SCIENCE CHINA
Information Sciences

• **RESEARCH PAPER** •

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*

Abstract In this paper, we present NeuralReshaper, a novel method for semantic reshaping of human bodies in single images using deep generative networks. To achieve globally coherent reshaping effects, our approach follows a fit-then-reshape pipeline, which first fits a parametric 3D human model to a source human image and then reshapes the fitted 3D model with respect to user-specified semantic attributes. Previous methods rely on image warping to transfer 3D reshaping effects to the entire image domain and thus often cause distortions in both foreground and background. In contrast, we resort to generative adversarial nets conditioned on the source image and a 2D warping field induced by the reshaped 3D model, to achieve more realistic reshaping results. Specifically, we separately encode the foreground and background information in the source image using a two-headed UNet-like generator, and guide the information flow from the foreground branch to the background branch via feature space warping. Furthermore, to deal with the lack-of-data problem that no paired data exist (i.e., the same human bodies in varying shapes), we introduce a novel self-supervised strategy to train our network. Unlike previous methods that often require manual efforts to correct undesirable artifacts caused by incorrect body-to-image fitting, our method is fully automatic. Extensive experiments on both indoor and outdoor datasets demonstrate the superiority of our method over previous approaches.

Keywords Image Manipulation, Deep Neural Networks, Human-body Reshaping.

Citation Beijia Chen, Yuefan Shen, Hongbo Fu, Xiang Chen, Kun Zhou, Youyi Zheng.
. Sci China Inf Sci, for review

1 Introduction

Semantic retouching of human bodies in images, such as increasing the height, slimming the body, etc., has been long coveted. Yet, the problem is essentially ill-posed as one needs to anticipate a set of articulated and non-rigid deformations of different body parts, given that the deformations are inherently three-dimensional. The situation is more complicated when images are taken in unconstrained environments where occlusions, complex interactions between the human body and its surroundings are presented.

Early attempts [27,77] have tried to tackle this problem by interactively fitting a 3D parametric human model to human bodies in images and let the fitted 3D model delegate the transformation. The final reshaping effect is achieved via image warping. Although compelling results are generated, these methods could suffer from the following two aspects. First, manual efforts are often required to iteratively correct the body-to-image fitting to ensure plausible retouching. Second, warping-based reshaping unavoidably introduces distortions in both the foreground and the background under large deformations.

The recent progress in generative neural networks like generative adversarial networks (GANs) [19, 49, 50], variational auto-encoders (VAEs) [14, 66], and their followups [12, 22, 31, 51] has facilitated many photo-realistic image synthesis tasks such as semantic image generation [55] and face editing [18]. Nevertheless, retouching of body shapes in a single image still poses a set of challenges for generative models. First, supervised end-to-end training requires paired data. However, such paired data is very scarce in practice. Previous pose transfer methods like [46] tend to create task-specific datasets which are non-flexible and lack expansion to different datasets. Second, human images are complex, highly varied, constrained to kinematic constraints and often exhibit complex self-interactions and interactions with

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

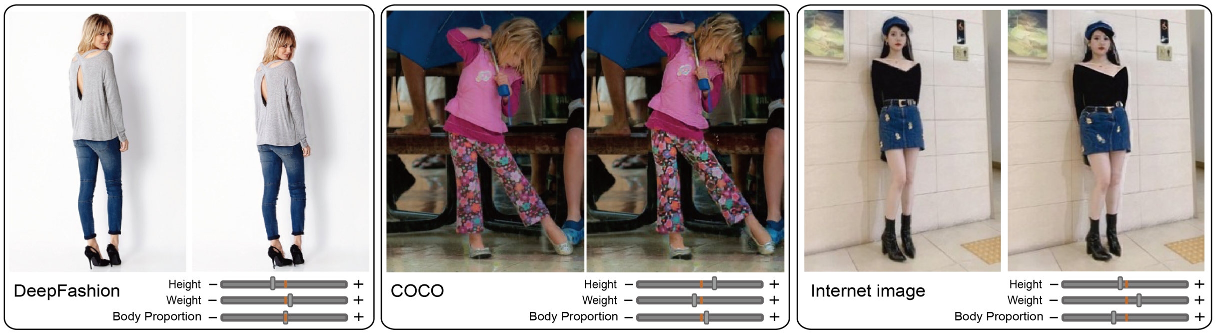

Figure 1 Our approach allows users to reshape human bodies by adjusting disentangled parameters, e.g., height and weight.

the background [39]. To enable an explicit control over the body shape, a disentangled representation that factorizes the human shape from other image factors, such as appearance and background, is needed. Last but not least, the reshaping of human bodies involves changes in image spatial layout and requires the synthesis of deformed foreground and dis-occluded image background. Due to the structure misalignment between the input and output images [17], simply applying typical network architectures like traditional UNet [61], which serve as solid baselines for many image generation tasks, are not workable here, as proved in [63].

In this paper, we present NeuralReshaper (Fig. 2), the first self-supervised learning based method for realistic human body reshaping in a single RGB image following a fit-then-reshape pipeline [27,53,60,77]. We first automate the fitting process using a hybrid learning-and-optimization based method with the Skinned Multi-Person Linear (SMPL) [48] model. Then in an essential stage, we exploit the 3D geometric deformation derived from the SMPL model to guide a subsequent generative model to synthesize reshape results in the image. As to our reshape learning, holistically, we separate the synthesis process in the background and the foreground and present two independent encoders: one for the foreground (i.e., body) and the other for the image background. To deal with the structure misalignment, we incorporate the 3D deformations of body within our network via feature space warping for foreground encodings. The warped foreground encodings are further integrated with the background encodings and passed to a decoder to produce a final reshaped image. To address the lack-of-paired-data problem, we introduce a novel self-supervised strategy to train our network with our pseudo paired data, as shown in Fig. 3.

