Simple Techniques for a Novel Human Body Pose Optimisation Using Differentiable Inverse Rendering

M. Battogtokh [1] and R. Borgo[1]

[1]King's College London, United Kingdom

Abstract

Human body 3D reconstruction has a wide range of applications including 3D-printing, art, games, and even technical sport analysis. This is a challenging problem due to 2D ambiguity, diversity of human poses, and costs in obtaining multiple views. We propose a novel optimisation scheme that bypasses the prior bias bottleneck and the 2D-pose annotation bottleneck that we identify in single-view reconstruction, and move towards low-resource photo-realistic 3D reconstruction that directly and fully utilises the target image. Our scheme combines domain-specific method SMPLify-X and domain-agnostic inverse rendering method redner, with two simple yet powerful techniques. We demonstrate that our techniques can 1) improve the accuracy of the reconstructed body both qualitatively and quantitatively for challenging inputs, and 2) control optimisation to a selected part only. Our ideas promise extension to more difficult problems and domains even beyond human body reconstruction.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction; Computer vision; Rendering; Ray tracing;***

1. Introduction

Human body 3D reconstruction has applications in 3D-printing, virtual/augmented reality, games, film, sport analysis, and more. This is a challenging problem due to 2D ambiguity and the diversity of human body poses [KBJM18, PCG*19].

In general, existing approaches to 3D reconstruction are either domain-agnostic or -specific. The former includes stereo vision [Ali19], which requires multiple images that may be simply unavailable or expensive. On the other hand, the latter often requires only a single image but remains challenged by images with rare/extreme poses. This is because they (especially learning approaches [GLK*20, HXL*20, JMT18]) generally rely on existing datasets of 3D bodies, which naturally have a limited distribution of poses. Furthermore, the non-learning techniques [PCG*19, AMX*18] effectively ignore the source image once a 2D-pose annotation is obtained, making them vulnerable to faulty estimations and blind to other features.

Unlike existing works, we propose to use a single source image *directly* as the target in human body reconstruction and optimise using an emerging domain-agnostic approach that is free of learnt biases, called inverse rendering [LADL18], combined with a domain-specific technique, SMPLify-X [PCG*19]. We propose our novel optimisation scheme accompanied by our two simple yet powerful techniques: *upstream parameter optimisation* and *selective inverse rendering*. We demonstrate our techniques by case studies on challenging in-the-wild images from judo, in which out-of-distribution poses due to extreme dynamism of the sport are common and motivate the need for an alternative to the existing methods. We achieve quantitative/qualitative improvements in the accuracy of reconstructed body poses over SMPLify-X baseline and over a concurrent work that uses stronger (kinematic) constraints [SBC21]. We also demonstrate how to optimise only a selected part without affecting the rest. In summary, our contributions are as follows:

1. A novel optimisation scheme for 3D human body reconstruction that uses the source image directly as a target. Bypasses reliance on prior biases or 2D annotations and opens a way towards utilising the source image fully, improving reconstruction accuracy.
2. *Selective inverse rendering*: a novel "select-by-looking" approach to limit the optimisation to an intended part of the body.
3. *Upstream parameter optimisation*: a novel approach to optimise an arbitrary upstream parameter of a 3D scene parameter.

2. Related works

Our work combines the expressive body shape and pose modelling technique, *SMPLify-X* [PCG*19], with the differentiable rendering solution: *redner* [LADL18]. SMPLify-X captures expressive pose features including hand and face pose by optimising a parameterised body model against estimated 2D joints, but is blind to other features, e.g., clothing, hair, and colour. SMPLify-X (like [AMX*18, KBJM18]) relies on 2D pose annotation as optimisation target, making quality of pose estimation the bottleneck to reconstruction quality. Furthermore, their optimisation is constrained by a learnt body pose prior, which limits generalisation to rare out-of-distribution poses.

DOI: 10.2312/egs.20221018

Alternatively, learning (e.g., [GLK*20, HXL*20]) approaches are recently common, but they are generally limited for rare poses due to reliance on training data. High-quality training dataset is obtained through 3D scans in [JMT18], but it is expensive and limited to restricted poses. Work by [SBC21] enforces human kinematic constraints and enriches training with synthetic data like [XZT19]. However, real and out-of-distribution data remain challenging.

