Joint 3D Human Mesh and Texture Estimation from A Single RGB Image

Yutian Lei
Carnegie Mellon University
yutianle@andrew.cmu.edu

Luyuan Wang
Carnegie Mellon University
luyuanw@andrew.cmu.edu

Abstract

Three-dimensional human mesh reconstruction is widely used in a variety of industries. Previous works mainly focus on estimating the human geometry itself, thus only predicting a mesh without texture information. Recent works are able to estimate the textures or clothes as well, but they either require input images from different viewing angles or the ground truth textures. We propose an end-to-end model that jointly estimates both the 3D human mesh and its corresponding texture, without the ground truth texture as supervision. Specifically, our model utilizes a differentiable renderer so that the model can learn to regress the input image itself. Our preliminary results show that the algorithm can output reasonable textures and has the potential to generate more photorealistic rendering results.

1. Introduction

Image-based 3D human mesh reconstruction has broad applications, including telepresence [4], gaming, virtual dressing [15], and VR/AR [2]. For example, users can build their 3D virtual avatars in the Metaverse by simply giving a photo taken by their web cameras or mobile devices. In this paper, we propose a novel method that can directly and jointly reconstruct 3D human bodies and their corresponding textures from a single RGB input image in an end-to-end manner, without ground truth textures as supervision.

Recently, as sophisticated and powerful models are proposed to represent and capture the whole human body [10, 13], the deep learning algorithms to estimate 3D human mesh directly from RGB images are well developed for a single person or even multiple people. However, the predicted human meshes are generally "naked" without corresponding textures to recover and render a realistically looking and articulated human avatar.

Some recent works try to generate textures from RGB image inputs. But many of them need high-quality ground-truth textures as supervision [3,9,20], which need to be purchased from commercial 3D datasets like AXYZ, RenderPeople or twindom and are generally cost $20 to $100 each. Although other optimized methods do not require ground truth textures, they are either slow for inference [5] or require predefined garment templates [11]. Moreover, none of the previous works can jointly generate the human body mesh with the corresponding texture in an end-to-end fashion —— they either use the provided body mesh or predict the body mesh and texture separately.

To tackle the above-mentioned problems, we need a model that can directly estimate both the 3D human mesh and its corresponding texture from a single RGB image, without the help of any ground truth textures. Our motivation is that the 3D human mesh estimation and texture prediction can benefit from joint training and produce more photorealistic rendering results. Using PyTorch3D [16], our entire rendering pipeline can be fully differentiable, such that we can use the input image itself as supervision and get rid of the ground truth textures.

We summarize our main contributions as follows:

- We propose an end-to-end model that can estimate the textures of human meshes and output colorful and photorealistic rendering results.
- The texture and the 3D human mesh are estimated jointly, such that the performances can benefit from each other.
- The texture prediction is self-supervised from the input image, and there is no need to purchase expensive texture data from commercial datasets.

2. Related Works

2.1. 3D Human Body Representation

One of the state-of-the-art 3D human models is Skinned Multi-Person Linear Model (SMPL) [10], which parameterizes the mesh by pose parameters $\theta \in \mathbb{R}^{3K}$ and shape parameters $\beta \in \mathbb{R}^{10}$. The pose parameters θ represent how individuals vary in height, weight, and body proportions, while the shape parameters β are defined by a standard skeletal rig and describe how the 3D surface deforms.