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 high-level semantic control over several main attributes of human bodies such as height and weight. We compare our method with several previous works and possible deep learning baselines. The evaluation on both the indoor DeepFashion dataset [47] and in-the-wild dataset (consisting of the data from COCO [43], MPII [1] and LSP [28]) shows the superiority of our proposed method over the prior art and the alternative solutions.

To summarize, our paper makes the following contributions:

- We present NeuralReshaper, the first self-supervised learning based method for realistic human reshaping in single images. Our tool is easy-to-use, fully automatic, and works well for images taken in unconstrained environments.
- We introduce a novel self-supervised training strategy which does not require any extra ground-truth annotations to handle the lack-of-data problem.
- We introduce an effective network with a two-headed UNet-like generator for orthogonal encoding of foreground and background to hallucinate textures in the deformed body area and the dis-occluded areas, respectively. The warped feature maps in the foreground head are merged with the feature maps in the background head and passed to a decoder to produce globally coherent results.

2 Related works

Our work is closely related to several topics, including 3D human recovery from images, body reshaping in images, deep image generation networks that involve geometric transformations, and image inpainting.

3D Human Recovery. A faithful recovery of a 3D human body from images could ensure the fidelity of a subsequent manipulation of the underlying body shape in images and serves as a key ingredient in our task. 3D human recovery has been a long-standing research topic. The existing solutions to this problem may be divided into two main categories: optimization-based methods [10, 20, 37, 64] and learning-based methods [29,30,54,58,67,69], both of which are heavily built upon parametric human body representations [2, 48]. The optimization-based methods find the best-fit parameters by projecting the

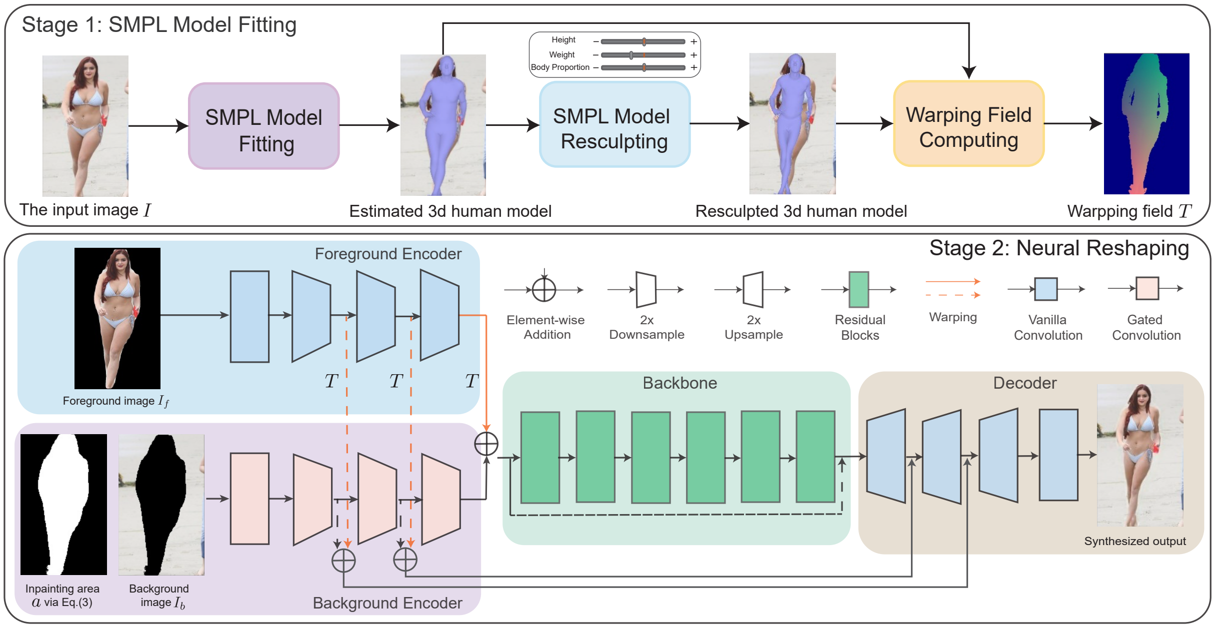

Figure 2 The overall pipeline of our proposed method, consisting a SMPL model fitting stage and a Neural Reshaping stage.

rendered 3D models to match 2D cues such as 2D keypoints and silhouette under kinematic constraints, to produce accurate estimations. In contrast, the learning-based methods resort to deep neural networks for fast and more stable estimations, but possibly at the cost of recovery accuracy. In our paper, we take advantage of both types of methods: we obtain an initial estimation using a pre-trained deep neural network followed by an optimization step for refinement, to strike a good balance between accuracy and efficiency.

Human-body Reshaping in Images. Pioneer works [27, 53, 60, 77] achieve semantic shape editing of human bodies images or videos via image warping [27, 77]. Zhou et al. [77] introduce the first semantic body retouching method. They first fit a 3D parametric human model to images and then deform it to a target 3D shape, followed by a body-aware warping approach to obtain reshaping results. Although an initial fit might be estimated automatically, user assistance for manual adjusting is often required in their method for higher accuracy. The idea of parametric body reshaping in [77] has been successfully extended by Jain et al. [27] for body reshaping in videos. These warping-based methods may introduce distortion artifacts in both the foreground and the background (Fig. 5). To address the distortion in the background, the works in [53, 60] propose to use a combination of foreground warping and background inpainting. However, such approaches would easily cause incoherent results between the foreground and background regions due to the inaccuracies of image matting and inpainting. To address these issues, we leverage the recent advances in deep neural networks to generate both the foreground and the background, enabling more robust reshaping effects. Very recently, a parallel work [59] presents a paired dataset with professionally retouched targets for supervised training. However, reshaping in [59] is not flexible (e.g., they are not able to modify body height or leg length) because they only have training data for limited deformation (lacking deformation of height growth, etc.), and the proposed orthogonality relation learning is not workable for reshaping body skeleton length. In contrast, our proposed self-supervised method allows us to pursue free-form reshaping.

