SCIENCE CHINA
Information Sciences

• LETTER •

September 2023, Vol. 66 199101
https://doi.org/10.1007/s11432-022-3675-1

NeuralReshaper: Single-image Human-body Retouching with Deep Neural Networks

Beijia Chen[1], Yuefan Shen[1], Hongbo Fu[2], Xiang Chen[1], Kun Zhou[1] & Youyi Zheng[1*]

[1] *State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China;*
[2] *School of Creative Media, City University of Hong Kong, Hong Kong, China*

Received 8 July 2022/Revised 7 October 2022/Accepted 29 November 2022/Published online 1 August 2023

Citation Chen B J, Shen Y F, Fu H B, et al. NeuralReshaper: Single-image Human-body Retouching with Deep Neural Networks. Sci China Inf Sci, 2023, 66(9): 199101, https://doi.org/10.1007/s11432-022-3675-1

Semantic retouching of human bodies in images, such as increasing the height and slimming the body, has been long desired. However, the problem is essentially ill-posed because one should anticipate a set of articulated and nonrigid deformations of different body parts, given that the deformations are inherently three-dimensional. This situation becomes more complicated when images are captured in unrestricted environments with occlusions, and complex interactions between the human body and its surroundings. Early attempts [24] have attempted to address this issue by interactively fitting a 3D parametric human model to human bodies in images and allowing the fitted 3D model to delegate the transformation via image warping. Although compelling results are produced, these methods may suffer from laborious interactions and foreground and background distortions.

Inspired by the impressive synthesized images from GANs [12], we present NeuralReshaper (Fig. 1), the first self-supervised learning-based method for realistic human-body reshaping in a single RGB image following a fit-then-reshape paradigm. The fitting process was first automated using a hybrid learning-and-optimization-based method with the skinned multiperson linear (SMPL) [16] model. Then, in an essential stage, the 3D geometric deformation derived from the SMPL model was used to guide the synthesis of the image reshape results. A specific set of design strategies are incorporated into our pipeline to achieve a faithful reshaping result. First, the synthesis process was divided into foreground and background and presented with two independent encoders to ease the reshape learning holistically. Second, to address structure misalignment, the 3D body deformations within our network via feature space warping were incorporated for foreground encodings. The warped foreground encodings are further combined with the background encodings before being passed to a decoder to produce a final reshaped image. Finally, to address the lack-of-paired-data problem, a novel self-supervised strategy was introduced to train our network with our pseudo-paired data.

NeuralReshaper is fast to use, fully automatic, and robust for images taken in unconstrained environments. The independent nature of SMPL parameters enables us to provide users with high-level semantic control over several key attributes of the human body, such as height and weight. We compare our method to several previous works and possible deep learning baselines. The evaluation of the indoor and in-the-wild datasets shows the superiority of our proposed method over the previous art and alternative solutions.

Method. We *automate* the parametric body fitting process using a data-driven initial fitting followed by a finetuning optimization step. For realistic reshaping, we introduce a novel *two-headed* neural architecture.

Skinned Multiperson Linear Model Fitting. SMPL is a differentiable mapping from the shape $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{72}$ to a 3D human model $M(\beta, \theta)$. Given a human image I, we obtain the initial shape and pose parameters (β^0, θ^0) with the pretrained model of [8]. To further refine the SMPL parameters, we used an optimization-based paradigm to iteratively align the fitted model with respect to the image cues. The overall optimization consists of two steps: optimizing β and θ using 2D key points (following [11]) and optimizing β for better silhouette matching. Because of the inherently decoupled shape and pose parameters in SMPL, users can intuitively achieve the desired body shape by directly adjusting the shape parameters β.

We project the corresponding 3D deformation onto the image space for the resculpted 3D human model to prepare a dense warping field T for the subsequent image reshaping stage. Given the dense warping field T, a naive idea would be to directly warp at the pixel level. However, a direct image warping approach based on T and its extrapolation from the human region to the entire image would easily result in noticeable artifacts in the background and foreground. In contrast, we choose to use T to guide the subsequent image synthesis via a neural generator to avoid distortions.

NeuralReshaper. As shown in Fig. 1, our network is designed as a two-headed UNet-like structure containing two encoders and a decoder to disentangle the complex foreground–background interactions. The foreground encoder $\mathcal{E}_f(I_f)$ consumes a foreground image I_f, the background encoder $\mathcal{E}_b(I_b, a)$ consumes a background image I_b

[*] Corresponding author (email: youyizheng@zju.edu.cn)

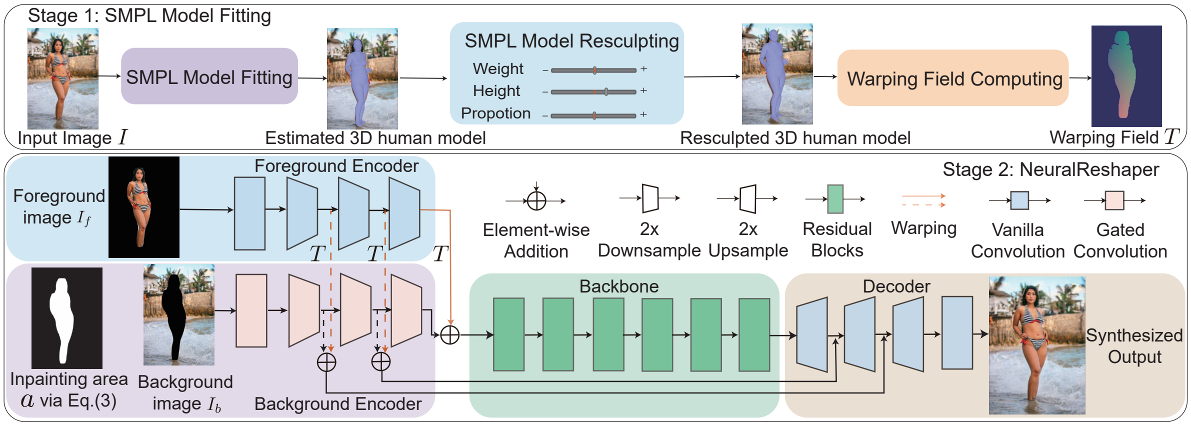

Figure 1 Overall pipeline of our proposed method, consisting of model fitting and neural reshaping stages.

with masked regions and a union foreground mask a (including occluded and disoccluded areas induced by deformation), and the decoder $\mathcal{D}$ takes the encoded codes from $\mathcal{E}_f$ and $\mathcal{E}_b$ and generates the final retouched result I^t. In an essential step, we take the warping field T derived from the body deformation to warp foreground features and fuse-warped foreground features with encoded background features. In addition to the above generator, during the training stage, a discriminator is used to enforce the overall realism of the generated image.

We introduce a unique warp-guided mechanism to integrate the features of the two encoders to generate desired retouching results based on T. Specifically, let f_i denote the intermediate feature produced by the i-th layer of the foreground encoder $\mathcal{E}_f$, i.e., $f_i = \mathcal{E}_f^i(I_f)$. We warp it to create a distorted feature $f_i^t = warp(f_i, T)$ (shown by the arrows in orange in Fig. 1), which is roughly aligned with the target shape. Then, we combined the warped foreground feature f_i^t with the corresponding background feature b_i to obtain a complete feature map $\varphi_i = f_i^t + b_i$ of the target.

With the warping-aware integration strategy, the generator succeeds in producing a synthesized image $I_{\text{out}} = \mathcal{D}(\mathcal{E}_f(I_f), \mathcal{E}_d(I_b, a), T)$ with the person in the appropriate shape. Ultimately, we obtain the target image

$$I^t = a * I_{\text{out}} + (1 - a) * I, \tag{1}$$

where $*$ denotes the element-wise multiplication.

