LEAP: Learning Articulated Occupancy of People
– Supplementary Material –

A. Overview

In this supplementary document, we first provide details about the proposed neural network modules (Sec. B) and their training procedure (Sec. C). Then, we present more qualitative and quantitative results that were not included in the paper due to the page limit (Sec. D). We also discuss the limitations of the proposed method (Sec. E). Lastly, we summarize the notations used in the paper to improve paper readability (Sec. F).

B. Network architecture

We first present details about the proposed neural encoders that are used to create the global feature vector $z \in \mathbb{R}^{596}$ (128, 312, and 156 dimensions for structure, shape, and pose features respectively) and then detail the proposed neural network architectures.

B.1. Structure encoder

The structure encoder consists of 52 small multi-layer perceptrons that are organized in a tree structure, where each node of the tree corresponds to one joint in the human skeleton and outputs a compact feature vector $b_k \in \mathbb{R}^6$.

Each MLP node (Table B.1) takes as input a 19-dimensional feature vector – 6 dimensions for the parent feature, 9 for the rotation matrix, 1 for the bone length, and 3 for the joint location – and outputs a small bone code. These bone codes are concatenated to form one structure feature vector as explained in the main paper.

Since the root node does not have a parent node to be conditioned on its feature vector $b_0 \in \mathbb{R}^6$, we create b_0 with a single linear layer that takes as input vectorized θ pose parameters and joint locations $\mathbf{J}$.

B.2. PointNet encoder

We implement a PointNet encoder (Figure B.1) to encode a point cloud into a fix-size feature vector. This network is used in Sec. 4.1.1 to create a 128-dimensional shape feature vector ($P = 128$ in Figure B.1) and two 100-dimensional feature vectors for the inverse and the forward LBS networks ($P = 100$).

Linear(19, 19) + bias
ReLU
Linear(19, 6) + bias
ReLU

Table B.1. **The MLPs of the structure encoder**. This compact neural network architecture consists of 500 trainable parameters.

B.3. Bone projection layers

The bone projection layers $\Pi_{\omega_k} : \mathbb{R}^{596} \mapsto \mathbb{R}^{12}$ create small per-bone features z_k and are implemented as efficient grouped 1D convolutions [26].

B.4. ONet

The architecture of the occupancy network is similar to the one proposed in [42] and is illustrated in Figure B.2.

B.5. Linear blend skinning networks

The inverse and the forward LBS networks are similar and implemented as MLPs conditioned on a latent feature vector (Figure B.3).

Forward LBS network. The latent feature vector for the forward LBS network $c_{\text{fwd}} \in \mathbb{R}^{200}$ is created as a concate-

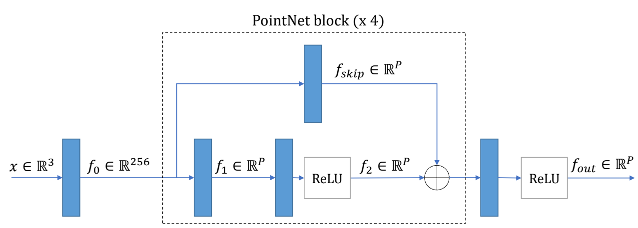

Figure B.1. **PointNet architecture.** The PointNet encoder encodes an arbitrary set of input points $\{x \in \mathbb{R}^3\}$ into a feature vector of length P. All layers except for the skip connections use bias. Blue rectangular blocks are trainable linear layers. $\oplus$ operator is the element-wise sum.

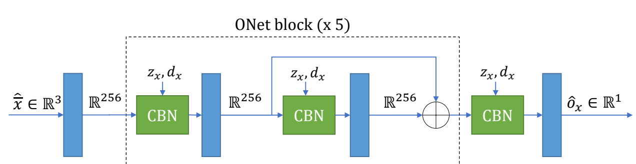

Figure B.2. **ONet architecture.** The occupancy neural network is implemented as a multi-layer perceptron that consists of five ONet blocks. Blue rectangular blocks are trainable linear layers. $\oplus$ operator is the element-wise summation. CBN blocks are Conditional Batch-Normalization blocks [14, 16, 42] followed by ReLU activation function that are conditioned on the local point feature $z_x \in \mathbb{R}^{12}$ and the cycle-distance feature d_x.