Modeling Spatial Deformations in Generative Adversarial Networks. Human body reshaping brings changes to the spatial layout in an image, and thus an essential task is to anticipate the deformed foreground. Compelling results have been achieved in image translation tasks [3, 8, 26, 45, 55] using GANs [19]. However, the ability of modeling spatial deformations between structurally misaligned image pairs is still limited [13, 42, 72], since a commonly used backbone network architecture, UNet [61], is not capable of capturing structure deformations. To model such geometric deformations in neural networks, several methods estimate geometric warp parameters [42] or warping fields to enable geometric transformations [72] between the input images and its structurally misaligned output. Adversarial learning is further employed to ensure the realism of deformations in [42, 72]. However, such methods are only able to learn subtle or affine transformations, which are not applicable in our case where transformations are often large, complex, and non-rigid.

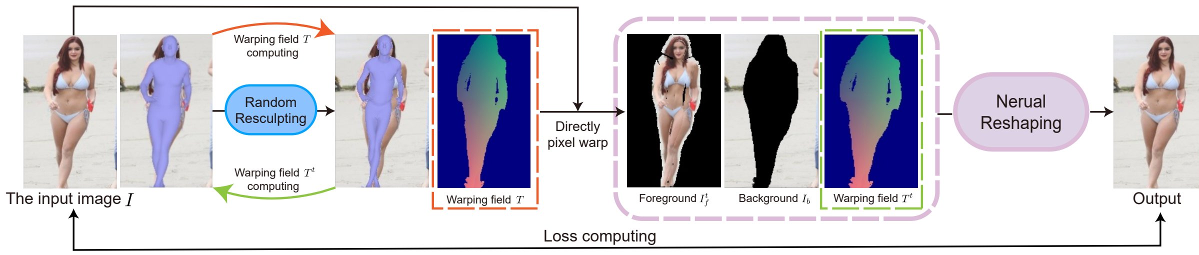

Figure 3 Our self-supervised training strategy. Given an input image I and its fitted SMPL model counterpart, we randomly resculpt the SMPL model and directly warp pixels with the warping field T to get a triplet of pseudo training inputs consisting of the deformed foreground image I_f^t, the background 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.

A series of studies [4, 46, 63, 78] in pose-guided human image synthesis have focused on modeling large deformations during synthesis. Early works [4,63,78] decompose a full deformation into a set of local affine transformations, and introduce deformable skip connections [63] or pose-attentional transfer blocks [78] to manipulate deformations in a feature space. These methods estimate deformations solely based on 2D cues, leading to unrealistic results. Recently, Liu et al. [46] resort to 3D deformable models to estimate 2D deformations via projection, achieving superior results. We also exploit 3D deformable models to model spatial deformations in our network, where 3D deformations are projected to the 2D space to guide shape transfer. [62] presents a re-rendering method based on UV feature map prediction. However, predicting UV texture maps from images will involve potential kinematic problems which may affect our downstream reshaping.

Image Inpainting. Our task also involves hallucinating textures for dis-occluded areas in the background, where a direct warping based approach often introduces large distortions [27, 77]. We resort to recent image inpainting techniques to speculate realistic background using generative neural networks. Early image inpainting methods employ diffusion techniques [5,9,38] or patch match [6,7,15,16,36,65,71] to fill small holes based on low-level image features, having limited ability for hallucinating new textures when complex and non-repetitive textures are presented [75]. Deep generative based methods [25, 40, 44, 56, 73–76] resolve this problem by exploiting high-level semantics learned from a large amount of data. For example, contextual attention module [75] and dilated convolution [25] have been introduced to capture dependencies from long-range patches. Latest works [44, 76] suggest that vanilla convolutions used in [25, 75], which treat the pixels inside and outside a missing region equally, are sub-optimal for image inpainting. Hence, partial convolution [44] and gated convolution [76] have been employed to learn better feature encodings. The former one modulates feature learning dependent on an input mask in an un-learnable way, while the latter learns a dynamic feature gating mechanism conditioned on an input mask and achieves better performance. Thus, in our work, we exploit gated convolution for background encoding.

3 Method

We adopt a fit-then-reshape pipeline. Unlike previous methods [27, 77], which require manual tweaks, we *automate* the parametric fitting process by using an initial-fitting model learned from data, followed by a fine-tuning optimization step (Section 3.1). For realistic reshaping of both the foreground and the background, we introduce a novel *two-headed* architecture based on neural networks. Meanwhile, we present a self-supervised strategy to enable our reshaping training under the lack-of-data situation (Section 3.2). The key is to learn how to generate realistic images from input foreground and background images guided by warping fields.

3.1 SMPL Model Fitting

The Skinned Multi-Person Linear Model (SMPL) is a differentiable mapping from shape-parameters $\beta \in \mathbb{R}^{10}$ and pose-parameters $\theta \in \mathbb{R}^{72}$ to a 3D human model $M(\beta, \theta)$ with 6,890 mesh vertices and 24 joints. Fitting the SMPL to a human image essentially solves for β and θ that produce a body tightly aligns with the image under its camera setting. State-of-the-art learning-based methods offer fast and robust estimations of SMPL parameters in images, in this paper, we build our hybrid paradigm upon the method of [29].

