SCIENCE CHINA
Information Sciences

• LETTER •

September 2023, Vol. 66 199101:1–199101:2
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* 310000*, China;*
[2]*School of Creative Media, City University of Hong Kong, Hong Kong* 999077*, 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 non-rigid 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 [1] 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 [2], we present NeuralReshaper (Figure 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) [3] 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 reshaping 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 studies 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 fine-tuning optimization step. For realistic reshaping, we introduce a novel two-headed neural architecture.

SMPL 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 [4]. 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 [5]) 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 Figure 1, our network

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

© Science China Press 2023

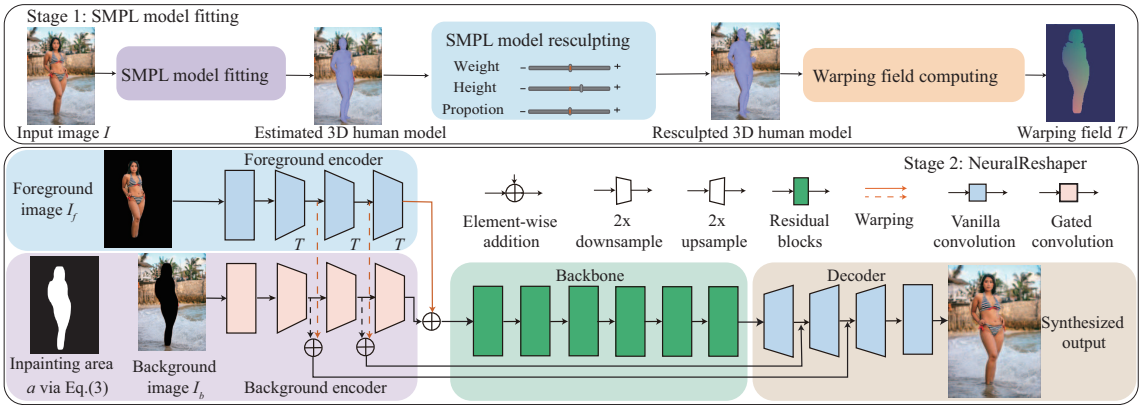

Figure 1 (Color online) Overall pipeline of our proposed method, consisting of model fitting and neural reshaping stages.

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 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 $f_i^t = \text{warp}(f_i, T)$ (shown by the arrows in orange in Figure 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,

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

We used the hinge loss [2] for GAN. Overall, we obtain an alternating minimization:

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

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 National Natural Science Foundation of China (Grant No. 62172363).

Supporting information Appendixes 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 Zhou S, Fu H, Liu L, et al. Parametric reshaping of human bodies in images. ACM Trans Graph, 2010, 29: 1–10
2 Lim J H, Ye J C. Geometric GAN. 2017. ArXiv:1705.02894
3 Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model. ACM Trans Graph, 2015, 34: 1–16
4 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, 2018. 7122–7131
5 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, 2019. 2252–2261