HUMAN CARVING: A PARSING-BASED FRAMEWORK FOR 3D HUMAN RECONSTRUCTION

Baoxing Li, Xu Zhao

Department of Automation, Shanghai Jiao Tong University
Key Laboratory of System Control and Information Processing, Ministry of Education of China

ABSTRACT

Human-centric computer vision tasks often benefit from each other. In this paper, we propose a novel framework called *Human Carving* to explore the relationships between human parsing and multi-view 3D human reconstruction, which is the first method to consider the two related tasks. It consists of three modules: 1) *Pose-aware Multi-view Human Parsing*, 2) *Semantic Visual Hull Carving* and 3) *Hierarchical Human Model Fitting*. Taking the sparse multi-view images as input, the framework automatically generates a *Part-Aware Visual Hull* (PAVH) of human body parts and then estimates the human shape and pose simultaneously. Experimental results on real scenes demonstrate the effectiveness of our framework.

Index Terms— Human Parsing, 3D Human Reconstruction, Multi-view Geometry, Visual Hull, SMPL Model

1. INTRODUCTION

Imaged-based *Multi-view 3D Human Reconstruction* is a challenging task, which requires estimating the human shape and pose simultaneously. Most of the existing methods are based on 2D human pose and silhouette information [1, 2, 3], while ignoring the semantic and detailed edge information of human body parts. In this paper, we explore the function of human parsing in the reconstruction task and propose the first parsing-based method to estimate 3D human shape and pose.

Human parsing, also known as *human body part segmentation*, is fundamental to many human-centric computer vision tasks, the goal of which is to segment the pixels of different human body parts given an RGB image. Recently, human parsing has been promoted by efficient CNN-based semantic segmentation approaches[4, 5]. Auxiliary tasks including edge detection [6, 7] and human pose estimation [8, 9] have improved the accuracy and generalizability of human parsing.

The relationships between human parsing and multi-view 3D human reconstruction can be summarized as follows: **(1) Silhouette Relationship**. The pixels of 2D human parsing result and the voxels of 3D human reconstruction result share the same silhouette information in the specific viewpoint. And such relationship between pixels and voxels can be bridged by multi-view geometry methods such as visual hull [10]. **(2) Semantic Relationship**. Human parsing provides a part-specific correspondence between image pixels and human surface [11], which can be captured by parametric human models, e.g. the Skinned Multi-Person Linear (SMPL) model [12]. The SMPL model is a deformable template mesh under the control of shape and pose parameters. And each vertice of the mesh corresponds to a specific body part. Based on such correspondence, the semantic relationship between human parsing and 3D human body can be built consequentially. **(3) Structure Relationship**. The result of human parsing also provides the rough locations of body joints because of the prior of the kinematic tree [13] of human body. And this tree structure is also the basic mechanism of the SMPL model. Utilizing the silhouette, semantic and structure information of human parsing, we estimate the shape and pose parameters of the SMPL model, reconstructing the naked human body.

Inspired by the three relationships, we propose a novel framework called *Human Carving* to bridge the gap between human parsing and multi-view 3D human reconstruction, the name of which is borrowed from the space carving theory [14]. As shown in Fig. 1, our framework is composed of three modules: 1) *Pose-aware Multi-view Human Parsing*. It provides the human body part segmentation of each view. 2) *Semantic Visual Hull Carving*. A part-aware visual hull is carved using human parsing results. 3) *Hierarchical Human Model Fitting*. The SMPL model is aligned with the part-aware visual hull as output. Our method is evaluated on a real scene multi-view human dataset [15]. The quantitative and qualitative results demonstrate the effectiveness of our method compared to the previous methods.

To summarize, our contributions are three fold. (1) We propose to utilize human parsing in multi-view 3D human reconstruction, which is novel to solve this problem. (2) We extend the concept of visual hull to a part-aware semantic model, enhancing its relevance to the parametric human model, e.g. SMPL. (3) We design a four-stage optimization approach to estimate the shape and pose parameters of the SMPL model hierarchically and efficiently.