Given a human image I, we obtain the initial shape and pose parameters (β^0, θ^0), and the camera parameters α with the pre-trained model of [29]. To further refine the SMPL parameters, we employ an optimization based paradigm to iteratively align the fitted model w.r.t. the image cues. The overall optimization consists of two steps: optimizing β and θ using 2D keypoints (following [34]) and optimizing

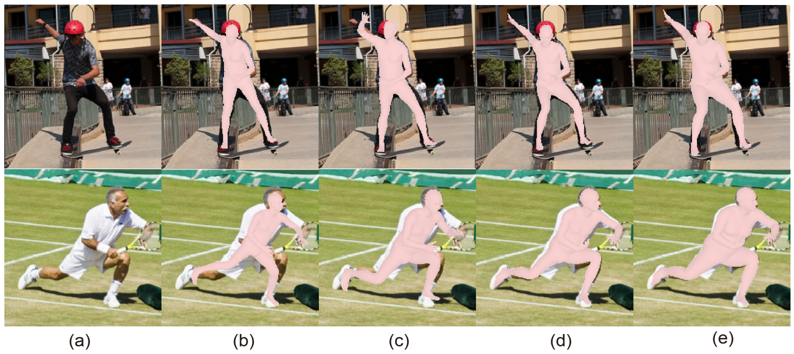

Figure 4 Fitting results obtained from different steps in the SMPL model fitting stage. (a) The original image. (b) The optimization result obtained from [10] (initialized with mean parameters). (c) The direct inference result obtained from the pre-trained model [29]. (d) The fitting result obtained after the 2D keypoints optimization step. (e) The fitting result obtained after the silhouette optimization step.

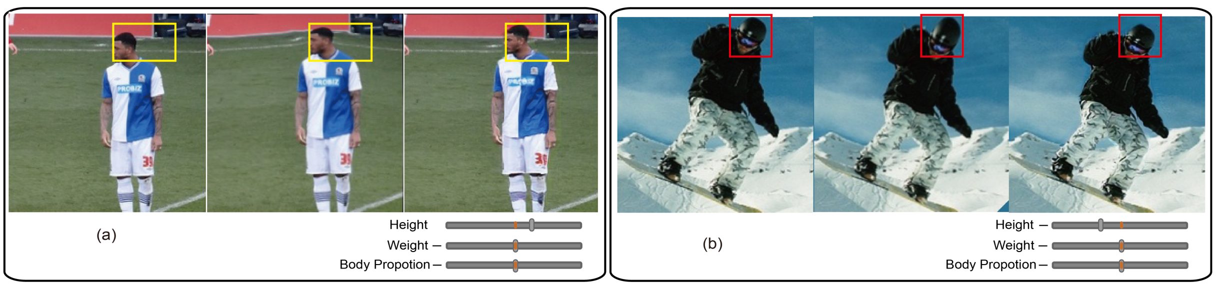

Figure 5 A comparison between direct warping with extrapolation and our method. From left to right: the original image, the reshaping result by direct warping, and the reshaping result by our method.

β for better silhouette matching. We keep α fixed throughout the optimization since the initial value of α predicted by [29] is already reliable.

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 [11] 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 [10] to reduce the optimization time while not harming the accuracy.

$$E_{\text{joint}}(\beta, \theta) = \sum_{\text{joint } i} w_i \rho \left(\Pi_\alpha M_{\text{joint}}(\beta, \theta) - J^i_{\text{est}} \right), \tag{1}$$

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 [21]). 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.

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

where M_{mesh} denotes the SMPL body mesh, and the operator $\mathrm{NR}_\alpha[\cdot]$ denotes the differentiable silhouette rendering [32] using the camera parameters α.

Fig. 4 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 [29] (Fig. 4 (b)) and [10] (Fig. 4 (c)), respectively. Note here we do not use the parametric model of SCAPE [2], which is used in [77], since the SMPL model is more accurate and compatible with modern rendering pipelines [48].

Thanks to the inherently decoupled shape and pose parameters in SMPL, users can intuitively get desired body shape by directly adjusting shape parameters β. 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. 6), 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 6 The reshaping results generated by our method with attributes adjusted individually.

For the resculpted 3D human model, we project the corresponding 3D deformation onto the image space to prepare a dense warping field T for the subsequent image reshaping stage. Given a dense warping field, a naive idea is to warp directly on pixel-level. However, a direct image warping approach based on T and its extrapolation from the human region to the entire image would easily lead to noticeable artifacts in both the background and the foreground (Fig. 5). In contrast, we choose to exploit T to guide the subsequent image synthesis via a neural generator to avoid distortions and obtain fine-grained details in both the foreground and the background.

3.2 NeuralReshaper

Ideally, our NeuralReshaper accepts a foreground body image I_f, a background image I_b which exhibits occluded and dis-occluded regions due to the potential deformation of the body, and the warping field T expressing the deformation, and should generate an output image with desired retouching result that inpaint the background regions. This idea is well followed with our network designs. As illustrated in Fig. 2, our network is a two-headed UNet-like structure containing two encoders and a decoder: A foreground encoder $\mathcal{E}_f(I_f)$ which consumes a foreground image I_f and a background encoder $\mathcal{E}_b(I_b, a)$ which consumes a background image I_b with masked regions (including occluded and dis-occluded areas induced by the deformation) and a union foreground mask a, and a decoder $\mathcal{D}$ which takes the encoded codes from both $\mathcal{E}_f$ and $\mathcal{E}_b$ and generates the final retouched result. 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 training stage, a discriminator is employed to enforce the overall realism of the generated image. Below we describe all the necessary components in details.

3.2.1 *Encoding*

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ of a person, we aim to synthesize a new image $I^t \in \mathbb{R}^{H\times W\times 3}$ of the same person but with the user-resculpted shape. To disentangle the complex foreground-background interactions, we introduce a two-headed UNet-like generator that separately encodes the information from foreground and background, respectively. For the foreground encoder $\mathcal{E}_f$, we extract the human region $I_f = I * m$ (m is the body mask) and employ vanilla convolutions [35] for feature encoding. As for the background encoder $\mathcal{E}_b$, we first compute a target body mask $m^t = warp(m, T)$ by warping the source body mask m (warping field derived from 3D SMPL transformation), and then compute a union foreground mask a:

$$a = m \cup m^t. \tag{3}$$

We send the concatenation of the masked background image $I_b = I * (1 - a)$ and the mask a itself to the background encoder. For encoding, we employ gated convolution [76], which learns a dynamic feature gating mechanism conditioned on the input mask. It modulates the input features at each channel and each spatial location according to the mask, providing more flexibility than alternatives like [44].