An emerging approach to supervise reconstruction without prior biases while using the full information of the source image is *inverse rendering*, which refers to inferring 3D scene parameters like geometry, material, and camera from a target 2D image [LHJ19]. Gradient-based solutions [LADL18, LHJ19] do this with differentiable rendering by re-projecting 3D scenes to 2D images and back-propagating error (from comparison with the target) to scene parameters. Recently, [LHJ19] showed impressive results optimising approximate 3D models to photo-realistic reconstructions. Unfortunately, the reliance on initial approximation is a limitation.

2.1. Similar works

The work by [LLCL19] briefly demonstrates human body fitting with novel differentiable renderer, but within a limited setting without details of the underlying theory/method. The main limitation is that their generic learning-based reconstruction relies on training data, which may be unavailable. ARCH [HXL*20] uses differentiable rendering to recover surface details of a volumetric body model but not the pose. Our work differs by proposing to use off-the-shelf solutions for a challenging domain with extreme body poses and describing the theory/method behind how to apply them in detail. Other works have also explored inverse rendering but only for faces [DBA*21] or hands [KKOT21]. Uniquely, we do so for the whole human body and optimise upstream parameters. Furthermore, we are the first to showcase selective inverse rendering.

3. Method

Related methods suffer from the following major problems (see 2):

1. Rare out-of-distribution poses
2. Bottleneck due to reliance on 2D-pose annotation
3. Blindness to other features (e.g., clothing, and colour)

We solve these problems simultaneously by combining domain-specific human body modelling [PCG*19] with domain-agnostic inverse rendering [LADL18] into a novel optimisation scheme. This is possible because inverse rendering (*redner*) supervises reconstruction with the source image directly, as aforementioned in 2. In 3.1, we present the method behind the optimisation scheme. Then, we present the idea of selective inverse rendering in 3.2.

3.1. Upstream parameter optimisation

With differentiable rendering, we can infer 3D scene parameters $\boldsymbol{\Phi}$ (e.g. geometry and camera pose) from a target 2D image $\boldsymbol{I}_{target}$, by rendering a synthetic image $\boldsymbol{I}$, from which we subsequently calculate a scalar loss L (Equation 1) and back-propagate gradients.

$$L(\boldsymbol{I}_{target}, \boldsymbol{I}) = \|\boldsymbol{I}_{target} - \boldsymbol{I}\|_2^2 \tag{1}$$

We use the differentiable rendering solution *redner* [LADL18] to create a synthetic image $\boldsymbol{I}$ (Equation 2).

$$\boldsymbol{I} = redner(\boldsymbol{\Phi}); \;\; \boldsymbol{\Phi} = (\boldsymbol{\Phi}_{geometry}, \boldsymbol{\Phi}_{camera}, ...) \tag{2}$$

Our key insight for 3.1 is that we can optimise parameters beyond just the scene parameters $\boldsymbol{\Phi}$ if these themselves are parameterised by and differentiable with respect to upstream parameters. Generally, this is because chain-rule allows calculating the partial derivative of $\boldsymbol{I}$ with respect to an upstream parameter $\boldsymbol{\alpha}_i$, for a scene parameter $\boldsymbol{\Phi}_i(\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ...\boldsymbol{\alpha_n})$ parameterised by upstream parameters $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ...\boldsymbol{\alpha}_n$.

In case of rendering the human body reconstruction of SMPLify-X, since the scene geometry $\boldsymbol{\Phi}_{geometry}$ is the reconstructed human body $\boldsymbol{M}$ parameterised by upstream parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}$ [PCG*19] for pose, shape, and facial expression respectively (Equation 3), it is possible to compute the partial derivative of L with respect to e.g., body pose $\boldsymbol{\theta}_b$ using the chain rule in Equation 4.

$$\boldsymbol{\Phi}_{geometry} = \boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{3}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta_b}} = \frac{\partial L}{\partial \boldsymbol{I}} \frac{\partial \boldsymbol{I}}{\partial \boldsymbol{M}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{\theta}_b} \tag{4}$$

In practice, since both SMPLify-X and *redner* are differentiable with respect to their parameters and implemented in PyTorch, computing the derivative in Equation 4 is handled by automatic differentation. This gives us the advantage that our novel optimisation process can be implemented simply: we pipeline SMPLify-X and *redner* together and instruct PyTorch to backpropagate gradients to the body pose $\boldsymbol{\theta}_b$.