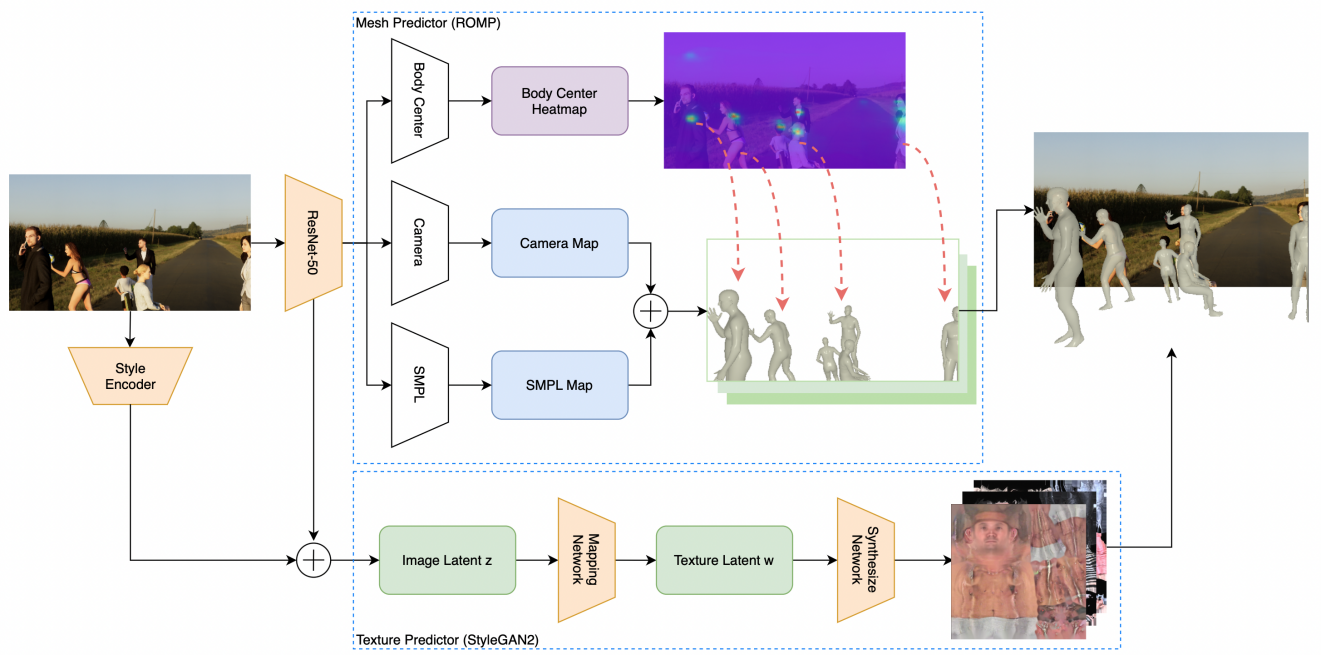

Figure 1. The overview of proposed model

SMPL is a fully differentiable function and the output is the body mesh $M \in \mathbb{R}^{N\times 3}$, with $N = 6890$ vertices.

SMPL model has good expressivity for human bodies, but it lacks the details of hands and faces. The follow-up work, SMPL-X [13], extends SMPL with fully articulated hands and an expressive face. SMPL-X is based on a template mesh that has $N = 10475$ vertices, and a kinematic tree with 54 joints. This model is controlled by 162 pose parameters θ which correspond to the 3D rotation of each model joint, shape parameters β, and the expression parameters ψ, which correspond to 10 coefficients of a PCA shape space and expression space respectively. Although SMPL-X is a more expressive representation, we are not focusing on the details of human faces and hands. Hence, we still adopt SMPL to model the human body in this research and leave using SMPL-X as future work.

2.2. 3D Human Mesh Prediction

Simplify [1] provides a framework to automatically estimate the 3D pose of the human body as well as its 3D shape from a single unconstrained image. It's a multi-stage approach, which first estimates 2D joint locations using a CNN-based method, DeepCut [14]. Then, it estimates the 3D model parameters of SMPL from these joint locations by optimizing the distances between the projected joints of the 3D model and the 2D joints. However, such a stepwise network architecture is not optimal, and it also has the disadvantage of its eventual regressor does not get to exploit the original image pixels thus the errors made by the proxy task cannot be overcome.