3.2.2 *Feature Integration and Decoding.*

To generate desired retouching result based on T, we introduce a novel warp-guided mechanism to integrate the features of the two encoders. Specifically, let f_i denote the intermediate feature produced by the i-th layer of the foreground encoder $\mathcal{E}_f$, i.e., $f_i = \mathcal{E}^i_f(I_f)$. We warp it to create a deformed feature $f^t_i = warp(f_i, T)$ (illustrated by the arrows in orange in Fig. 2), which is roughly aligned with the target shape. Then we add the warped foreground feature f_i^t with the corresponding background feature b_i to obtain a complete feature map φ_i of the target

$$\varphi_i = f_i^t + b_i. \tag{4}$$

We employ skip connections in the generator to consume the features $\varphi_1, \varphi_2, \ldots, \varphi_n$ ($n = 4$ is the number of layers in both encoders), similar to the vanilla UNet (see Fig. 2).

With the warping-aware integration strategy, the generator succeeds to produce a synthesized image $I_{\text{out}} = \mathcal{D}(\mathcal{E}_f(I_f), \mathcal{E}_d(I_b, a), T)$ with the person in a desired shape. Note that we only synthesize pixels within the mask a and leave the contents outside it untouched. Ultimately, we obtain the target image

$$I^t = a * I_{\text{out}} + (1 - a) * I, \tag{5}$$

where $*$ denotes the element-wise multiplication.

3.2.3 *Self-supervised Training Strategy*

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, such data is hard, if not impossible, to obtain in the real world. This lack-of-data problem also prevents us from employing the state-of-the-art image translation methods, e.g., CycleGAN and its successors [23, 24], since they usually require images from multiple domains and focus on learning transformations between two domains rather than two specific images. To this end, we introduce a novel self-supervised training strategy, where we exploit deformed source image as input and in seek to generate the original source image. In this way, the source image can serve naturally as supervision information without any additional annotation.

Specifically, as shown in Fig. 3, 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. 3.1. We then resculpt the fitted 3D model by randomly altering the shape parameters constrained in valid ranges to ensure the deformation plausibility. 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 as described in Sec. 3.2.1. 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.

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

$$L_R = \left\| I - G(I_b, I_f^t, T^t) \right\|_1. \tag{6}$$

We exploit the hinge loss [41] for GAN to enforce the synthesis realism of the generator. The loss for the generator is

$$L_G = -\mathbb{E}_{z \sim \mathbb{P}_z(z)}[D(G(z))], \tag{7}$$

where D is the spectral-normalized patch discriminator [76] whose loss function is formulated as

$$\begin{aligned} L_D = \mathbb{E}_{x \sim \mathbb{P}_{data}(x)}&[\mathrm{ReLU}(1 - D(x))] + \\ &\mathbb{E}_{z \sim \mathbb{P}_z(z)}[\mathrm{ReLU}(1 + D(G(z)))]. \end{aligned} \tag{8}$$

Overall, we obtain an alternating minimization:

$$\begin{aligned} &\min_G \, (\lambda_{\text{recovery}} L_R + \lambda_{\text{gan}} L_G) \\ &\min_D \, L_D \end{aligned} \tag{9}$$

where $\lambda_{recovery}$ and λ_{gan} are trade-off parameters for different losses. Note that, to enforce the generator's synthesis realism, we adopt a spectral-normalized patch discriminator [76] to train our generator in an adversarial way.

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. 3), 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. 14 for comparisons between our approach and an alternative solution by directly warping in the image space (instead of the feature space).

Figure 7 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.

In the testing phase, we take (I_b, I_f, T) as the input, the generator should output a synthesized image with the human in a user-desired shape. Since the mask obtained from Mask R-CNN does not match the fitted area perfectly due to the imperfections of fitting, we diffuse the warping field from the fitted area to the whole mask following the warping method in Geng et al. [18]. Specifically, we sample 300 points on the contour of the fitted mask and image edges, and triangulate the whole image exploiting Delaunay triangulation. We fix the points on the image edges and let the points on the contour to drive deformations to complete the deformations in unfitted areas.

4 Experiments

4.1 Dataset

We have conducted extensive experiments on an indoor dataset DeepFashion [47] (512×512 resolution) and an outdoor dataset consisting of images from COCO [43], MPII [1], and LSP [28] (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.

As mentioned in Sec. 3.1, we run OpenPose [11] and Mask R-CNN [21] to obtain the 2D keypoints and human masks. 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. We use 85% of them for training and the rest for testing. For the outdoor dataset, we have $29,353$ valid images left ($24,806$ for training and the rest for testing), which are cropped and resized to 256×256.

4.2 Implementation Details

4.2.1 *Fitting Stage*

The initial SMPL fitting network [29, 34] cannot handle images of varying input size, we crop and resize all training images into 224×224. For the optimization step, we use Adam [33] 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.

Figure 8 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).

4.2.2 *Reshaping Stage*

The UNet-like generator (Fig. 2) 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 [75]. We use 6 layers of residual blocks in the bottleneck and 4 layers for the decoder. The spectral-normalized [52] 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 [33]. 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 MaskRCNN [21], keypoint detection by OpenPose [11], and SMPL fitting optimization. Then the modification of the parameters to derive reshaping results runs at an instant rate (less than 1s).

4.3 Qualitative Results

4.3.1 *Results on the DeepFashion Dataset*

Fig. 7 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. 6 illustrates the individual effect of each attribute, showing the well-disentangled reshaping semantics.

Figure 9 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.

4.3.2 *Results on the Outdoor Dataset*

We also show results on outdoor dataset where complex occlusions and backgrounds can present. Fig. 8 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 [76]. 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. 8 shows that our approach provides fine-grained and holistic reshaping effects. The last row of Fig. 8 exemplifies reshaped body images with challenging poses, severe occlusions, and complex backgrounds.

4.3.3 *Results on the Online Images*

Results, exhibited in Fig. 9, 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 [11] and Mask R-CNN [21]). Our method is applicable to a wide range of real situations.