This work has been supported in part by the funding from NSFC (61673269, 61273285) and the project funding of the Institute of Medical Robotics at Shanghai Jiao Tong University. (Corresponding author: Xu Zhao. E-mail: zhaoxu@sjtu.edu.cn)

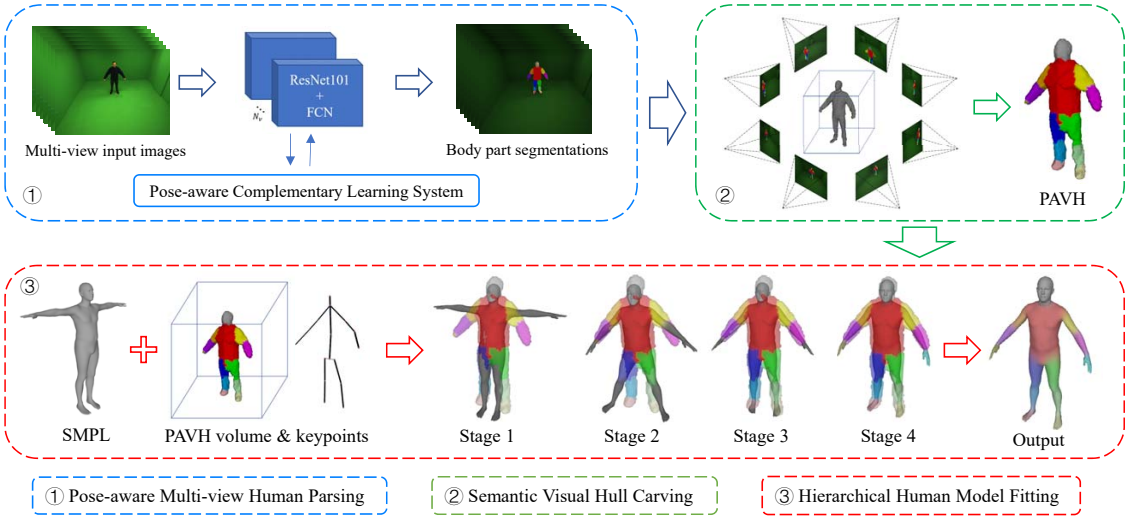


Fig. 1. The proposed framework. Firstly, we obtain the body part segmentations of the multi-view input images using a multi-view human parsing network trained by a pose-aware complementary learning system. Secondly, the body part segmentations are fed into the semantic visual hull carving module, producing a Part-Aware Visual Hull (PAVH). Thirdly, the parametric human body model, SMPL, is used to fit the constructed PAVH in a hierarchical way.

2. OUR FRAMEWORK

2.1. Pose-aware Multi-view Human Parsing

Following the idea of [16], we build a multi-view human parsing neural network to segment the human body part for multi-view input images, which is trained under the supervision of both human 2D pose and body part segmentation.

1) **Multi-view Human Parsing Network**. For the input images $I_i \in \mathbb{R}^{H\times W\times 3}, i = 1,\cdots,N_v$ of N_v views, the ResNet101 [17] backbone with pyramid connections [18], denoted as r, is firstly used to extract the feature maps $F_i = r(I_i)$. Then a fully convolutional network, denoted as α, is used to compute the body part scores $B_i = \alpha(F_i), B_i \in \mathbb{R}^{(N_p+1)\times H\times W}$, where $N_p + 1 = 15$ is the number of predicted labels, following the order of {background, head, torso, left&right upper arms, left&right fore arms, left&right hands, left&right thighs, left&right shanks, left&right feet} from 0 to 14. The argmax values of B_i derive the body part segmentation results $S_i \in \mathbb{R}^{H\times W}$. And the above ResNet101+FCN structure is stacked for N_v times to parsing the images of N_v views in parallel.

2) **Pose-aware Complementary Learning System**. The 2D human pose estimation is adopted as an auxiliary task to train the human parsing network using both synthetic data with part labels and real data with pose labels, which makes the trained model learn human body structure priors.

2.2. Semantic Visual Hull Carving

Using the results of human parsing, we construct a *Visual Hull*-based semantic model of human body, which is called the *Part-Aware Visual Hull* (PAVH). The construction of PAVH be divided into three steps:

1) **Carving Space Initialization**. Firstly, a triangulation-based method is used to determine the 3D boundary of carving space based on the 2D bounding boxes derived from the body part segmentations $S_i, i = 1,\cdots,N_v$ and the calibrated camera metrixes M_i. Given a pre-defined voxel size, the initial voxel grid V^{init} can be obtained, where $\Phi(v_k) = 0$ denotes the initial value for each $v_k \in V^{init}$.