We formulate inferring the body pose $\boldsymbol{\theta}_b$ as constrained optimisation problem and seek to minimize the objective function:

$$\min_{\boldsymbol{\theta}_b, \boldsymbol{\alpha}} L(\boldsymbol{I}_{target}, \boldsymbol{I}) + \lambda E(\boldsymbol{\theta}_b) \tag{5}$$

where $E(\boldsymbol{\theta}_b)$ is a simple squared L_2 norm prior of the body pose $\boldsymbol{\theta}_b$ and λ is the weight of the regularisation term. Note that we also optimise an additional upstream parameter for camera angle $\boldsymbol{\alpha}$, which we define to change the camera pose $\boldsymbol{\Phi}_{camera}$ scene parameter without changing the camera distance.

We simultaneously optimise the body pose $\boldsymbol{\theta}_b$ and the camera angle $\boldsymbol{\alpha}$ using stochastic gradient descent (SGD) for 200 iterations starting from an approximate camera pose with regularisation weight $\lambda = 3$ unless otherwise stated. We use two SGD optimisers for $\boldsymbol{\theta}_b$ and $\boldsymbol{\alpha}$ with momentum 0.9 and learning rates 1×10^{-4} and 1×10^{-8} respectively.

3.2. Selective inverse rendering

With *selective inverse rendering*, we refer to inverse rendering with the purpose of optimising only a certain part of the body. Doing so

is motivated by the observation that it is often useful to optimise one part in isolation from the other parts, e.g., when we do not want the optimisation to affect a part of the body that is already in the right position. To our knowledge, we are the first to address such cases.

Practically, our question is: how do you "select" the vertices and triangles constituting the intended part of a given 3D model? Our attempt to the answer is intuitive and simple. It is to simply set the scene parameter for camera pose $\mathbf{\Phi}_{camera}$ such that only the selected part is visible in the rendering. In other words, our answer is to quite literally "look".

In theory, this selecting-by-looking is possible because the computation of a pixel of the rendered image $\boldsymbol{I}(x,y)$ is traceable (by the definition of ray tracing) to only those triangles of the mesh that are visible (directly or through refraction/reflection) to the rendering camera. This means that the triangles that are not visible to the camera and therefore do not contribute to $\boldsymbol{I}$ will receive no error backpropagated to them. This is useful because we can "select" parameters (including upstream parameters) that are responsible for some selected parts of the rendering (e.g., the hands or the face) by "pointing" the camera at those parts without worrying about vertices and triangles, or the model architecture.

4. Case study

4.1. Challenging domain - Judo

We apply our techniques to judo images, which have close-contact people with extreme (e.g., twisted, or upside-down) poses. We may want to 3D reconstruct judo scenes for artistic purposes or technical analysis. The rarity of the poses makes learning approaches and the 2D-pose estimation (which SMPLify-X relies on) unreliable and motivates the need for a novel method. Unfortunately, the rarity also means absence of annotations and limited evaluation [SBC21].

4.2. Pose optimisation results and evaluation

Figure 1 shows results obtained by our novel optimisation method. We clamp image pixels between zero and one before calculating the loss in Equation 1. Figure 1 second row uses looser regularisation with weight $\lambda = 2$ (see Equation 5) to allow optimisation to consider more "extreme" poses the target requires. The results achieve clear qualitative improvements including more hunched pose for Figure 1 first row, twisted legs with raised arms for the second, and forward lean with raised right leg for the third, all of which are more accurate and crucial to identifying the judo moves for technical analysis. We quantitatively compare our method to the SMPLify-X baseline and to a concurrent work that utilises human kinematic constraints [SBC21] on the judo images. Table 1 shows standard L_2 normal re-projection errors (like [HXL*20,PCG*19] at dense pixel-level) regularised by the image dimensions. We achieve up to 45.5% error reduction from SMPLify-X baseline and up to 22.3% error reduction from the concurrent work [SBC21].