4.4 Comparisons

In this section, we compare our approach to the warping-based human reshaping method [77], the state-of-the-art human image editing method [46] using deep learning, and an alternative solution that combines foreground warping with background inpainting.

Fig. 10 shows that the reshaping method [77] introduces distortions in the background, leading to unrealistic and implausible results. As suggested in their paper, one possible strategy to alleviate such distortion artifacts is to manually adjust the saliency map. This, however, involves dense human intervention. In contrast, our method is fully automatic, and the decomposition of foreground and background information ensures realistic results.

The human image editing method [46] 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 [43]. We compare our method with [46] on their dataset, COCO dataset and online images, respectively. Fig. 11 (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. 11 (b), their method leads to seriously degraded results when facing non-seen cases. We also do quantitative comparison with [46] in Table. 1. 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.

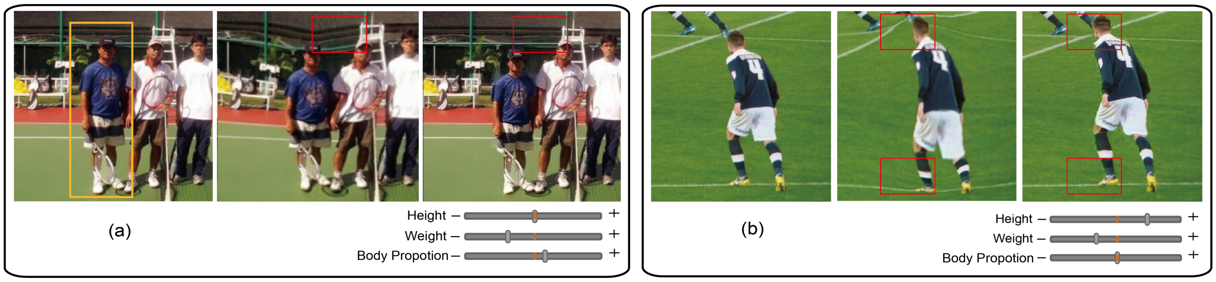

Figure 10 Comparison with Zhou et al. [77]. For each case, from left to right are the original image, the reshaping result by the method in [77], and the reshaping result by our method. The sliders below each case indicate the attributes and their degrees that have been edited. The person that has been reshaped is highlighted with a yellow bounding box in (a). We also highlight the distorted areas in [77] and their corresponding patches in our result with red bounding boxes.

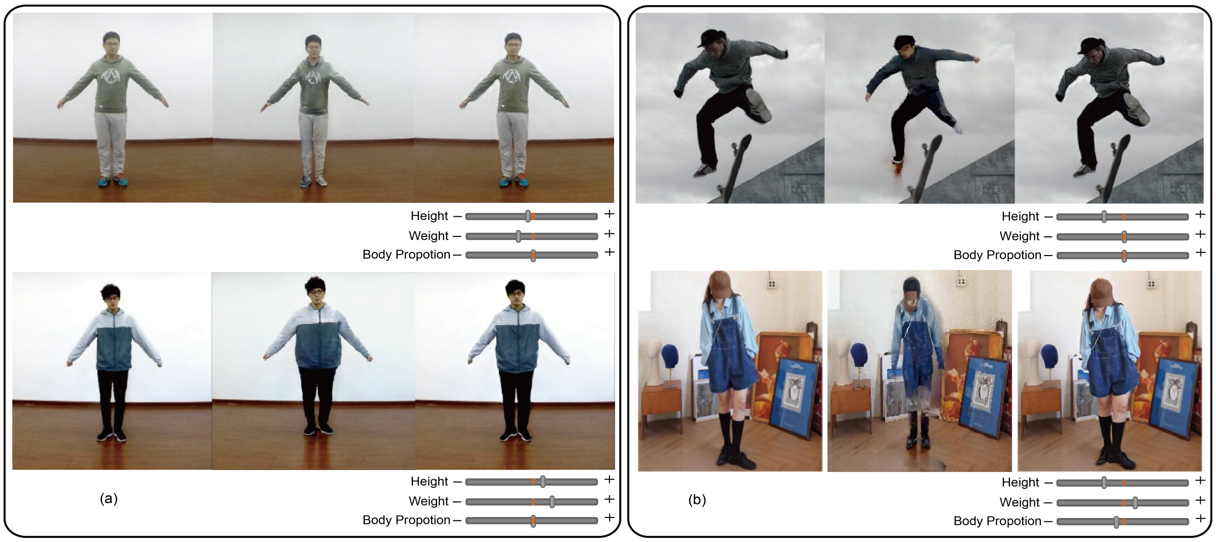

Figure 11 Comparison with Liu et al. [46]. For each case, from left to right are the original image, the reshaping result by the method in [46], and the reshaping result by our method.

Method	FID ↓
Liquid Warping [46]	89.41673
Ours	**80.28321**

Table 1 Quantitative evaluation of the generated images with the retouched new poses.

4.5 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. 12, 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.

In reshaping stage, we replace the gated convolution with vanilla convolution for the background branch. We denote the resulting model as variant **G-**. Fig. 13 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 a popular alternative that merges two feature maps in a mask-guided

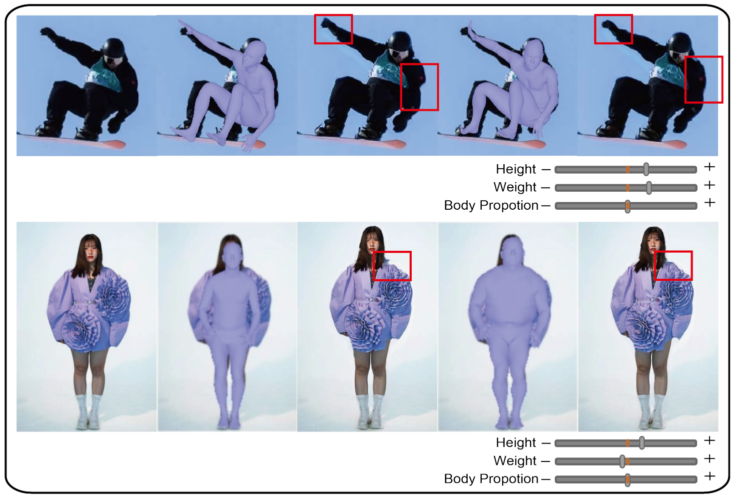

Figure 12 Ablation study on the SMPL fitting optimization. The first row shows the fitted SMPL model and synthesized image without 2D keypoints optimization and the corresponding results with our full optimization. The second row shows the comparison on the silhouette optimization.