A recent approach that addressed the above-mentioned limitations is ROMP (Monocular, One-stage, Regression of Multiple 3D People) [17]. ROMP is a one-stage network for regressing multiple 3D people in a per-pixel prediction fashion. It directly estimates multiple differentiable maps from the whole image, from which we can easily parse out the 3D meshes of all people. Since ROMP achieves superior performance on challenging benchmarks, including 3DPW [19] and CMU Panoptic [7], as well as providing an open-source and real-time (runs over 30 FPS on an Nvidia 1070TI graphic card), we exploit ROMP as one of our backbone networks.

2.3. Textured Human Mesh Prediction

Both Simplify and ROMP can only recover the geometric shapes of the human body from a single input image, but the texture is also important to building photo-realistic 3D human models. Chaudhuri *et al.* proposed a Region-adaptive Adversarial Variational AutoEncoder (ReAVAE) [3] that learns the probability distribution of the style of each region individually so that the style of the generated texture can be controlled by sampling from the region-specific distributions. This model can generate high-resolution texture maps in a semi-supervised setup, but the ground truth texture map is still required.

StylePeople [5], which is built on top of the StyleGAN2 and Neural dressing presents a generative model able to sample random neural textures as well as optimize latent

code in one-shot or few-shot mode. The StyleGAN2 network is used to generate a multi-channel neural texture map. After the neural texture is imposed on the SMPL-X human model, the Neural dressing network will be used to render images. In this work, we follow a similar approach of using StyleGAN2 as a generator to produce the human texture. The latent vector from the feature extractor of ROMP is also fed into the texture generator, such that the two networks are stitched together.

3. Method

3.1. Overview

The overall framework of our proposed model is illustrated in Figure 1. The model consists of two predictors: the Mesh Predictor built on the SOTA multi-person 3D mesh reconstruction network ROMP [17] to predict the 3D topology of each person in the input images, and the Texture Predictor built on StyleGAN2 [8] to predict the texture information of the persons in the input images. Finally, the differentiable renderer will take the full texture as well as the generated body mesh as input to render an output image.

3.2. 3D Human Mesh Predictor

The mesh predictor adopts a simple multi-head design with a backbone and three head networks: Body Center Header, Camera Header, and SMPL header. The network takes a single RGB image as input, and outputs a Body Center heatmap $C_m \in \mathbb{R}^{1\times H\times W}$, Camera map $A_m \in \mathbb{R}^{3\times H\times W}$, and SMPL map $P_m \in \mathbb{R}^{142\times H\times W}$ with three corresponding headers, describing the detailed information of the estimated 3D human meshes. The Body Center heatmap predicts the probability of each position being a human body center. Each body center is represented as a Gaussian distribution in the heatmap, while the Gaussian kernel size k of each person center is computed given the diagonal length of the bounding box and the width of the heatmap to better incorporate the scale information of the body. Specifically, k is derived as

$$k = k_l + (\frac{d_{bb}}{\sqrt{2}W})^2 k_r \tag{1}$$

where $k_l = 2$ is the minimum kernel size and $k_r = 5$ is the variation range of k.

In the Camera map, the camera parameters of the person that takes the position as the center are predicted pixel-wise. The camera parameters $A_m \in \mathbb{R}^{3\times H\times W}$ contains 2D scale s and translation $t = (tx, ty)$ of each person in the image.

In the SMPL map, the SMPL parameters of the person that takes the position as the center are predicted pixel-wise. The SMPL parameters $P_m \in \mathbb{R}^{142\times H\times W}$ contains the 142-dim SMPL parameters, which describe the 3D pose and shape of the body mesh.

After predicting the Body Center heatmap, Camera map and SMPL map from three corresponding headers, we need to sample the camera and SMPL parameter results from the Camera map and SMPL map according to the Body Center heatmap. Specifically, the local maxima of the Body Center heatmap are selected using Non Maximum Suppression (NMS). And then the confidence scores at each local maxima center are ranked and the top N points are taken as the final centers. During training, each estimated center are matched with the nearest ground truth body center according to the L2 distance. Finally, the SMPL and camera parameters are sampled from corresponding maps into the SMPL model using the selected centers to generate the 3D body meshes. [17]