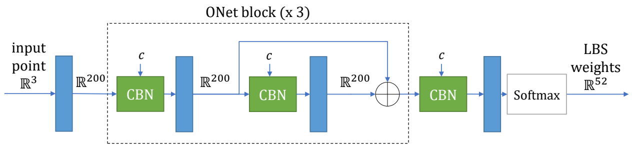

Figure B.3. **The architecture of the inverse and the forward LBS networks.** Both networks use this architecture to regress skinning weights. Blue rectangular blocks are trainable linear layers. $\oplus$ operator is the element-wise sum. ONet and CBN blocks are introduced in Figure B.2. The feature vector c is $c_{\text{inv}} \in \mathbb{R}^{280}$ for the inverse LBS network and $c_{\text{fwd}} \in \mathbb{R}^{200}$ for the forward LBS network.

nation of two 100-dimensional feature vectors created by the PointNet encoder. One is created by encoding the estimated canonical vertices $\hat{\bar{V}}$, and the other one is created by encoding the estimated posed vertices $\hat{V}$.

Inverse LBS network. The latent feature vector for the inverse LBS network $c_{\text{inv}} \in \mathbb{R}^{280}$ is created as a concatenation of the conditional features produced for the forward LBS network c_{fwd} and an additional 80-dimensional feature vector created by a single linear layer that takes as input concatenated vectorized pose parameters and joint locations.

C. Training

We provide additional details for three independent training procedures. First, we train the inverse and the forward LBS networks and then use these two modules as deterministic differentiable functions for the occupancy training.

Occupancy training. All modules except the LBS networks are trained together and have a total of about 1.5M trainable parameters. For each batch, 1536 points are sampled uniformly and 1536 near the surface (1024 in the posed and 512 in the canonical space). Points that are sampled directly in the canonical space are not propagated through the forward LBS network and are associate with pseudo ground truth skinning weights $w_{\bar{x}}$ to calculate local codes z_x.

The inverse LBS network has about 1.5M parameters. Each training batch consists of 1024 uniformly sampled points and 1024 points sampled around the mesh surface.

The forward LBS network has about 1.2M parameters. Each training batch consists of 1024 points sampled in the canonical space (512 uniformly sampled and 512 sampled near the surface) and points that are sampled for the training of the inverse LBS network. The latter set of points is mapped to the canonical space via the proposed pseudo ground truth weights.

D. Additional experiments and results

We supplement experiments for generalization (Figure D.4), for learning LBS (Sec. D.1), and placing people in scenes (Sec. D.2).

D.1. Evaluation of linear blend skinning networks

Our forward LBS network operates in the canonical space and does not need to deal with challenging human poses as the inverse LBS network. Here, we quantify the performance gap between these two networks on the unseen portion of query points for three experimental setups presented in the main paper.

As an evaluation metric, we report the l_1 distance between pseudo ground truth weights and predicted weights by the inverse l_1^{inv} and the forward l_1^{fwd} LBS networks.

Quantitative results (Table D.2) show that the forward LBS network consistently outperforms the inverse LBS network across all settings. Furthermore, the inverse LBS network performs worse when subjects are not seen during the training.

D.2. Generating people in scenes

We provide additional qualitative results (Figure D.5) for the experiment presented in Figure 5 and visualize SDF (Figure D.7) that is used to compute the *human-scene* collision score. Although the SDF is very noisy, we still use it to compute this score for a fair comparison with PLACE [72].

We further provide an experiment on a larger Replica room [61]. Similar to Sec. 6.4, we sample 50 people from PLACE [72] and select 60 human body pairs that interpenetrate. These pairs are then optimized with our method by minimizing the proposed point-based loss (20).