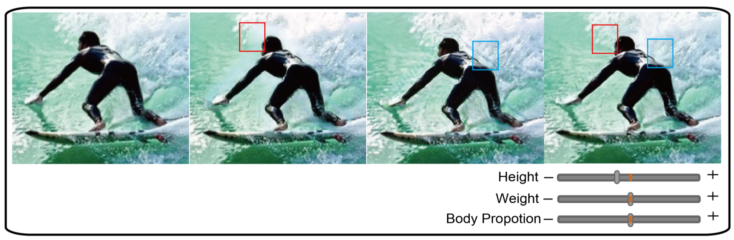

Figure 13 Ablation study. For each case, from left to right are the original image, the reshaping result generated by **G-**, the result by **M+**, and the result by our full model. The sliders below each case indicates the attributes and their degrees that have been edited. Red and blue rectangles are used to highlight the areas with artifacts.

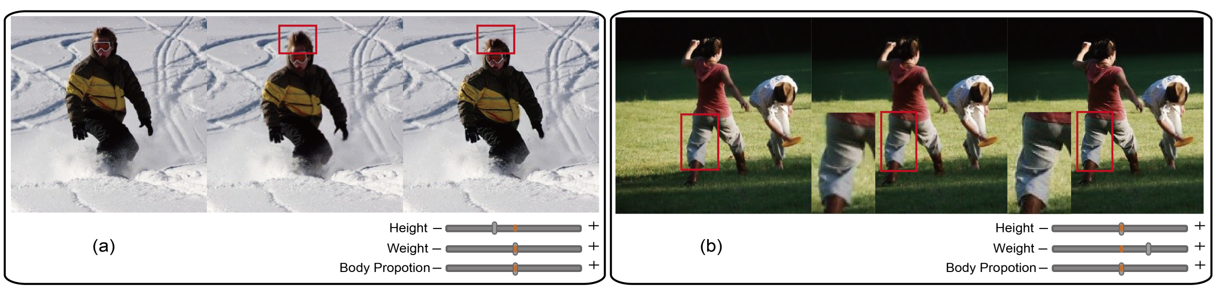

Figure 14 Comparison with an alternative approach by combining direct warping on foreground with background inpainting. For each case, from left to right are the original image, the reshaping result by the alternative approach and the reshaping result by our method.

way [68]. Specifically, we compute a mask for the warped foreground, denoted as m^t. We then integrate the features as

$$\varphi_i = b_i + f_i^t * m^t, \tag{10}$$

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. 10 in the full model to obtain the variant **M+**. Fig. 13 shows that the mask-guided merging tends to introduce artifacts around synthesized humans (see the area highlighted with a blue rectangle).

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 [18] and employ the network presented in [76] for inpainting. Fig. 14 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. 14 (a)) or blurriness (Fig. 14 (b)).

5 Limitations

Our method has a few limitations. Our method might introduce artifacts under extreme deformations (see Fig. 15 (a)). The reasons are two-fold. First, the resculpted 3D model might become implausible,

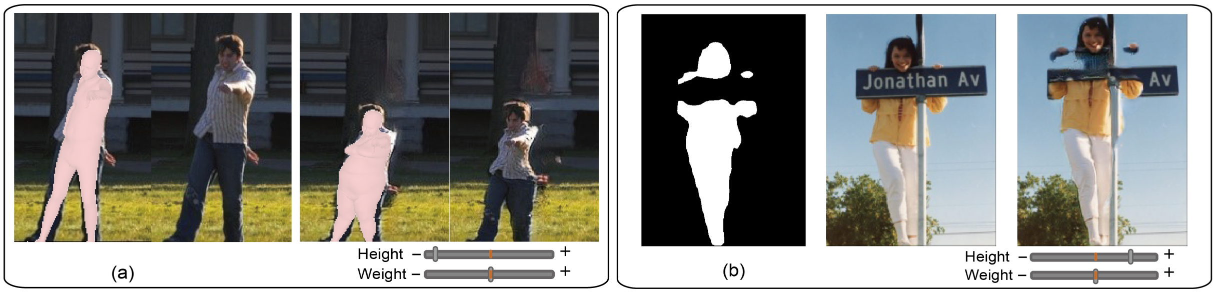

Figure 15 Failure results generated by our method. (a) Synthesized results under extremely large deformations. (b) Synthesized results when the interacting objects lie outside the foreground mask.

leading to unnatural image warping effects for body and/or face areas. Second, 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 SMPL model often fails at fine-grained regions, such as hands, which may lead to undesirable synthesized results at those regions (Fig. 7, 8, 9). A more fine-grained parametric model [57] might help.

Our method deals well with parts tightly coupled with the body, like cloth, hat, or hair, since Mask R-CNN [21] 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. 15 (b)). To appropriately represent the interactions between objects, we could introduce the scene graphs into the generation process.

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 [70] 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.