3.3. Texture Predictor

The texture predictor is derived from StyleGAN2 [8]. To generate the $\mathbb{R}^{512}$ latent code for StyleGAN2 input, we train a ResNet-18 as a style encoder to extract the latent style information from the input images. And then the encoded features are concatenated to the features extracted from the Mesh Predictor backbone to integrate the 3D structure information of the person. Then following StyleGAN2 [8], we train a multi-layer perceptron as mapping network $M(z) : \mathbb{R}^{512} \rightarrow \mathbb{R}^{512}$ to map the latent code $z \in \mathbb{R}^{512}$ to the style code $w \in \mathbb{R}^{512}$. The Synthesize Network takes a set of 512-dimensional style vectors as input controlling generation at different resolutions via modulation-demodulation mechanism as well as the set N of noise maps at these resolutions. Finally, the differentiable renderer will take the full texture as well as the generated body mesh as input to render an output image.

3.4. Loss

To jointly supervise the mesh and texture predictor, we develop individual loss functions for two predictors.

3.4.1 3D Mesh Loss

Body Center loss The body center loss encourages a high confidence value at the body center c of the Body Center heatmap and low confidence elsewhere, which is defined as the L2 difference between the predicted and ground truth center heatmap.

$$\mathcal{L}_C = ||C_m - \hat{C_m}||_2 \tag{2}$$

Camera and SMPL Parameter Loss During the training, each ground truth body is matched with a predicted parameter result for supervision as mentioned before. And for each human mesh predictor, the loss is derived as

$$\mathcal{L}_m = \lambda_{SMPL}\mathcal{L}_{SMPL} + \lambda_{cam}\mathcal{L}_{cam} + \lambda_{pj2d}\mathcal{L}_{pj2d} \tag{3}$$

where $\mathcal{L}_{SMPL}$ is the L2 loss of the SMPL parameters, $\mathcal{L}_{cam}$ is the L2 loss of the camera parameters, and L_{pj2d}

is the L2 loss of the projected 2D joints to incorporate the SMPL and camera information. The total Mesh Parameter loss is defined as the sum of the mesh loss each person in the image, which is

$$\mathcal{L}_M = \sum_{n=1}^{N} \mathcal{L}_{m,n} \tag{4}$$

where n is the total number of persons in the image.

3.4.2 Texture Loss

Perceptual Loss The perceptual loss will be calculated between the rendered images and the ground truth images. Specifically, the middle features of a pre-trained VGG will be extracted to compare the high-level similarity of the generated and ground truth images, which is

$$\mathcal{L}_{per} = \sum_{m=1}^{M} ||VGG_l(x) - VGG_l(G(x))||_1 \tag{5}$$

where m is the number of hidden features, x is the input image, and $G(x)$ is the rendered output image.

Reconstruction Loss The pixel-wise reconstruction loss is also adopted to improve the photo-realism of the rendering results. Specifically, we compare the intensity domain and gradient domain between the generated and ground truth images by

$$\mathcal{L}_{rec} = ||x - G(x)||_1 + ||\partial x - \partial G(x)||_1 \tag{6}$$

KLD loss The Kullback–Leibler divergence loss is used to approximate the learned latent texture distribution to a standard normal distribution $\mathcal{N}(0, I)$ and is formulated as

$$\mathcal{L}_{KL} = KLD(v, \mathcal{N}(0, I)) \tag{7}$$

Finally our total loss will be

$$\mathcal{L} = \mathcal{L}_{per} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_C\mathcal{L}_C + \lambda_M\mathcal{L}_M \tag{8}$$

4. Experiments

Due to the limitation of the computational resources and time, we can only train our model to generate the texture images of 64×64 on a single GeForce RTX 2080 GPU. The loss weights are set to balance each loss to a similar scale. And we train our model with an ADAM optimizer with a batch size of 16 and a learning rate of 0.0001 for 40000 iterations. The learning rate decays in the 20000 and 25000 iterations by a scale of 10.