Ideally, we should train our network with paired images (I, I^t) of the same persons under the same poses but in different shapes. However, obtaining such paired data is difficult, if not impossible. For that purpose, we introduce a novel self-supervised training strategy in which we use a deformed source image as the input and attempt to generate the original source image. As a result, the source image can naturally serve as supervisory information without any additional annotation.

To this end, we adopt an L1 loss to encourage this source image recovery

$$L_R = \left\| I - G(I_b, I_f^t, T^t) \right\|_1 . \tag{2}$$

We used the hinge loss [12] for GAN. Overall, we obtain an alternating minimization:

$$\begin{aligned} &\min_G \, (\lambda_{\text{recovery}} L_R + \lambda_{\text{gan}} L_G) \\ &\min_D L_D \end{aligned} \tag{3}$$

where λ_xs are tradeoff parameters for different losses.

Appendix C shows more experimental details. Note that our method is desired for altering the human shape parameters such as height and weight. Thus, the human pose is maintained untouched throughout the reshaping.

Conclusion. NeuralReshaperis a practical method for the realistic reshaping of the human body in single images using deep generative networks. Our method enables users to reshape human images by moving several sliders and receive immediate feedback. Extensive findings on the indoor and outdoor datasets and online images have shown our method's superiority compared with alternative solutions. Furthermore, we believe that our method can serve as automatic dataset generation for future supervised learning-based methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation China (Grant No. 62172363).

Supporting information Appendix A-C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Andriluka M, Pishchulin L, Gehler P, et al. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

2 Anguelov D, Srinivasan P, Koller D, et al. SCAPE: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers. 408–416

3 Bogo F, Kanazawa A, Lassner C, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European conference on computer vision (ECCV). Springer, 561–578

4 Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 43: 172–186

5 Geng J, Shao T, Zheng Y, et al. Warp-guided gans for single-photo facial animation. ACM Transactions on Graphics (TOG), 2018. 37: 1–12

6 He K, Gkioxari G, Dollár P, et al. Mask r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2961–2969

7 Johnson S, Everingham M. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In:

Proceedings of the British Machine Vision Conference. Doi:10.5244/C.24.12

8 Kanazawa A, Black M J, Jacobs D W, et al. End-to-end recovery of human shape and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7122–7131

9 Kato H, Ushiku Y, Harada T. Neural 3d mesh renderer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3907–3916

10 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980, 2014

11 Kolotouros N, Pavlakos G, Black M J, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2252–2261

12 Lim J H, Ye J C. Geometric gan. arXiv preprint arXiv:170502894, 2017

13 Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In: Proceedings of the European conference on computer vision (ECCV). Springer, 740–755

14 Liu W, Piao Z, Min J, et al. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 5904–5913

15 Liu Z, Luo P, Qiu S, et al. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

16 Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 2015. 34: 1–16

17 Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:180205957, 2018

18 Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10975–10985

19 Ren J, Yao Y, Lei B, et al. Structure-Aware Flow Generation for Human Body Reshaping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7754–7763

20 Tan Z, Chai M, Chen D, et al. MichiGAN: multi-input-conditioned hair image generation for portrait editing. ACM Transactions on Graphics (TOG), 2020. 39: 95–1

21 Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8798–8807

22 Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5505–5514

23 Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 4471–4480

24 Zhou S, Fu H, Liu L, et al. Parametric reshaping of human bodies in images. ACM transactions on graphics (TOG), 2010. 29: 1–10

Appendix A Method Details

Appendix A.1 SMPL Fitting Optimization

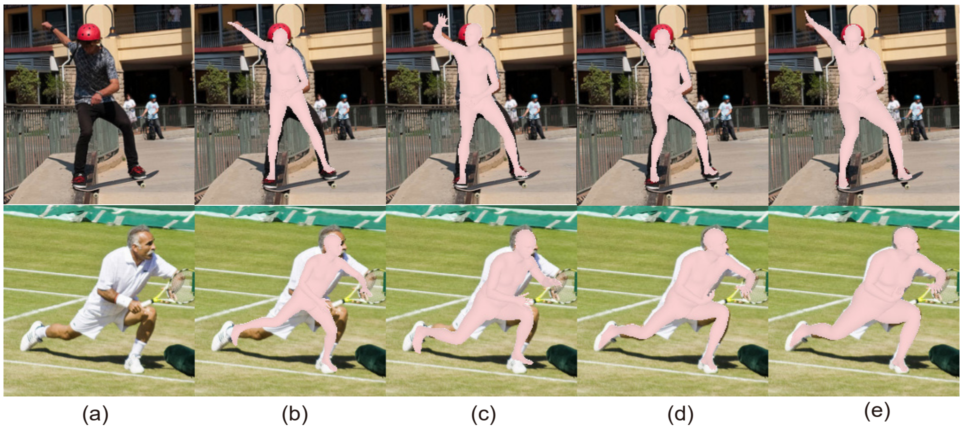

Figure A1 Fitting results obtained from different steps in the SMPL model fitting stage. (a) The original image. (b) The optimization result obtained from [3] (initialized with mean parameters). (c) The direct inference result obtained from the pre-trained model [8]. (d) The fitting result obtained after the 2D keypoints optimization step. (e) The fitting result obtained after the silhouette optimization step.

Refining β and θ using 2D Keypoints. We first extract from I the 2D keypoints $J_{\text{est}} \in \mathbb{R}^{2K}$ along with their confidence $w \in \mathbb{R}^K$ by using OpenPose [4] and optimize β and θ to match the image projections of 3D joints with the estimated 2D keypoints J_{est}. Since the initial value of β and θ are also reliable, we adopt a simplified energy of [3] to reduce the optimization time while not harming the accuracy.

$$E_{\text{joint}}(\beta,\theta) = \sum_{\text{joint } i} w_i \rho\left(\Pi_\alpha M_{\text{joint}}(\beta,\theta) - J^i_{\text{est}}\right), \tag{A1}$$

where M_{joint} denotes the 3D SMPL joint positions, Π_α denotes the orthogonal camera projection and ρ is the robust Geman-McClure penalty function.

Refining β using 2D Silhouette. We perform an additional step to further optimize β to minimize the L2 loss between the projected silhouette of the body and the estimated 2D binary mask m of the human body (extracted from I using Mask R-CNN [6]). We keep θ intact at this step since optimizing θ w.r.t. the silhouette may affect the actual positions w.r.t. the 2D key points.

$$E_{\text{silhouette}}(\beta) = \|\text{NR}_\alpha[M_{\text{mesh}}(\beta,\theta)] - m\|_2^2, \tag{A2}$$

where M_{mesh} denotes the SMPL body mesh, and the operator $\text{NR}_\alpha[\cdot]$ denotes the differentiable silhouette rendering [9] using the camera parameters α.

Fig. A1 shows an example with fitting results in different phases. We observe considerable improvements by using both 2D keypoints and silhouette alignment compared to the direct inference results of [8] (Fig. A1 (b)) and [3] (Fig. A1 (c)), respectively. Note here we do not use the parametric model of SCAPE [2], which is used in [24], since the SMPL model is more accurate and compatible with modern rendering pipelines [16].

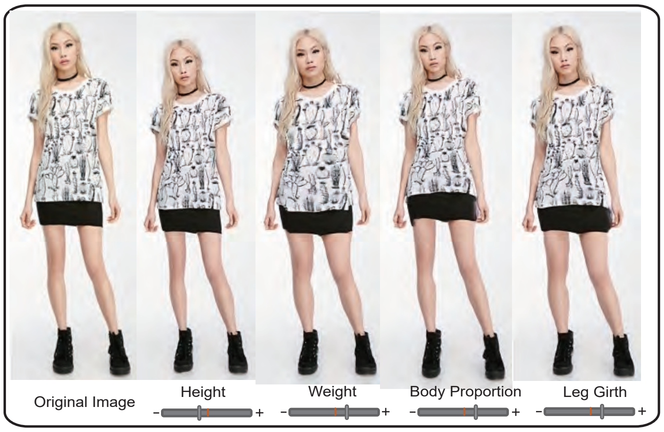

Figure A2 The reshaping results generated by our method with attributes adjusted individually.