2) **Voting-based Volumetric Data Fusion**. Secondly, the body part segmentations S_i and scores B_i of each view are projected into the voxel grid V^{init}. For each voxel $v_k \in V^{init}$, the projected values $s_{k,i} \in S_i$ and $b_{k,i} \in B_i$, where $i = 1,\cdots,N_v$, are fused by a *weighted voting function* to determine the corresponding 3D body part label:

$$\Phi(v_k) = f(s_{k,1},\cdots,s_{k,N_v}; b_{k,1},\cdots,b_{k,N_v}) \tag{1}$$

where $s_{k,i}$ are the 2D part labels, as votes, and $b_{k,i}$ are the corresponding predicted probabilities, as weights. The voting function f is used to reduce the influence of self-occlusion by fusing the probabilities of other views, shown in equation (2).

$$\begin{aligned} f(s_k; b_k) &= argmax(w_k),\ w_k \in \mathbb{R}^{N_p} \\ w_{k,j} &= \sum_{s_{k,i}=j} b_{k,i},\ j = 1,\cdots,N_p \end{aligned} \tag{2}$$

where $w_{k,j} \in w_k$ is the sum of weighted votes for the j-th body part from all of the N_v views.

3) **Non-zero Voxels Combination**. Finally, the non-zero valued voxels are selected to compose the PAVH, denoted as V^{part}. As shown in Fig. 1, PAVH contains both human silhouette and semantic body parts information.

2.3. Hierarchical Human Model Fitting

The goal of this step is to fit the SMPL model to the constructed PAVH by minimizing an objective function.

1) **The SMPL Model**. The SMPL (Skinned Multi-Person Linear) model [12] is a linear human mesh model. It uses an artist-created human body mesh with 6890 vertices as the template $\bar{T}$, which is then deformed under the control of shape parameters, $\beta \in \mathbb{R}^{10}$, pose parameters, $\theta \in \mathbb{R}^{24\times 3}$ and global translation parameters, $t \in \mathbb{R}^3$ in an additive way, formulating the model $M(\beta, \theta, t)$:

$$M(\beta,\theta,t) = W((\bar{T} + B_S(\beta) + B_P(\theta)), J(\beta), \theta, t, \mathcal{W}) \quad (3)$$

where $W(\cdot)$ is a linear blend skinning function and other details refer to [12]. We divide the vertices of the SMPL mesh into 14 parts, denoted as $V_j^{SMPL}, j = 1, \cdots, 14$, corresponding to the 14 body parts defined in section 2.1, to make it easier to be fitted to the PAVH data.

The human model fitting problem is thus equivalent to the estimation of β, θ and t, solved by minimizing an objective function. To form the objective function, we extract 14 keypoints and construct 14 volumes from the PAVH data.

2) **PAVH keypoints Extraction**. For each voxel v_k in V^{part}, if the value of its adjacent voxels $v_{k,adj}$ is not the same as v_k, it is marked as a key voxel. And the corresponding key value is denoted as an unordered tuple, $(v_k, v_{k,adj})$. The average position of each set of key voxels is taken as the location of *PAVH keypoints*, $J_i, i = 1, \cdots, 14$, as shown in table 1.

Table 1. Relationships between J_i and body parts.

key values	keypoints	key values	keypoints
(1,2)	J_1: neck	(2,9,10)	J_2: pelvis
(2,3)	J_3: L shoulder	(2,4)	J_4: R shoulder
(3,5)	J_5: L elbow	(4,6)	J_6: R elbow
(5,7)	J_7: L wrist	(6,8)	J_8: R wrist
(2,9)	J_9: L hip	(2,10)	J_{10}: R hip
(9,11)	J_{11}: L knee	(10,12)	J_{12}: R knee
(11,13)	J_{13}: L ankle	(12,14)	J_{14}: R ankle

3) **PAVH Volume Construction**. Inspired by the TSDF (Truncated Signed Distance Field) volume used in [19], we construct the *PAVH volumes*, $V_j^{PAVH}, j = 1, \cdots, 14$, for 14 body parts, to represent the distance between any voxel outside the j-th body part and the closest inside voxel. For each voxel $v_{j,k} \in V_j^{PAVH}$ and $v_i \in V^{part}$, the value of $v_{j,k}$ is determined by equation (4).

$$\Phi_j(v_{j,k}) = \begin{cases} \min\limits_{\Phi(v_i)=j} \{d(v_{j,k}, v_i)\}, & if\ v_{j,k} \notin V^{part} \\ 0, & otherwise \end{cases} \quad (4)$$

where $d(\cdot)$ calculates the euclidean distance between two voxels. The PAVH volume is visualized in Fig. 2.