Quantitative results (Table D.3) demonstrate that our approach improved collision scores over the baseline [72], except for the human-scene score which is unreliable due to the aforementioned noisy SDF (Figure D.7). The qualitative results displayed in Figure D.6 show that our method successfully resolves deep interpenetrations with scene geometry which could not be straightforwardly achieved with differentiable mesh-based collision methods – for example, a modified version of the approach presented in [62]. This indicates that our volumetric error signal is more effective than the surface error signal imposed by mesh-based methods.

E. Limitations

We observed some challenging scenarios in which our learned inverse linear blend skinning network may fail to correctly map a query point to the canonical space and consequently distort occupancy in the posed space. This problem occurs when the network is not well trained and two body parts are close to each other or even self-intersect. An example of a failure case of an unseen subject is displayed in Figure E.8. Therefore, a promising future direction is to explicitly model self-contact for learning the occupancy representation.

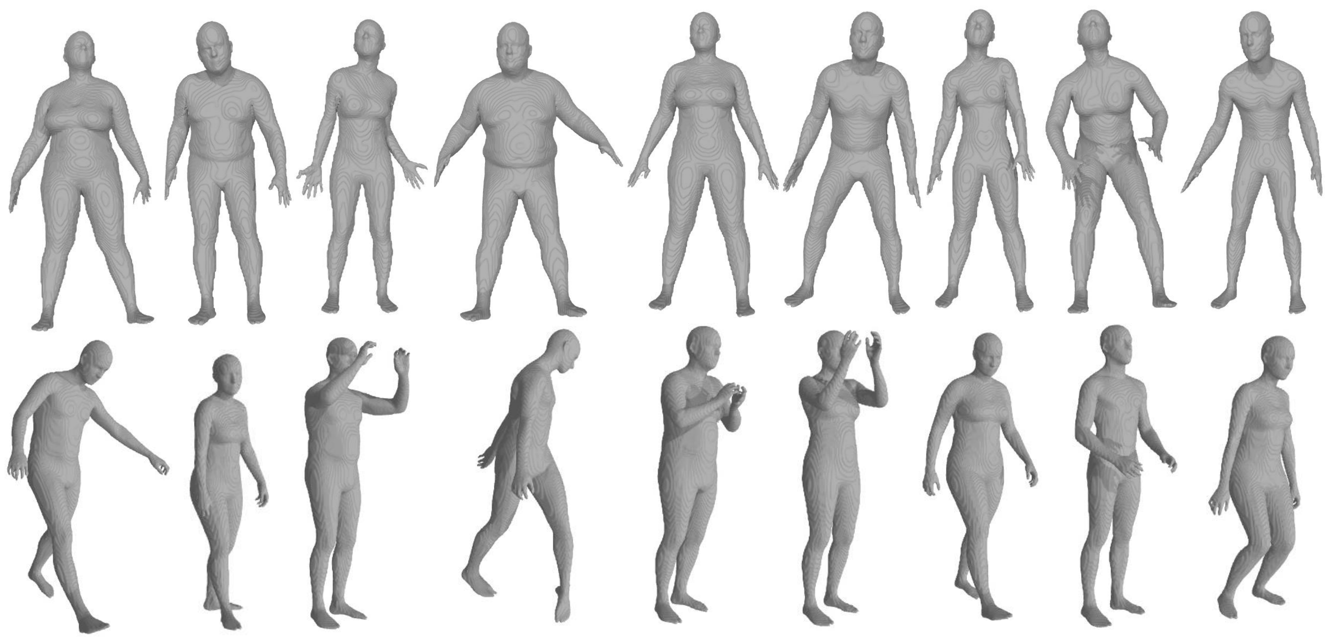

Figure D.4. **Generalization experiment.** Qualitative results for the generalization experiment (Sec. 6.3, Table 2) for DFaust [5] unseen poses (top row) and MoVi [20] unseen subjects (bottom row).