We create an easy-to-use interface for semantic reshaping of human bodies by offering users a set of sliders corresponding to the aforementioned semantic attributes (Fig. A2), and the users can get reshaping results instantly after adjusting. Specifically, we pick four key shape parameters which separately represent the height, weight, leg girth, and body proportion for simple usage. Note that, increasing the body proportion indicates the lengthening of legs while keeping the height unchanged, and vice versa. We also allow users to pick a particular subject by drawing a bounding box before the 3D fitting, if multiple people exist in a source image.

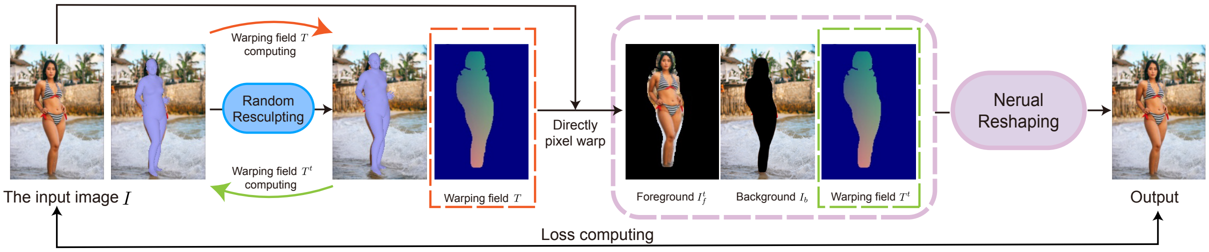

Figure A3 Our self-supervised training strategy. Given an input image I and its fitted SMPL model counterpart, we randomly resculpt the SMPL model and directly warp pixels with the warping field T to get a triplet of pseudo training inputs consisting of the deformed foreground image I_f^t, the background image I_b and the inverse warping field T^t. The neural reshaping module is trained to recover the original image I from the triplet inputs in which case I provides a neutral supervision.

Appendix A.2 Self-Supervised Training Strategy

Specifically, as shown in Fig. A3, for each image from an existing dataset (we denote these images as source images in the following), we first fit SMPL to it as described in Sec. Appendix A.1. We then resculpt the fitted 3D model by randomly altering the shape parameters. To ensure the deformation plausibility and generating large deformations simultaneously, we define valid ranges for SMPL shape parameters. Next, we randomly sample shape variations from these valid ranges following the unit Gaussian distribution. All the shape variations are controlled to weight/height changes in the range of [0-20] kg/cm, which are consistent with previous work [24] and satisfy most of the real-world reshaping requirements. The deformation from the source shape to the deformed one is projected onto the image space to form a warping field, denoted as T. Next, we directly warp the source image by T (at pixel-level) to generate the deformed foreground I_f^t, and compute the background image I_b. Finally, we obtain a paired data $((I_b, I_f^t, T^t), I)$ for training where T^t is the inverse of T. By taking the tuple (I_b, I_f^t, T^t) as input, our NeuralReshaper generator G should produce a complete image as close to the source image I as possible.

We benefit from this training strategy in a significant way. With this strategy, we have sufficient data for training. We keep a neat network and concise loss functions. The network learns how to produce the warped foreground and inpaint the background. It also learns how to composite the two parts together naturally.

Note that directly warping on pixel-level usually brings significant distortions (shown as I_f^t in Fig. A3), which makes our training foreground different from our testing foreground. Therefore, to relieve this ambiguity, we choose to warp on high-level features instead of directly warping on pixel-level and use GAN loss for realistic image generation. See Fig. B12 for comparisons between our approach and an alternative solution by directly warping in the image space (instead of the feature space).

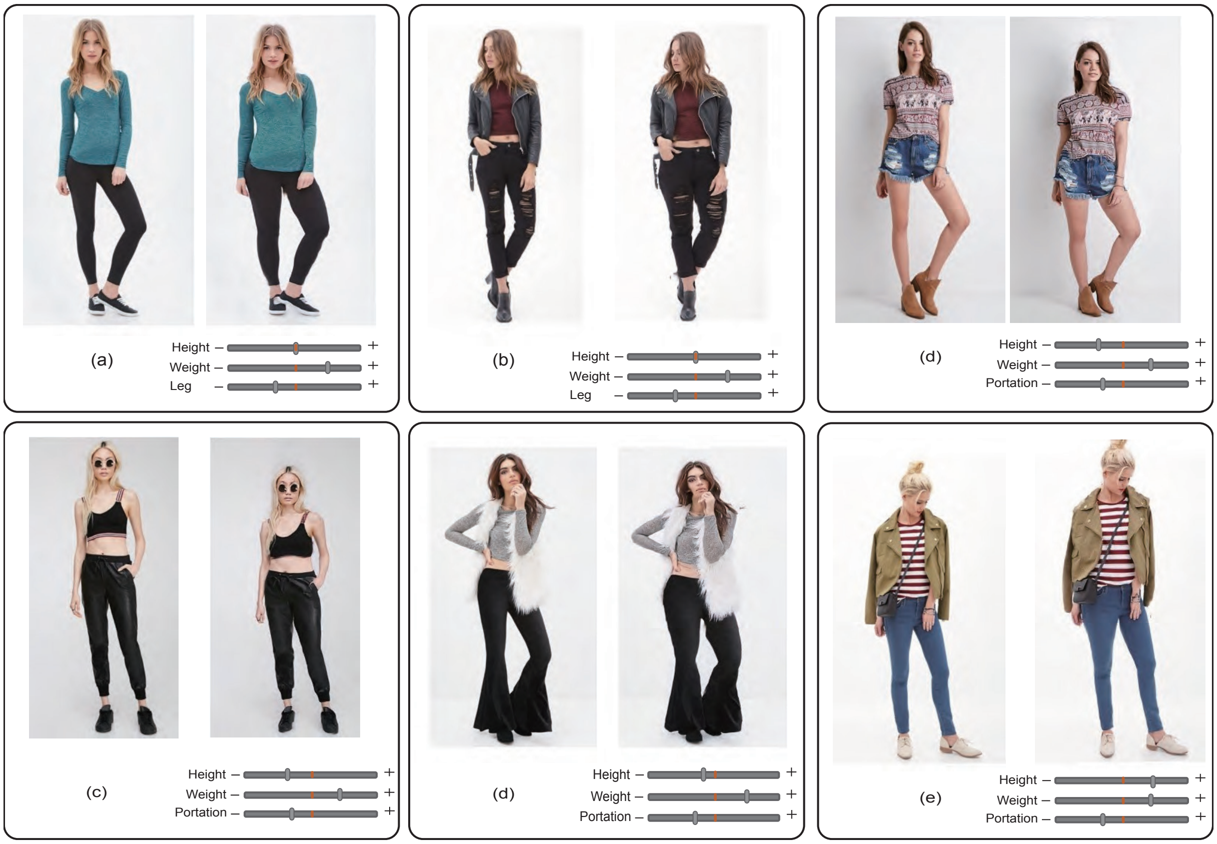

Figure A4 A series of representative reshaping results on the DeepFashion dataset. For each case, the left side is the original image and the right side is a reshaped image with the sliders below each case indicating the attributes and their degrees that have been edited.