4.3. Selective inverse rendering

To first demonstrate our novel selective inverse rendering method, we optimise without constraints ($\lambda = 0$) the scene parameter geometry $\mathbf{\Phi}_{geometry}$ directly (rather than its upstream parameter like body pose $\boldsymbol{\theta}_b$) in two conditions: *full* and *selective*. In the *full* condition, we render the entire body whereas in the *selective*, we render the lower body only. The target images are the corresponding parts. Figure 2 shows the results that the selective condition has effectively controlled optimisation to only the lower body, which has prevented the unwanted changes we see in Figure 2d upper row.

(a) *Target* **(b)** *Iter. 1* **(c)** *Error* **(d)** *Iter. 200* **(e)** *Error*

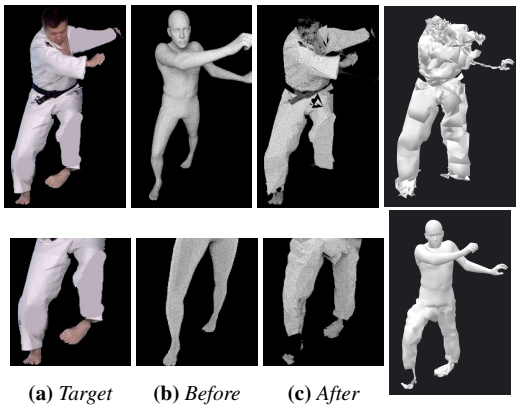
Figure 1: *Each column shows a) the raw target b) rendering of faulty SMPLify-X reconstruction before optimisation at 1st iteration c) error at the 1st iteration d) rendering at 200th iteration and e) error after the entire optimisation at the 200th iteration.*

Target image	Ours	SMPLify-X	[SBC21]
Figure 1a top	**15.7**	28.8	20.2
Figure 1a mid	**18.1**	24.7	20.1
Figure 1a bottom	**19.4**	29.6	20.4
Figure 3a	**19.6**	24.9	23.9

Table 1: *L_2 normal error as percentage of the product of image dimensions (height, width, and number of channels).*

Although the rendering in the top row Figure 2c is incredibly photo-realistic to the target image, the reconstructed 3D mesh appears unintelligible when rendered differently, especially the upper body (Figure 2d top). The bottom row has selectively optimised the lower body to generate much more intelligible result by preserving the upper body (contrast Figure 2d top and bottom). This demonstrates selective inverse rendering is a promising novel approach to isolate a selected part and regulate the other parts.

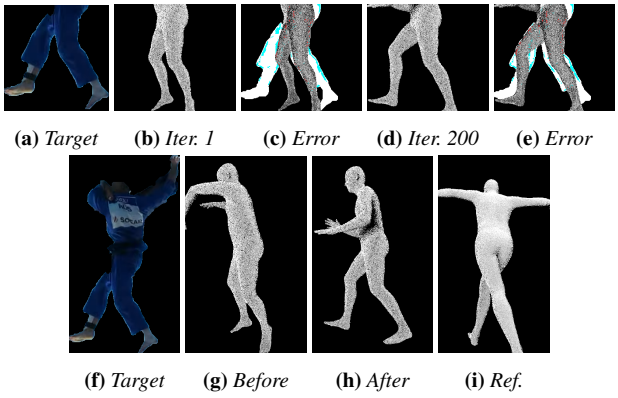
We also demonstrate selective inverse rendering for optimising an upstream parameter, the body pose $\boldsymbol{\theta}_b$. Figure 3 shows the results of selectively optimising the pose of the lower body. The lower body error is reduced by 21.3% and a more accurate lower body



(a) *Target* **(b)** *Before* **(c)** *After* **(d)** *3D mesh*

Figure 2: *Selective inverse rendering of geometry*

pose than non-selective optimisation result is achieved. Figure 3h achieves *lower body* pose more accurate to the target image (Figure 3f) than in Figure 3i, with intentionally less changes in the upper body. Note, the upper body is not fully isolated from the lower body due to 1) the pose prior regularisation term in our objective function (Equation 5) and 2) the lack of disentangling in the pose embedding that we adopted from SMPLify-X. To improve the isolation, we reduced the regularisation term weight to $\lambda = 1$. More work is needed to investigate how to better isolate the selected part during optimisation of parameters with such entangled latent embeddings.