6 Conclusion

We have presented NeuralReshaper, a practical method for realistic reshaping of human bodies in single images using deep generative networks. Our method enables users to reshape human images by moving several sliders with instant feedback. Although we share the fit-then-reshape pipeline with Zhou et al. [77], we improve the whole process in several aspects, including a fully-automatic human-to-image fitting pipeline and more realistic edited results in various situations. An essential ingredient of our pipeline lies in employing GANs for realistic reshaping instead of applying direct-warping like the previous methods. To enhance the robustness, we design a two-headed UNet-like generator for orthogonal feature-encoding of the background and the foreground, respectively. We integrate the features extracted from both branches in a warp-then-merge manner, resolving the vanilla UNet's inherent misalignment issue. Our generator performs inpainting of the dis-occluded areas and synthesizes the foreground in a single UNet, resulting in a compact yet powerful network. Moreover, we also introduce a novel self-supervised strategy to solve the lack-of-data problem. Extensive results on the DeepFashion dataset, outdoor datasets, and online images have demonstrated our method's superiority compared with alternative solutions. What's more, We believe that our method can serve as automatic dataset generation for future supervised learning based methods.

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 Baek K, Choi Y, Uh Y, et al. Rethinking the truly unsupervised image-to-image translation. arXiv preprint arXiv:200606500, 2020

4 Balakrishnan G, Zhao A, Dalca A V, et al. Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8340–8348
5 Ballester C, Bertalmio M, Caselles V, et al. Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing, 2001. 10: 1200–1211
6 Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM transactions on graphics (TOG), 2009. 28: 24
7 Barnes C, Shechtman E, Goldman D B, et al. The generalized patchmatch correspondence algorithm. In: European Conference on Computer Vision. Springer, 29–43
8 Bau D, Strobelt H, Peebles W, et al. Semantic photo manipulation with a generative image prior. arXiv preprint arXiv:200507727, 2020
9 Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. 417–424
10 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
11 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
12 Chen R T, Li X, Grosse R B, et al. Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. 2610–2620
13 Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 764–773
14 Doersch C. Tutorial on variational autoencoders. arXiv preprint arXiv:160605908, 2016
15 Efros A A, Freeman W T. Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. 341–346
16 Efros A A, Leung T K. Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE International Conference on Computer Vision. IEEE, volume 2, 1033–1038
17 Esser P, Sutter E, Ommer B. A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8857–8866
18 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
19 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems. 2672–2680
20 Guan P, Weiss A, Balan A O, et al. Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, 1381–1388
21 He K, Gkioxari G, Dollár P, et al. Mask r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2961–2969
22 Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations
23 Hoffman J, Tzeng E, Park T, et al. Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning. PMLR, 1989–1998
24 Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, 172–189
25 Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion. ACM Transactions on Graphics (TOG), 2017. 36: 1–14
26 Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1125–1134
27 Jain A, Thormählen T, Seidel H P, et al. Moviereshape: Tracking and reshaping of humans in videos. ACM Transactions on Graphics (TOG), 2010. 29: 1–10
28 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
29 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
30 Kanazawa A, Zhang J Y, Felsen P, et al. Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5614–5623
31 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4401–4410
32 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
33 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980, 2014
34 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
35 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012. 25: 1097–1105
36 Kwatra V, Essa I, Bobick A, et al. Texture optimization for example-based synthesis. In: ACM SIGGRAPH 2005 Papers, 2005. 795–802
37 Lassner C, Romero J, Kiefel M, et al. Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6050–6059
38 Levin A, Zomet A, Weiss Y. Learning how to inpaint from global image statistics. In: Proceedings Ninth IEEE International Conference on Computer Vision. 305–312
39 Li Y, Huang C, Loy C C. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3693–3702
40 Li Y, Liu S, Yang J, et al. Generative face completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3911–3919
41 Lim J H, Ye J C. Geometric gan. arXiv preprint arXiv:170502894, 2017
42 Lin C H, Yumer E, Wang O, et al. St-gan: Spatial transformer generative adversarial networks for image compositing. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9455–9464
43 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
44 Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, 85–100
45 Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks. arXiv preprint arXiv:170300848, 2017
46 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
47 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
48 Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 2015. 34: 1–16
49 Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2794–2802
50 Metz L, Poole B, Pfau D, et al. Unrolled generative adversarial networks. arXiv preprint arXiv:161102163, 2016
51 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:14111784, 2014
52 Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:180205957, 2018
53 Nidhi A, Kumar H, Dhaliwal J S, et al. Improved Interactive Reshaping of Humans in Images. In: International Conference on Computer Graphics. IEEE, 230–238
54 Omran M, Lassner C, Pons-Moll G, et al. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). IEEE, 484–494
55 Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2337–2346
56 Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2536–2544
57 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
58 Pavlakos G, Zhu L, Zhou X, et al. Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 459–468
59 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
60 Richter M, Varanasi K, Hasler N, et al. Real-time reshaping of humans. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission. IEEE, 340–347
61 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, 234–241
62 Sarkar K, Mehta D, Xu W, et al. Neural re-rendering of humans from a single image. In: Proceedings of the European conference on computer vision (ECCV). Springer, 596–613
63 Siarohin A, Sangineto E, Lathuiliere S, et al. Deformable gans for pose-based human image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3408–3416
64 Sigal L, Balan A, Black M J. Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Advances in Neural Information Processing Systems. 1337–1344
65 Simakov D, Caspi Y, Shechtman E, et al. Summarizing visual data using bidirectional similarity. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 1–8
66 Sønderby C K, Raiko T, Maaløe L, et al. Ladder variational autoencoders. In: Advances in Neural Information Processing Systems. 3738–3746
67 Sun X, Xiao B, Wei F, et al. Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, 529–545
68 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
69 Tung H Y, Tung H W, Yumer E, et al. Self-supervised learning of motion capture. In: Advances in Neural Information Processing Systems. 5236–5246
70 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
71 Wexler Y, Shechtman E, Irani M. Space-time completion of video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007. 29: 463–476
72 Wu R, Tao X, Gu X, et al. Attribute-driven spontaneous motion in unpaired image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 5923–5932
73 Yang C, Lu X, Lin Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6721–6729
74 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:151107122, 2015
75 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
76 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
77 Zhou S, Fu H, Liu L, et al. Parametric reshaping of human bodies in images. ACM transactions on graphics (TOG), 2010. 29: 1–10
78 Zhu Z, Huang T, Shi B, et al. Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2347–2356