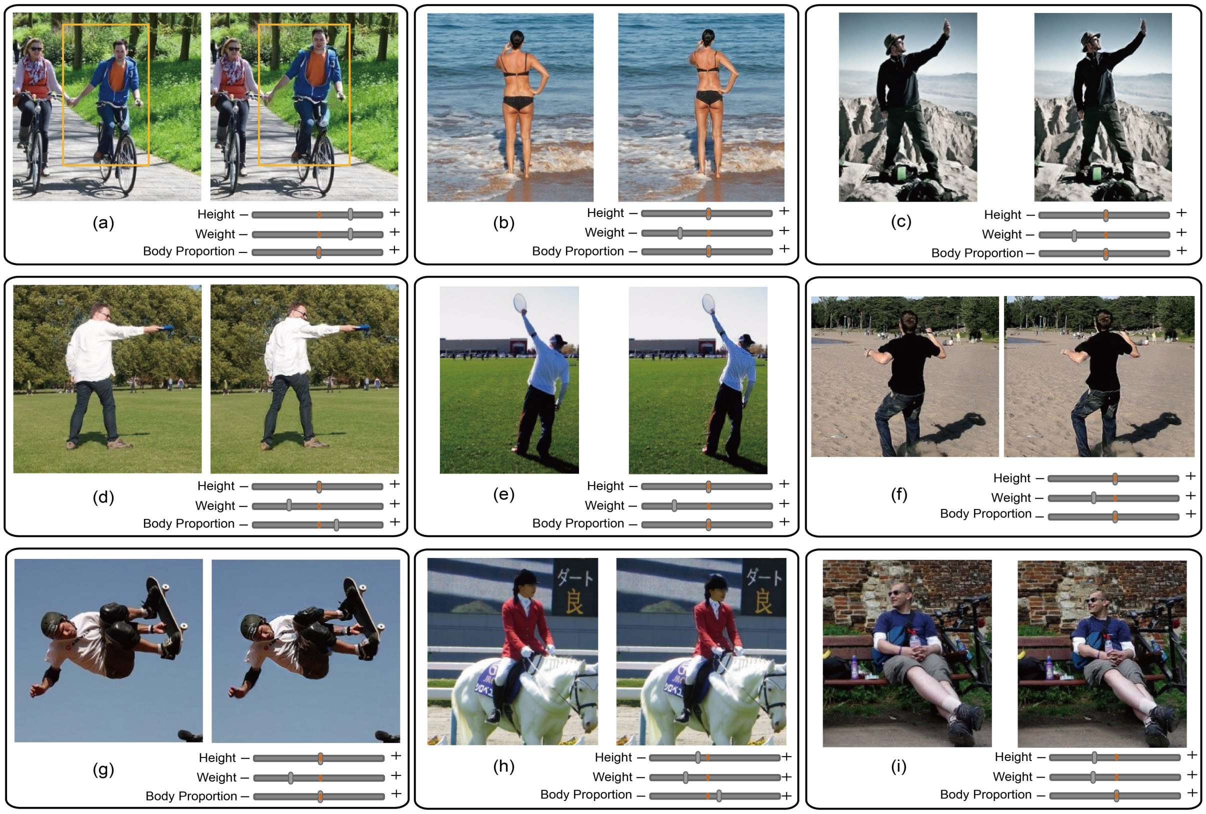

Figure B1 A series of representative reshaping results for outdoor images. For each case, the left side is the original image and the right side is a reshaped result with the sliders below each case indicating the attributes and their degrees that have been edited. The person that has been reshaped is highlighted with a yellow bounding box in (a).

Appendix B Experiments

Appendix B.1 Dataset

We have conducted extensive experiments on an indoor dataset DeepFashion [15] (512 × 512 resolution) and an outdoor dataset consisting of images from COCO [13], MPII [1], and LSP [7] (256 × 256 resolution). Since there are fewer high-quality human images in the outdoor dataset, we demonstrate our network's ability to generate photo-realistic images of higher resolution on DeepFashion. Both the training images and testing images are coming from the same source dataset. For DeepFashion, we use 85% of them for training and the rest for testing. For the outdoor dataset, we collected $29,353$ images in total. Specifically, we randomly pick $24,806$ for training and the rest for testing.

As mentioned in Sec. Appendix A.1, we run OpenPose [4] and Mask R-CNN [6] to obtain the 2D keypoints and human body silhouette for all images. For simplicity, we discard images that have less than six visible keypoints. We fit the SMPL model to each image and get rid of the images if the fitted region does not cover half of the human mask. For each image in DeepFashion, we compute the bounding box of the human and use it to crop the image and resize it to 512 × 512. For the outdoor dataset, images are cropped and resized to 256 × 256. Though our method proposes to refine SMPL shape parameter according to human body mask, it is still difficult and error-prone to achieve desirable fitting results for loose-fitting cases since SMPL is a naked body model. Thus, we do not consider the case of very loose dresses or skirts in the following experiments.

Appendix B.2 Implementation Details

Appendix B.2.1 *Fitting Stage*

The initial SMPL fitting network [8, 11] cannot handle images of varying input size, we crop and resize all training images into 224 × 224. For the optimization step, we use Adam [10] for both 2D keypoints and silhouette optimization with a learning rate of 0.01 and 0.05, respectively. We set 100 as the total iteration number for both steps.

Appendix B.2.2 *Reshaping Stage*

The UNet-like generator consists of two parallel encoders, a bottleneck, and a decoder symmetric to the encoders. Specifically, we employ 4 layers of convolutions for each encoder. The input layer increases the number of channels to 64, and the following three downsampling layers decrease the spatial dimension by a factor of 2 while increasing the numbers of channels to 128/256/512. Each convolution layer is followed by an instance normalization and a leaky ReLU activation. The only difference between the two encoders is that we use gated convolution for the background and vanilla convolution for the foreground since gated convolution provides more flexible feature learning for background inpainting [22]. We use 6 layers of residual blocks in the bottleneck and 4 layers for the decoder. The spectral-normalized [17] patch discriminator consists of 6 layers of vanilla convolution.

We implement our system in PyTorch with one NVIDIA 1080Ti GPU (11 GB memory). We train the network for 100 epochs with the Adam optimization [10]. We set the learning rates for the generator and discriminator to 0.0001 and 0.0004, respectively. We use a batch size of 8 for the outdoor dataset and 2 for the DeepFashion dataset. We set the weight $\lambda_{\text{recovery}}$ to 100 and λ_{gan} to 10. In the training stage, our method takes about 48 hours for training on the indoor dataset DeepFashion, 72 hours for training on the outdoor dataset. At runtime, our system takes about 15s for pre-processing each image, including semantic segmentation using