(a) *Target* **(b)** *Iter. 1* **(c)** *Error* **(d)** *Iter. 200* **(e)** *Error*

(f) *Target* **(g)** *Before* **(h)** *After* **(i)** *Ref.*

Figure 3: *Selective inverse rendering of upstream parameter*

5. Conclusion and future work

We have presented a novel optimisation scheme for human body reconstruction along with two simple yet powerful techniques. We build upon existing works by combining complementary and distantly-related reconstruction approaches towards photo-realistic reconstruction that captures all the features of its source/target image. Our novel optimisation scheme utilises the source image directly as the target and automatically optimises the 3D body pose through *upstream parameter optimisation*. We demonstrated that the new scheme can improve reconstructions both qualitatively and quantitatively (by as much as nearly halving the error from baseline). Moreover, we demonstrated that our intuitive selective inverse rendering can successfully control optimisation to a selected part and intentionally regulate changes in the other parts.

Ultimately, this work pushes towards low-resource and photo-realistic 3D reconstruction. In the future, we would like to incorporate kinematic constraints into our approach [GLK*20] and extend our work to different domains (even beyond human bodies) and different upstream parameters including shape, colour, and material, as well as investigate how to better isolate optimisation when an upstream parameter is latent and entangled.

References

[Ali19] ALICEVISION: Meshroom. https://alicevision.org/#meshroom, 2019. accessed November 2019. 1

[AMX*18] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Video Based Reconstruction of 3D People Models. In *CVPR* (2018), pp. 8387–8397. 1

[DBA*21] DIB A., BHARAJ G., AHN J., THÉBAULT C., GOSSELIN P., ROMEO M., CHEVALLIER L.: Practical Face Reconstruction via Differentiable Ray Tracing. *Computer Graphics Forum 40*, 2 (5 2021), 153–164. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.142622. 2

[GLK*20] GEORGAKIS G., LI R., KARANAM S., CHEN T., KOŠECKÁ J., WU Z.: Hierarchical Kinematic Human Mesh Recovery. *LNCS 12362 LNCS* (8 2020), 768–784. URL: https://link.springer.com/chapter/10.1007/978-3-030-58520-4_45. 1, 2, 4

[HXL*20] HUANG Z., XU Y., LASSNER C., LI H., TUNG T.: ARCH: Animatable Reconstruction of Clothed Humans. In *CVPR* (2020), pp. 3093–3102. 1, 2, 3

[JMT18] JACKSON A. S., MANAFAS C., TZIMIROPOULOS G.: 3D Human Body Reconstruction from a Single Image via Volumetric Regression. In *ECCV Workshops* (2018), pp. 0–0. 1, 2

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end Recovery of Human Shape and Pose Input Reconstruction. In *CVPR* (2018). URL: https://akanazawa.github.io/hmr/. 1

[KKOT21] KARVOUNAS G., KYRIAZIS N., OIKONOMIDIS I., TSOLI A.: Multi-view Image-based Hand Geometry Refinement using Differentiable Monte Carlo Ray Tracing. *arXiv* (2021). 2

[LADL18] LI T.-M., AITTALA M., DURAND F., LEHTINEN J.: Differentiable monte carlo ray tracing through edge sampling. *SIGGRAPH Asia 37*, 6 (2018), 222:1–222:11. 1, 2

[LHJ19] LOUBET G., HOLZSCHUCH N., JAKOB W.: Reparameterizing discontinuous integrands for differentiable rendering. *SIGGRAPH Asia 38*, 6 (Dec. 2019). doi:10.1145/3355089.3356510. 2

[LLCL19] LIU S., LI T., CHEN W., LI H.: Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *ICCV* (2019), pp. 7708–7717. URL: https://github.com/. 2

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *CVPR* (2019). 1, 2, 3

[SBC21] SENGUPTA A., BUDVYTIS I., CIPOLLA R.: Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild. In *ICCV* (2021). URL: https://github.com/akashsengupta1997/HierarchicalProbabilistic3DHuman. 1, 2, 3

[XZT19] XU Y., ZHU S. C., TUNG T.: DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. *ICCV 2019-October* (9 2019), 7759–7769. URL: https://arxiv.org/abs/1910.00116v2. 2