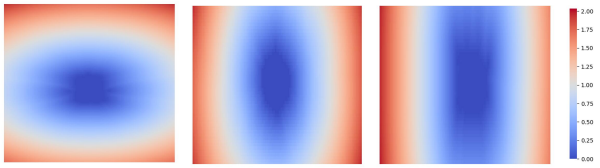


Fig. 2. 2D visualization of the 3D PAVH volume, V_j^{PAVH} where $j = 2$, along X,Y and Z coordinates.

4) **Objective Function**. The objective function is the sum of four error terms, including two data terms of PAVH keypoint and volume, two prior terms of SMPL shape and pose:

$$\begin{aligned} E(\beta,\theta,t) = E_{keypoint} + \lambda_1 E_{volume} \\ + \lambda_2 E_{shapePrior} + \lambda_3 E_{posePrior} \end{aligned} \quad (5)$$

PAVH Keypoint Term: We select 14 out of 24 joints of the SMPL model to fit the corresponding 14 PAVH keypoints as shown in table 1.

$$E_{joint}(\beta,\theta,t) = \sum_{i=1}^{14} ||J_i - R_\theta(J(\beta)_i)||_2 \quad (6)$$

where $R_\theta(J(\beta)_i)$ is the position of the i-th selected SMPL joint, obtained by global rigid transformation.

PAVH Volume Term: Our proposed PAVH constrains each body part in a limited space. And the constraint power is represented in PAVH volumes. So the pre-computed PAVH volumes are used here to punish the misplaced SMPL vertices of each body part.

$$E_{volume}(\beta,\theta,t) = \sum_{j=1}^{14} \sum_{i=1}^{N_j} \Phi_j(v_{j,i}) \quad (7)$$

where $v_{j,i} \in V_j^{SMPL}$ representing the i-th vertice of j-th body part of the SMPL mesh, and $\Sigma_{j=1}^{14} N_j = 6890$.

SMPL Shape and Pose Prior Terms: These two terms are the same as the prior terms used in SMPLify [20].

5) **Four-Stage Hierarchical Optimization**. Following the rule of kinematic chain of human body, we minimize the objective function using the SLSQP optimizer in four stages, hierarchically optimizing the shape and pose of the SMPL model. The design of each stage is shown below:

Stage 1: Global rotation and transformation.

Stage 2: Optimization of the pose of upper limbs, basic body shape, and global transformation.

Stage 3: Optimization of the pose of upper and lower limbs, basic body shape, and global transformation.

Stage 4: Refined optimization with consideration of all shape, pose and transformation parameters.

The kinematic tree of human body is considered in the four-stage optimization method, which is proved to be effective in our experiments.

3. EXPERIMENTS

3.1. Experimental settings

In our experiments, the multi-view human parsing network is based on the pre-trained CDCL model [16]. We make a trade-off between accuracy and efficency by setting the voxel size of the visual hull to $0.02m$ to construct a coarse initialization. As for model fitting step, the number of data terms and variables used in each stage of the hierarchical optimization process is shown in table 2. And the super parameters of the objective function Eq.5 follows the setting of [20]. Our experiment is carried out on an indoor human dataset [15], the images of which are captured by an 8-view stereo vision system. Totally 850 frames RGB images with groundtruth mesh are used to evaluate our framework.

Table 2. Data terms and variables used in each stage.

Stages	Keypoints	Volumes	β	θ	t
Stage 1	6	2	1	1×3	3
Stage 2	10	6	2	5×3	3
Stage 3	14	10	3	9×3	3
Stage 4	14	14	10	24×3	3

3.2. Quantitative Analysis

Metric. To quantitatively evaluate our reconstruction result, we calculate two per-vertex errors, SMPL-GT and SMPL-VH, using equation (8):

$$e_{verts} = \frac{1}{N_i} \sum_{i}^{N_i} \min_{1<j<N_j} ||v_{smpl,i} - v_{ref,j}||_2 \qquad (8)$$

where $v_{smpl,i}$ denotes the i-th of the N_i vertices of the fitted SMPL mesh, and $v_{ref,j}$ denotes the j-th of the N_j vertices of the other reference mesh, which is the groundtruth (GT) mesh provided by the dataset [15] and PAVH mesh obtained by marching cube algorithm [21].