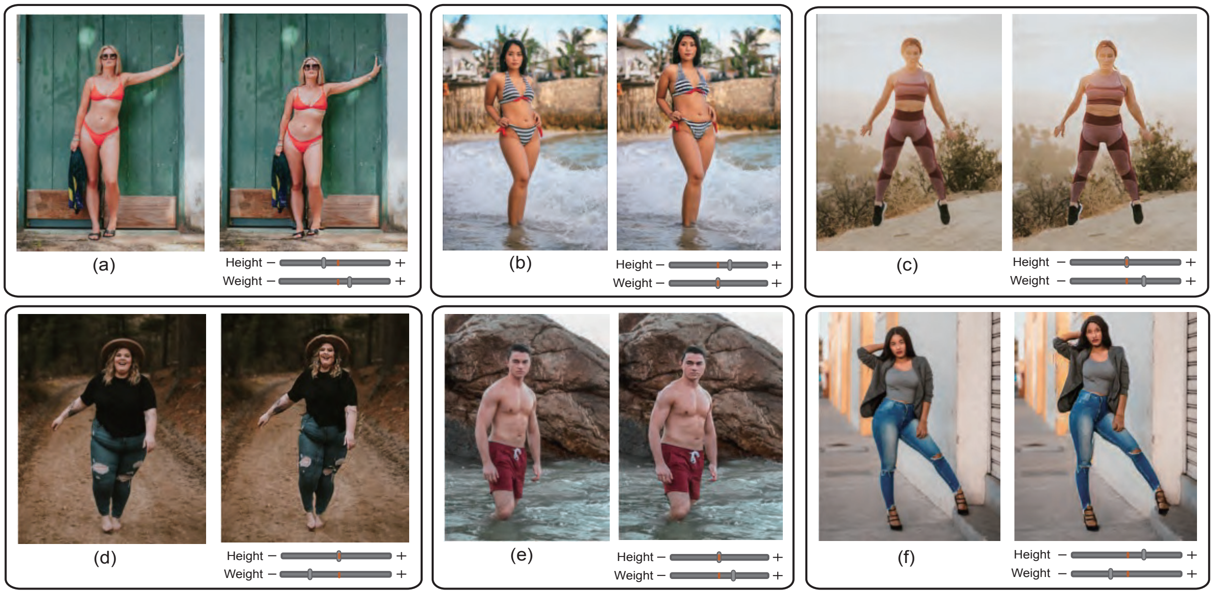

Figure B2 Several reshaping results tested on online images. For each case, the left side is the original image and the right side is a reshaped result with the sliders indicating the attributes and their degrees that have been edited.

MaskRCNN [6], keypoint detection by OpenPose [4], and SMPL fitting optimization. Then the modification of the parameters to derive reshaping results runs at an instant rate (less than 1s).

Appendix B.3 Qualitative Results

Appendix B.3.1 *Results on the DeepFashion Dataset*

Fig. A4 shows reshaping exemplars on the DeepFashion Dataset with persons in diverse poses, shapes, and appearances. The sliders below each case indicate the user-specified shape attributes. The system supports individual or joint manipulation of these attributes. We observe that the system produces natural and consistent reshaping results while being robust under large deformations. The synthesis results preserve high-frequency details of clothes, face, and hair in the original images. Fig. A2 illustrates the individual effect of each attribute, showing the well-disentangled reshaping semantics.

Appendix B.3.2 *Results on the Outdoor Dataset*

We also show results on outdoor dataset where complex occlusions and backgrounds can present. Fig. B1 shows reshaping exemplars under diverse outdoor situations. Though trained with the self-supervised data, our system is able to handle large reshaping effects. The synthesized textures in the dis-occluded areas and the deformed foreground are visually realistic. The color discrepancy along the boundary between the mask region and the background is the common artifacts in previous works of image inpainting [23]. In contrast, our method produces spatially consistent textures and colors in these regions. The synthesized foreground is blended naturally and seamlessly with the background without boundary artifacts. It indicates that our model succeeds in decoding the desired foreground and background in a reasonable spatial layout by merging both the branches' information. Fig. B1 shows that our approach provides fine-grained and holistic reshaping effects. The last row of Fig. B1 exemplifies reshaped body images with challenging poses, severe occlusions, and complex backgrounds.

Appendix B.3.3 *Results on the Online Images*

Results, exhibited in Fig. B2, show that our semantic reshaping method generalizes well to the randomly retrieved online images outside of the testing dataset. It is noteworthy that our method is fully automatic and does not require any annotations such as 2D keypoints or silhouette during testing, thanks to the robust 2D detection techniques (i.e., OpenPose [4] and Mask R-CNN [6]). Our method is applicable to a wide range of real situations.

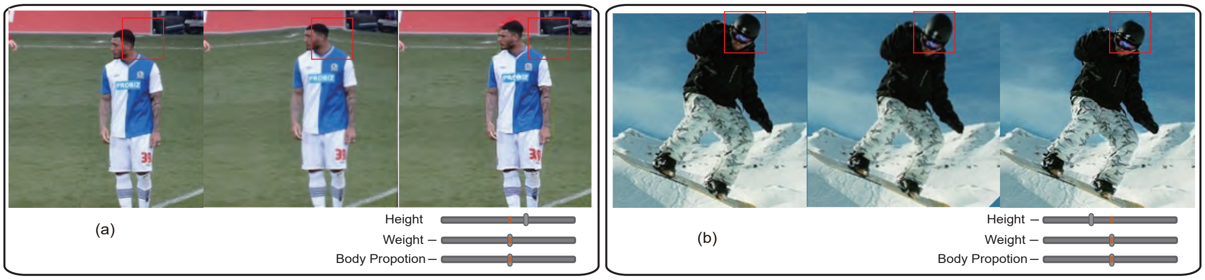

Figure B3 A comparison between direct warping with extrapolation and our method. From left to right: the original image, the reshaping result by direct warping, and the reshaping result by our method.

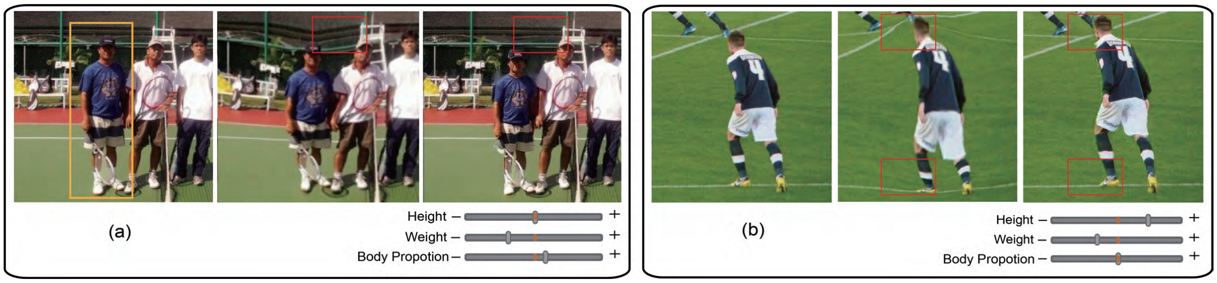

Figure B4 Comparison with Zhou et al. [24]. For each case, from left to right are the original image, the reshaping result by the method in [24], and the reshaping result by our method. The sliders below each case indicate the attributes and their degrees that have been edited. The person that has been reshaped is highlighted with a yellow bounding box in (a). We also highlight the distorted areas in [24] and their corresponding patches in our result with red bounding boxes.

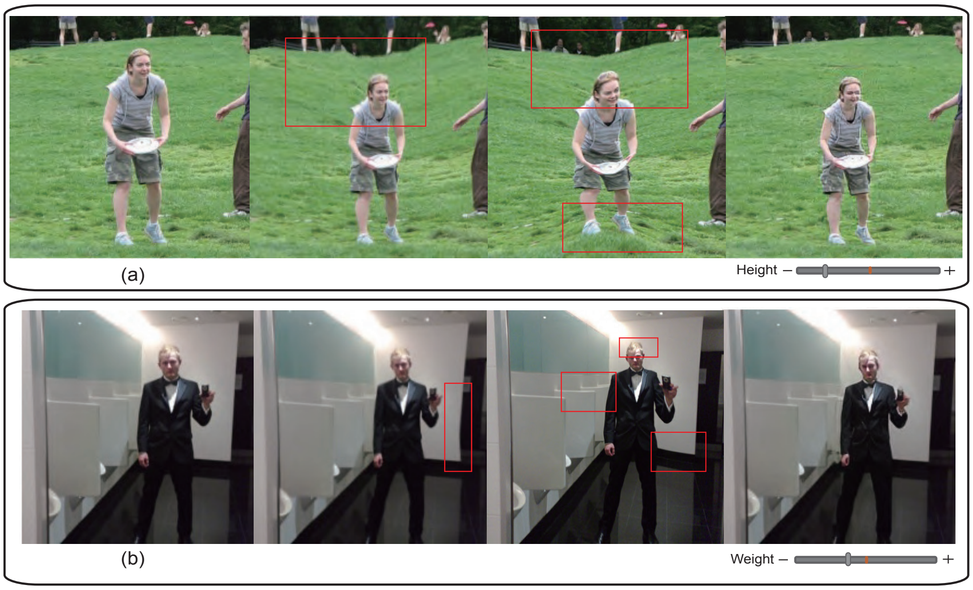

Figure B5 Comparisons with the direct warping method and Zhou et al. [24]. For each case, from left to right are the original image and the reshaping results generated by direct warping, Zhou et al. [24], and our method respectively. The sliders below each case indicate the attributes and their degrees that have been edited. We also highlight the distorted areas with red bounding boxes.