Experiment type	$l_1^{inv} \downarrow$	$l_1^{fwd} \downarrow$
Multi-person occupancy (Sec. 6.2)	0.1894	0.1252
Generalization: unseen poses (Sec. 6.3)	0.1997	0.0818
Generalization: unseen subjects (Sec. 6.3)	0.2138	0.1098

Table D.2. **Evaluation of the inverse and the forward linear blend skinning networks.** Reported l_1 distance shows that the forward LBS network consistently outperforms the inverse LBS network across all experiment settings. This is expected because the inverse LBS network reasons about different body shapes and poses, while the fwd-LBS network reasons only about body shapes since the pose in the canonical space is constant. "Multi-person" and "unseen poses" experiments are performed on the DFaust [5] dataset, while the "unseen subjects" experiment is performed on the MoVi [20] dataset. More details on the experimental setups are available in the paper (Sec. 6.2, Sec. 6.3).

F. Notation

Lastly, we summarize the key notation terms in Table F.4 for improved readability.

		Room 1 (Figure D.5)		Room 2 (Figure D.6)	
Collision score		PLACE [72]	Ours	PLACE [72]	Ours
human-scene	$\downarrow$	5.72%	5.72%	**0.34%**	1.20%
scene-human	$\downarrow$	3.51%	**0.62%**	0.98%	**0.77%**
human-human	$\downarrow$	5.73%	**1.06%**	7.64%	**1.09%**

Table D.3. **Improved PLACE [72].** Results on two Replica [61] rooms. Our proposed optimization method successfully mitigates interpenetrations between scene geometry and other humans. Note that the human-scene score is unreliable metric due to noisy scene SDF.

PLACE [72] Our optimization

Figure D.5. **Improved PLACE [72].** Additional viewpoints of a Replica room [61] for results presented in the paper (Figure 5). Our point-based loss effectively resolves collisions of the human pairs. Quantitative results are reported in Table D.3.

PLACE [72] | Our optimization

Figure D.6. **Improved PLACE [72].** Results demonstrate that our method can resolve challenging interpenetrations with scene geometry. Note that this complex penetrations with the thin mesh geometry cannot be straightforwardly fixed with mesh-based intersection methods [62] that impose a surface-based error signal. This demonstrates that our flexible volumetric point-based loss is more efficient, which is quantified by the improved collisions scores displayed in Table D.3.

Figure D.7. **Noisy SDF** that is used to compute the human-scene score for results presented in the paper (Sec. 6.4, Table 3, Figure 5).

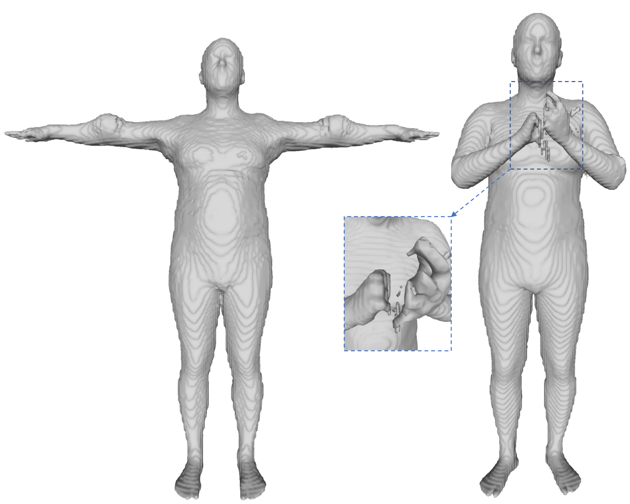

Figure E.8. **Failure case.** An example of an unseen MoVi [20] subject when two hands self-intersect. The inverse LBS network may incorrectly map a given query point to the canonical space for self-intersected regions which consequently distorts occupancy representation in the posed space.