4.1. Dataset

We use Avatars in Geography Optimized for Regression Analysis (AGORA) [12] as our dataset, which is a synthetic human dataset with high realism and accurate ground truth. It consists of around 14K training and 3K test images by rendering between 5 and 15 people per image using either image-based lighting or rendered 3D environments, taking care to make the images physically plausible and photoreal. In total, AGORA contains 173K individual person crops. AGORA provides (1) SMPL / SMPL-X parameters and (2) segmentation masks for each subject in images. To make the problem simpler, we mask out the background with the ground truth segmentation mask, so that each of our training images only contains one person within a white background. Figure 2 shows some samples from the dataset.

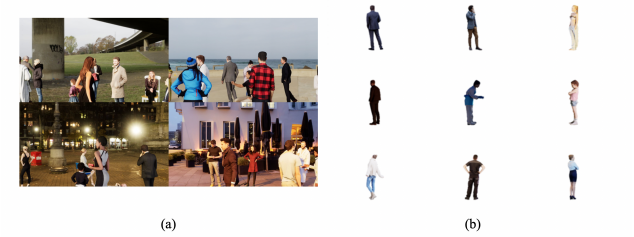

Figure 2. Samples from the AGORA dataset. (a) the original AGORA dataset. (b) the segmented dataset, each input image only contains one person.

4.2. Qualitative Results

From Figure 3 we can see that the predicted textures are able to preserve the general color of the input image and the new synthetic views are reasonable, although the rendered images are blurry and lack details. We summarize three potential reasons. First, due to the limitation of our computation resources, we are only able to train the network to generate the texture images of 64×64. If we can enlarge the output image size, the texture can have more details. Second, we only use the StyleGAN2 as a generator and didn't apply any adversarial loss. Thirdly, we can also pre-train the GAN on an SMPL texture dataset, so that the model can learn how to generate realistic textures. To further address this hypothesis, we trained the GAN on SURREAL [18] dataset, which contains synthetic human textures. Some examples of the generated textures are shown in Figure 4. We can see that these textures have much more detailed information, thus we believe fine-tuning this model to solve the human texture prediction task can give us a better result.

4.3. Quantitative Results

To further evaluate the performance of our model quantitatively, we numerically compare our predicted results and