Appendix B.4 Comparisons

In this section, we compare our approach to the direct warping method, the warping-based human reshaping method [24], the state-of-the-art human image editing method [14] using deep learning and the state-of-the-art human reshaping method [19].

Visual comparisons with the direct warping method and the warping-based human reshaping method are shown in Fig. B3, Fig. B5 and Fig. B4. As one can see, both directly warping-based methods introduce distortions in the foreground and background, leading to unrealistic and implausible results. The larger deformations conducted, the severe distortions induced. This is mainly because that directly warping-based methods first conduct delaunay triangulation on the full image and deform these triangles according to the 3d human model. Thus, the deformations of human model are propagated to the whole image, introducing unfavorable changes on background. Despite the background distortions, warping-based methods also induce distortions on unfitted body areas such as hair (see Fig. B5 (a)). This is mainly because that the unfitted body areas are treated as background. As suggested in [24], one possible strategy to alleviate such distortion artifacts is to manually adjust the saliency map. Nevertheless, this involves lots of human intervention. In contrast, our method is fully automatic in fitting, and preserves regular patterns occluded by humans (e.g., wire fence in Fig. B4 (a)) rather than distorting them. This is because that our method decomposes the generation of foreground and background parts and the generator learns to inpaint the missing areas.

The human image editing method [14] trains their network with paired images sampled from videos in a fully supervised manner and applies it to specific tasks like pose-guided human image transfer. For our body reshaping task, since we lack paired data, it is impossible to reproduce their method on outdoor in-the-wild datasets like COCO [13]. We compare our method with [14] on their dataset, COCO dataset and online images, respectively. Fig. B6 (a) shows that our self-supervised method even gets better results than their supervised training method on the reshaping task. One can see that our method can preserve original appearances well and our results look more realistic. As shown in Fig. B6 (b), their method leads to seriously degraded results when facing non-seen cases. We also do quantitative comparison with [14] in Table. B1. We show the Frechet Inception Distance (FID) between randomly sampled generated images and original images. The superiority proves that our method can generate more realistic images again.

We also compare our method with the state-of-the-art reshaping method [19]. Ren et al. [19] exploit neural networks to generate deformation flow for image warping. A single scalar is used to control the deformation extent. To train their network, they require a dataset of reshaped ground truth produced by artists. However, their method can only amend the weight and is unable to modify the height. Visual comparisons on COCO dataset are shown in Fig. B7. The results show that our method generates comparable results with B7 (see results in columns 2 and 3) while our method can also change the height (see results in column 4). Moreover, the deformation field generated by their method can introduce artifacts to the image. As highlighted in the red bounding boxes in Fig .B7 (c), undesirable deformations and distortions exist, while our method generates more realistic results. For a fairer comparison, we also present results on their released dataset BR-5K in Fig. B8. We again observe that Ren et al. [19] sometimes

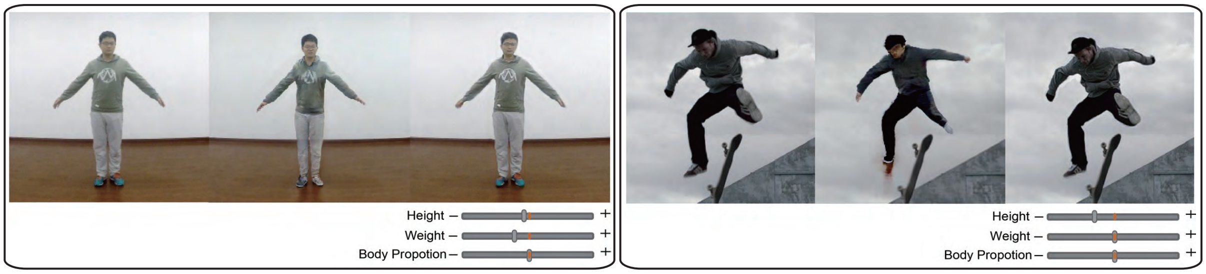

Figure B6 Comparison with Liu et al. [14]. For each case, from left to right are the original image, the reshaping result by the method in [14], and the reshaping result by our method.

generate unnatural reshaping effects, as highlighted in the red bounding box, while our method produces more global consistent results. Moreover, their methodis unable to produce height change, while our method generates realistic height-changing effects.

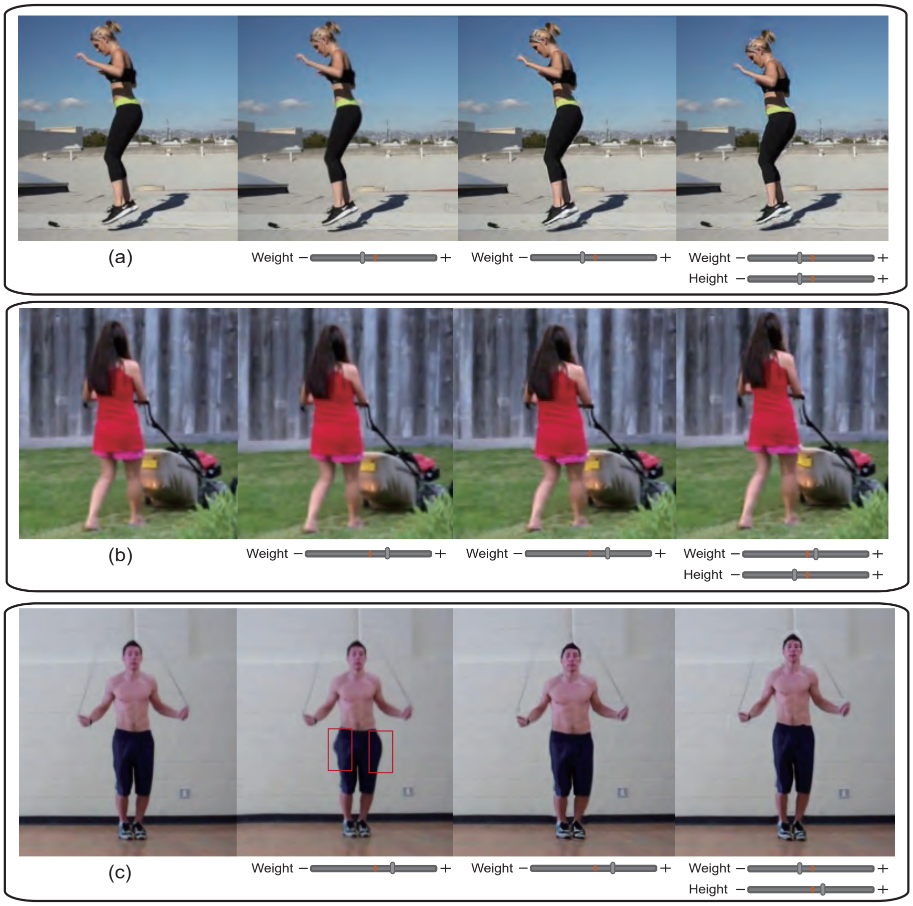

Figure B7 Comparison with Ren et al. [19]. For each case, from left to right are the original image, the reshaping result by [19], and two different reshaping results by our method.

