-2$	1,000 PGD iterations
Pattern Search	$\eta_p = 1.0.$	$1e-2$	1,000 PGD iterations

Table S6. Numerical metrics of DGP, ACGPN, PF-AFN, and VITON-HD on CMI and MPV datasets. ↓ indicates lower is better.

Methods	**CMI**		**MPV**	
	FID↓	**SWD↓**	**FID↓**	**SWD↓**
512×512 resolution				
ACGPN	137.9	121.3	81.1	90.4
PF-AFN	97.3	76.7	67.8	67.1
VITON-HD	87.5	56.1	**40.6**	52.7
DGP (Ours)	**51.6**	**22.4**	48.4	**36.7**
128×128 resolution				
ACGPN	115.7	68.4	48.0	39.2
PF-AFN	86.6	29.4	49.5	**24.9**
VITON-HD	95.0	27.4	**44.7**	25.6
DGP (Ours)	**56.5**	**18.8**	46.8	29.9

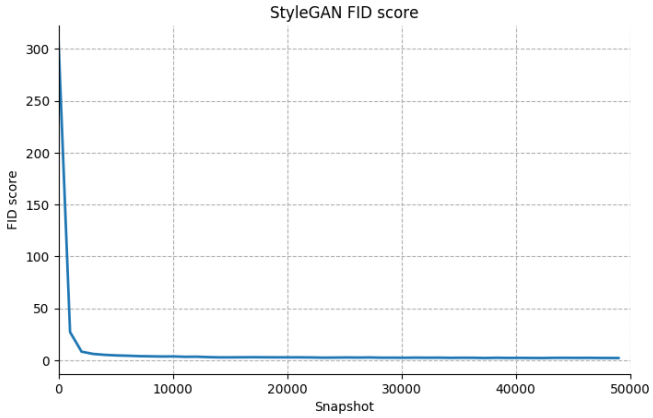

Figure S9. FID score during training.

L Training Details of StyleGAN2

The official StyleGAN2-ADA [9] is used to train the StyleGAN2 model. We train StyleGAN2-ADA on 8 NVIDIA A100 Tensor Core GPUs, where we set the hyper-parameter *batch_size* as 64, *ema_kimg* as 20.0, and *ema_rampup* as 0.05. The resolution of the generated images is 512×512, the ADA options of brightness, contrast, and saturation are turned on, and horizontal flipping is allowed. The optimizer of the generator is Adam, the learning rate is 0.0025, and the discriminator settings are the same. The loss function remains the same as in the official StyleGAN2-ADA. The training stops with 48,988,000 images, with an early stop strategy, and the final FID is 2.16. The FID score during training is shown in Fig. S9, and samples of generated images are shown in Fig. S6.

Figure S10. More visual results of qualitative comparison on the CMI dataset.

Figure S11. More visual results of qualitative comparison on the MPV dataset.

Figure S12. Performance under different optimization steps. Best viewed in large size.

Figure S13. Robustness of the DGP method. The mistakes like missing parts of clothing, wrong key point alignments, and zigzag clothing boundaries are easily corrected in the final results.

M Qualitative Comparison on CMI and MPV Dataset

More qualitative results on CMI and MPV datasets are reported of the proposed weakly supervised DGP method compared with the other three supervised competitors, VITON-HD, PF-AFN, and ACGPN. Please refer to Fig. S10 for more details on the CMI dataset and Fig. S11 on the MPV dataset.

N Limitations

The major limitation of the DGP algorithm is the time cost of semantic and pattern searches. The current setting costs around 1 minute to finish the wearing of a single batch of data with one GTX 1080Ti GPU. Shortening the optimization steps may lower the overall performance, as reported in Fig. S12. It is acceptable for clothing model generation where we do not need real-time and interactive computation of results, but accelerating the computation can certainly benefit the method to apply on broader occasions. Also, the common difficulty of existing SOTA VTO methods in handling extremely complicated poses still remains in the proposed method.

O Robustness of DGP

The imagination ability in Sec. 4.1 of the projector is very appealing for the clothing model generation. So this section further investigates how this ability can help the DGP method overcome mistakes in the preprocessing period. We deliberately feed the DGP method with flawed rough alignment images, such as missing parts of clothing, wrong key point alignments, and zigzag clothing boundaries. We then observe how the DGP method will perform with those mistakes. The results are reported in Fig. S13, which confirm that DGP can easily correct these tiny mistakes, and yield realistic synthesis in the final results.

References

[1] Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 157–164, 2000. 4

[2] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14131–14140, 2021. 6

[3] Kai Lai Chung and Kailai Zhong. *A course in probability theory*. Academic press, 2001. 2, 3

[4] 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 *Int. Conf. Comput. Vis.*, pages 9026–9035, 2019. 6

[5] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8485–8493, 2021. 6

[6] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5337–5345, 2019. 4

[7] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Int. Conf. Comput. Vis.*, pages 10471–10480, 2019. 6

[8] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM Trans. Graph.*, 24(3):1134–1141, 2005. 4

[9] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 9

[10] Henry Oliver Lancaster and Eugene Seneta. Chi-square distribution. *Encyclopedia of Biostatistics*, 2, 2005. 3

[11] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 3

[13] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 3

[14] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7850–7859, 2020. 6

[15] Xingxing Zou, Xiangheng Kong, Waikeung Wong, Congde Wang, Yuguang Liu, and Yang Cao. FashionAI: A hierarchical dataset for fashion understanding. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, pages 0–0, 2019. 5