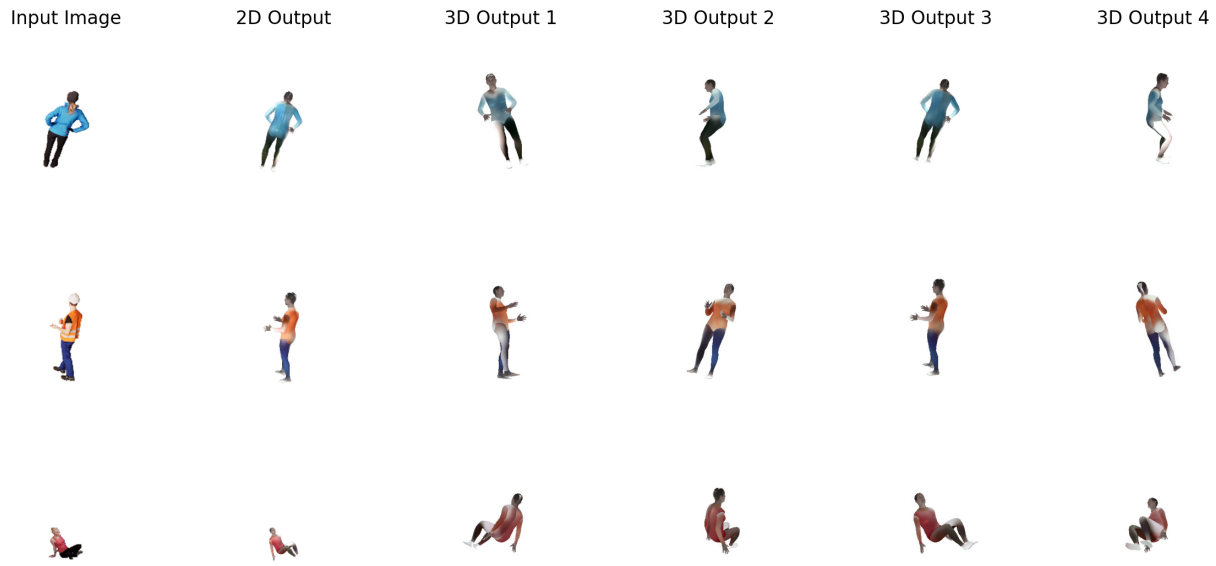

Figure 3. Examples of the prediction results. From left to right: the single RGB input image; the 2D output of the model, where the camera poses are estimated to let the rendering result best fit the input image; four snapshots of the textured mesh from different viewing angles, respectively.

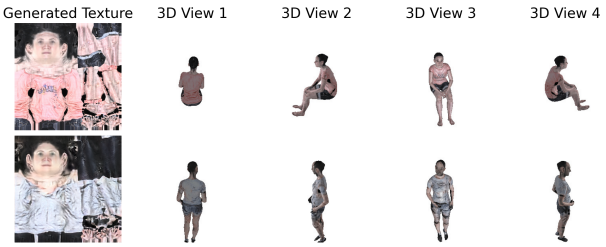

Figure 4. Textures generated by StyleGAN2. From left to right: the texture generated by StyleGAN2; four snapshots of the textured mesh from different viewing points.

ground truth using three metrics: SSIM, PSNR and FID.

SSIM Structural Similarity Index Measure (SSIM) is a widely employed metric to evaluate image processing algorithms currently, which not only considers image degradation as a perceived change in structural information but also incorporates important perceptual phenomena, including both luminance masking and contrast masking terms.

PSNR Peak signal-to-noise ratio (PSNR) computes the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation and thus is commonly used to quantify reconstruction quality for images.

SSIM	PSNR	FID
0.974	26.134	36.144

Table 1. The Numerical Results between our predicted results and ground truth

FID Fr´echet Inception Distance (FID) [6] which computes the Fr´echet distance between the distribution of the feature space for the synthetic images and real-world images, is also adopted to evaluate the realism of our image prediction result.

We show the quantitative results in Table 1. As we can see, the SSIM is pretty high, which results from the background removal and similar distribution of the generated and ground truth images. The value of PSNR and FID is in a reasonable range, which confirms the qualitative results we show above.

5. Conclusion

In this paper, we present our deep learning model that can jointly estimate the human mesh and the texture. One of the key contributions is that our training process does not require ground truth texture as supervision, so it's a lot easier to collect the training data, without purchasing the expensive commercial datasets. Compare with previous works of predicting human mesh itself, our rendering

results are more photorealistic. Although our preliminary results demonstrate our model is able to produce reasonable textures, which indicates the feasibility of the proposed method, there is still room for improvement. In future work, we can add face regularization loss to better generate facial textures. As the generated texture is blurry, we can add adversarial loss to the texture and mesh generator for better results. We also realize the texture generator is hard to train, thus one of the potential solutions is that we can pretrain the StyleGAN2 on an SMPL texture dataset, so that the model can learn the general distribution of human textures, and then fine-tune it for prediction based on the input image.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2

[2] Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, et al. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE transactions on visualization and computer graphics*, 24(11):2993–3004, 2018. 1

[3] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised synthesis of high-resolution editable textures for 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7991–8000, 2021. 1, 2

[4] Daniel Gotsch, Xujing Zhang, Timothy Merritt, and Roel Vertegaal. Telehuman2: A cylindrical light field teleconferencing system for life-size 3d human telepresence. In *CHI*, volume 18, page 552, 2018. 1

[5] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 1, 2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[7] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[8] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 3

[9] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 1

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1

[11] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020. 1

[12] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[13] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2

[14] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[15] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 1

[16] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[17] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 2, 3

[18] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 4

[19] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2

[20] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4491–4500, 2019. 1