Appendix B.5 Ablation Study

We put considerable efforts into the network design to keep it compact without compromising the reshaping performance. To analyze the impact of each component and justify its necessity, we perform two ablation studies. We refer to the network introduced by Fig. 2 in our main paper as the full model. We obtain the variants by replacing different components with alternatives and train them with the same protocol.

First, we demonstrate the necessity of our optimization during the SMPL model fitting stage. As shown in Fig. B9, without either of our optimization, there exist feature loss and distortions in the synthesized images. This is because if we fail to fit the SMPL model to the body region, we cannot compute the accurate warping field for reshaping. Specifically, using 2D keypoints for optimization will improve our performance when facing challenging poses, and using the 2D silhouette for optimization will improve our performance when handling fluffy clothes. Note that, although our method can handle fluffy clothes like Fig. B9, we will also have distortions for cases with very loose skirts or flying hairs.

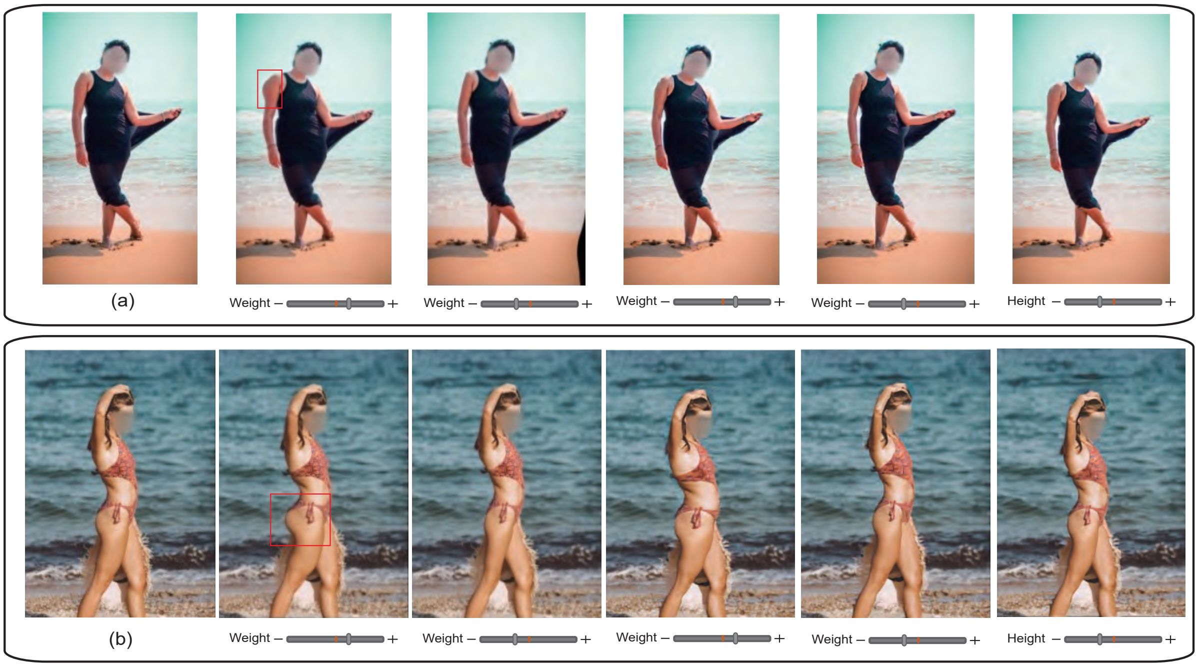

Figure B8 Comparison with Ren et al. [19] on BR-5K dataset. For each case, from left to right are the original image, the two reshaping results (i.e. weight-gaining and weight-losing) by [19], and three reshaping results (i.e. weight-gaining, weight-losing, the change of height) by our method.

Method	FID ↓
Liquid Warping [14]	89.41673
Ours	**80.28321**

Table B1 Quantitative evaluation of the generated images with the retouched new poses.

In reshaping stage, we replace the gated convolution with vanilla convolution for the background branch. We denote the resulting model as variant **G-**. Fig. B10 shows that the synthesized textures by **G-** tend to be blurry and exhibit color discrepancy with their surroundings. In contrast, our full model successfully produces rich and consistent details (see the area highlighted with a red rectangle).

The image reshaping requires an extensive spatial rearrangement on a complex background. The key is how to blend the deformed foreground with the inpainted background without introducing artifacts. Our method exploits the simple addition of the two branches in the feature space. To validate its effectiveness, we compare it with two alternatives. The first one is a popular combine method that merges two feature maps in a mask-guided way [20]. Specifically, we compute a mask for the warped foreground, denoted as m^t. We then integrate the features as

$$\varphi_i = b_i + f_i^t * m^t, \tag{B1}$$

where b_i and f_i are the background and foreground feature maps of the i-th layer, respectively, and $*$ denotes the element-wise multiplication. We replace Eq. 4 in our main paper with Eq. B1 in the full model to obtain the variant **M+**. Fig. B10 shows that the mask-guided merging tends to introduce artifacts around synthesized humans (see the area highlighted with a blue rectangle). The second way is that we only merge features after the last layer of encoding. Thus the skip-connections are not available in such situation. We denote the resulting model as variant **C-**. Fig. B11 shows that without per-layer merges and skip-connections, our method tends to introduce artifacts around synthesized humans (see areas highlighted with red rectangles).

We compare our method with an alternative that directly warps the foreground (instead of the feature space) and composes it with the inpainted background. We diffuse the warping field to the whole foreground mask via the warping strategy used in [5] and employ the network presented in [23] for inpainting. Fig. B12 shows that our method synthesizes realistic shapes and details while the direct-warping strategy often brings in distortions (e.g., the hair region in Fig. B12 (a)) or blurriness (Fig. B12 (b)).

Appendix C Limitations

Our method has a few limitations. Our method might introduce artifacts under extreme deformations (see Fig. C1 (a)). Most of the results we presented are within a weight/height change of 20 kg/cm to ensure the deformed SMPL model is valid in 3d space. However, when deformations larger than this range happens, our method may fail. The reasons are two-fold. First, the resculpted 3D model might become implausible, leading to unnatural image warping effects. Second, large deformations on human body can cause undesirable distortions on human faces when the SMPL face model is not fitted well with face areas on 2d images. Third, the inpainting may fail, since such a large occlusion would not appear in the training data. To alleviate such problems, we could learn a manifold that admits valid SMPL parameters, and reject implausible ones. Moreover, we could enhance our network's texture hallucination for large occlusions by improving the data and strategies.

Our method relies on the 3D fitting, whose imperfections may degrade the image reshaping effects. The fitting could fail in cases where severe occlusions or self-occlusions exist. We observe that the fitting quality decreases dramatically if more than half of the human body is dis-occluded. Manual corrections may help in such situations. On the other hand, the automatic fitting of