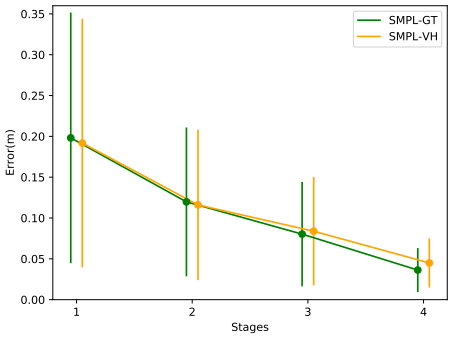
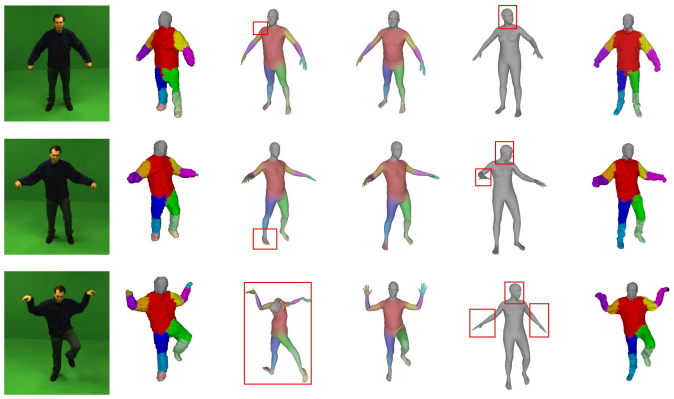


Fig. 3. Means and standard deviations of per-vertex errors.



(a) view 1. (b) PAVH. (c) 2-stage. (d) 4-stage. (e) [22]. (f) GT.

Fig. 4. Qualitative comparison. (a) Input image of one view (cropped). (b) the Part-Aware Visual Hull (PAVH). (c) the fitted SMPL model of two-stage method (ablation study). (d) the fitted SMPL model of four-stage method (final output). (e) the result of a multi-view SMPL fitting method based on smplify-x [22]. (f) the ground truth mesh colored by the PAVH volumes.

As shown in Fig. 3, the means and standard deviations of SMPL-GT and SMPL-VH reduce to a reasonable range in stage 4, illuminating the success of the SMPL fitting process.

3.3. Qualitative Comparison

To test the effectiveness of our hierarchical optimization method, we make an ablation study, which is a two-stage optimization method consisting only stage 1 and 4 in table 2, but with the same max iteration number as the four-stage method to make a fair comparison. As shown in Fig. 4 (c) and (d), the four-stage method converges to a better result especially for complex human poses. To compare our human parsing-based method to the 2D human pose-based method, we test the smplify-x [22] based multi-view SMPL fitting method on the same dataset, where the 2D human pose is estimated by [16]. As shown in Fig. 4 (d) and (e), our method outperforms [22] in both lower limbs and head orientation estimation. Because human parsing captures more detailed information than 2D human pose.

4. CONCLUSION

As far as we know, our framework is the first attempt to utilize human parsing in multi-view 3D human reconstruction. We extend the concept of visual hull to a part-aware semantic model, which provides a reliable initialization of 3D human reconstruction. A hierarchical optimization method is proposed to estimate the shape and pose of the SMPL model. Without any 3D training data, our framework reserves good generalizability in real indoor scenes.

5. REFERENCES

[1] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black, "Towards accurate marker-less human shape and pose estimation over time," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 421–430.

[2] Martin Oswald and Daniel Cremers, "A convex relaxation approach to space time multi-view 3d reconstruction," in *ICCV*, 2013, pp. 291–298.

[3] Armin Mustafa, Chris Russell, and Adrian Hilton, "U4d: Unsupervised 4d dynamic scene understanding," in *ICCV*, 2019, pp. 10423–10432.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[6] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018, pp. 770–785.

[7] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *AAAI*, 2019, vol. 33, pp. 4814–4821.

[8] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[9] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan, "Human pose estimation with parsing induced learner," in *CVPR*, 2018, pp. 2100–2108.

[10] Aldo Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018, pp. 7297–7306.

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[13] Xin Cao and Xu Zhao, "Anatomy and geometry constrained one-stage framework for 3d human pose estimation," in *ACCV*, 2020.

[14] Kiriakos N Kutulakos and Steven M Seitz, "A theory of shape by space carving," *International journal of computer vision*, vol. 38, no. 3, pp. 199–218, 2000.

[15] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović, "Articulated mesh animation from multi-view silhouettes," in *ACM SIGGRAPH 2008 papers*, pp. 1–9. 2008.

[16] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun, "Cross-domain complementary learning using pose for multi-person part segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[19] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *CVPR*, June 2018.

[20] 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 *ECCV*. Springer, 2016, pp. 561–578.

[21] William E Lorensen and Harvey E Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[22] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019, pp. 10975–10985.