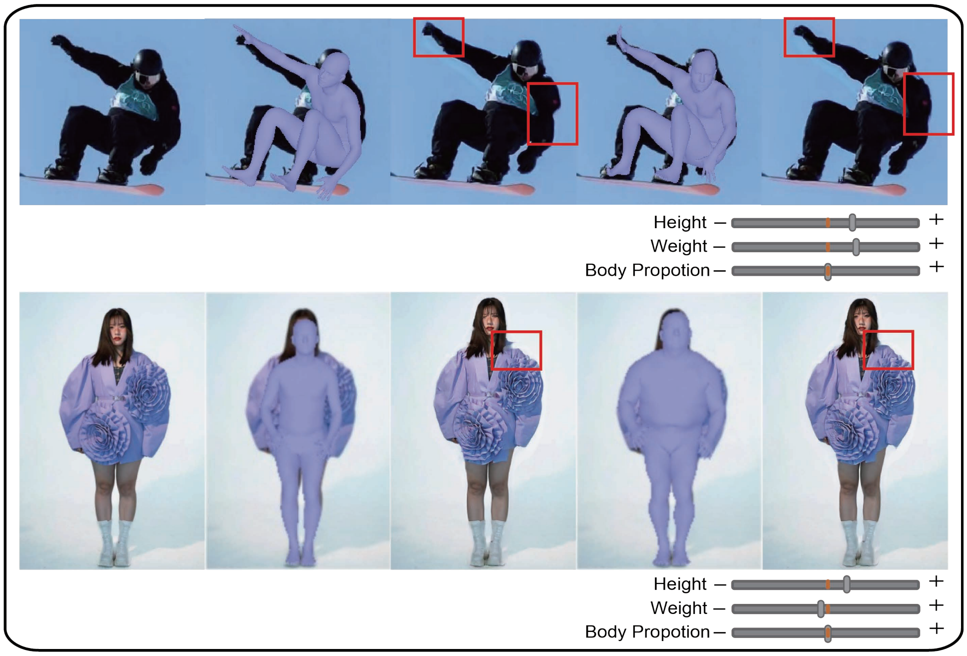

Figure B9 Ablation study on the SMPL fitting optimization. The first row shows the fitted SMPL model and synthesized image without 2D keypoints optimization and the corresponding results with our full optimization. The second row shows the comparison on the silhouette optimization.

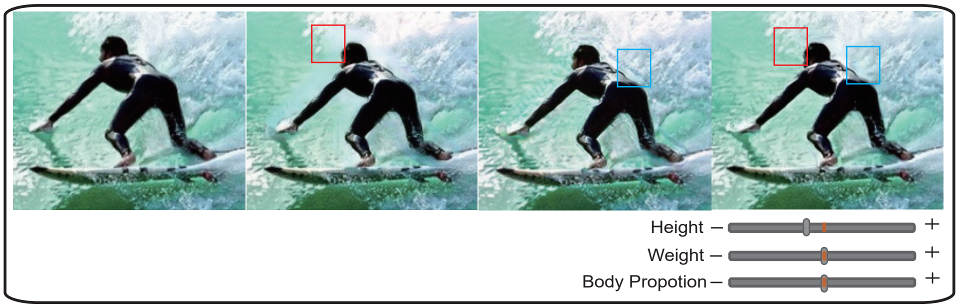

Figure B10 Ablation study. For each case, from left to right are the original image, the reshaping result generated by **G-**, the result by **M+**, and the result by our full model. The sliders below each case indicates the attributes and their degrees that have been edited. Red and blue rectangles are used to highlight the areas with artifacts.

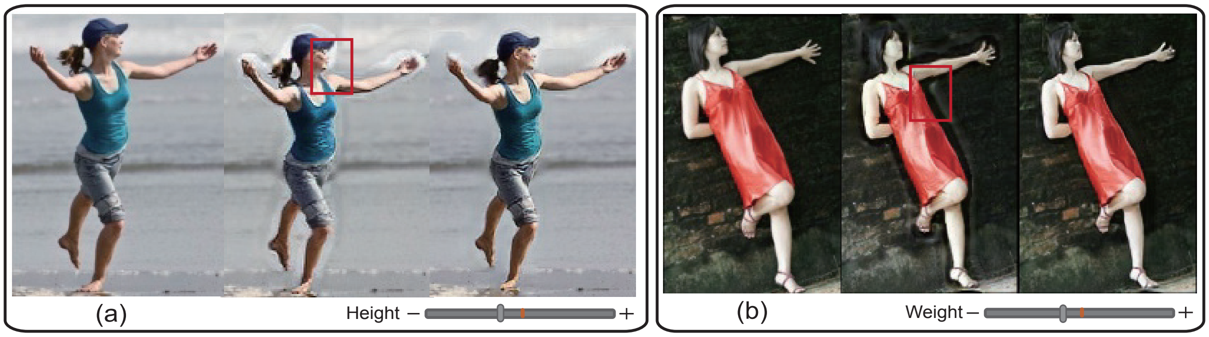

Figure B11 Ablation study. For each case, from left to right are the original image, the reshaping result generated by **C-** and the result by our full model. The sliders below each case indicates the attributes and their degrees that have been edited. Red rectangles are used to highlight the areas with artifacts.

SMPL model often fails at fine-grained regions, such as hands, which may lead to undesirable synthesized results at those regions (Fig. A4, B1, B2). A more fine-grained parametric model [18] might help.

Our method deals well with parts tightly coupled with the body, like cloth, hat, or hair, since Mask R-CNN [6] usually segments the foreground with these areas included. However, if the interacting objects with people are not covered by the segmentation mask, they remain where they were during the synthesis, thus decreasing the realism of the generated results (see Fig. C1 (b)). To appropriately represent the interactions between objects, we could introduce the scene graphs into the generation process.

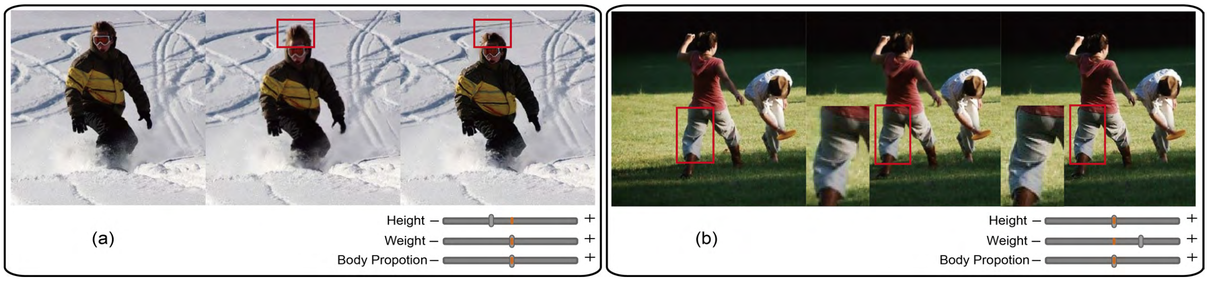

Figure B12 Comparison with an alternative approach by combining direct warping on foreground with background inpainting. For each case, from left to right are the original image, the reshaping result by the alternative approach and the reshaping result by our method.

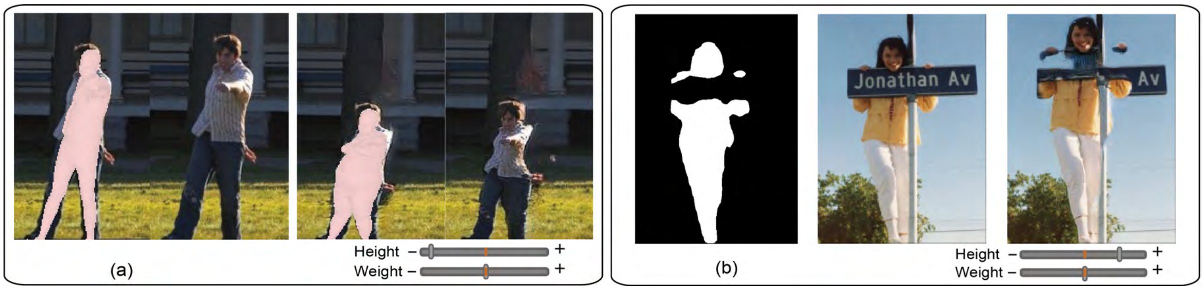

Figure C1 Failure results generated by our method. (a) Synthesized results under extremely large deformations. (b) Synthesized results when the interacting objects lie outside the foreground mask.

For generating high resolution results, our method needs high resolution datasets. However, training with high resolution data is very time-consuming and requires large GPU memories. Training with multi-scale generator and discriminator like [21] might help. Also, as to the human face, adding constraints to fix face features may help us avoid face changing.

Our method cannot handle the reshaping of multiple people simultaneously. The system requires the localization of the reshaped person via a bounding box to enable the sequential editing of each individual in multi-person images.