		Input parameters
$K \in \mathbb{R}$	:	the number of input bones (52)
$x \in \mathbb{R}^3$	:	query point
$G_k \in \mathbb{R}^{4,4}$	:	bone transformation matrix of part k
		SMPL parameters
N	:	the number of vertices (6890)
$\mathcal{S}$	:	shape blendshape parameters
$\mathcal{P}$	:	pose blendshape parameters
$\mathcal{W}$	:	blend weights
$\mathcal{J}$	:	joint regressor matrix
$\bar{\mathbf{T}} \in \mathbb{R}^{N,3}$	:	template mesh
$\mathbf{V} \in \mathbb{R}^{N,3}$	:	mesh vertices
$\bar{\mathbf{V}} \in \mathbb{R}^{N,3}$	:	canonical mesh vertices
		SMPL functions
B_P	:	pose blendshape function
B_S	:	shape blendshape function
		Estimated parameters
$\hat{\bar{x}} \in \mathbb{R}^3$	:	estimated canonical point
$\hat{\mathbf{V}} \in \mathbb{R}^{N,3}$	:	estimated mesh vertices
$\hat{\bar{\mathbf{V}}} \in \mathbb{R}^{N,3}$	:	estimated canonical mesh vertices
$w_{\hat{x}} \in \mathbb{R}^K$	:	weights predicted by the inverse LBS network
$w_{\hat{\bar{x}}} \in \mathbb{R}^K$	:	weights predicted by the forward LBS network

Table F.4. **Notation summary.**

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 3, 5

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[3] Matan Atzmon and Yaron Lipman. SALD: Sign agnostic learning with derivatives. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 3

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Proc. Neural Information Processing Systems (NeurIPS)*, December 2020. 3

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7, 8, 3

[6] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 3

[7] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 3

[8] Will Chang and Matthias Zwicker. Range scan registration using reduced deformable models. In *Computer Graphics Forum*, 2009. 2

[9] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[12] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 3

[13] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 3

[14] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017. 1

[15] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 5, 7

[16] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *Proc. International Conference on Learning Representations (ICLR)*, 2017. 1

[17] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005. 3

[18] Oren Freifeld and Michael J Black. Lie bodies: A manifold representation of 3d human shape. In *Proc. European Conference on Computer Vision (ECCV)*, 2012. 2

[19] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 5

[20] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multipurpose motion and video dataset. *arXiv preprint arXiv:2003.01888*, 2020. 7, 3, 6

[21] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proc. International Conference on Machine Learning (ICML)*, 2020. 3

[22] Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn, and Hans-Peter Seidel. Learning skeletons for shape and pose. In *Proc. ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2010. 2

[23] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 2

[24] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *Proc. German Conference on Pattern Recognition (GCPR)*, 2019. 1

[25] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[26] Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[27] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and JP Lewis. Skinning: Real-time shape deformation. In *ACM SIGGRAPH 2014 Courses*, 2014. 2

[28] Timothy Jeruzalski, David IW Levin, Alec Jacobson, Paul Lalonde, Mohammad Norouzi, and Andrea Tagliasacchi. NiLBS: Neural inverse linear blend skinning. *arXiv preprint arXiv:2004.05980*, 2020. 3

[29] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proc. Interna-*

tional Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2, 8

[30] Olaf Kähler, Victor Adrian Prisacariu, Carl Yuheng Ren, Xin Sun, Philip Torr, and David Murray. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Transactions on Visualization and Computer Graphics*, 2015. 3

[31] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[32] Tero Karras. Maximizing parallelism in the construction of bvhs, octrees, and k-d trees. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2012. 2

[33] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *Proc. International Conference on 3D Vision (3DV)*. IEEE, 2020. 3

[34] Ladislav Kavan and Jiří Žára. Spherical blend skinning: a real-time deformation of articulated models. In *Proc. Symposium on Interactive 3D graphics and games*, 2005. 2

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 7

[36] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 1

[37] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[38] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2000. 2

[39] Ming C Lin, Dinesh Manocha, Jon Cohen, and Stefan Gottschalk. Collision detection: Algorithms and applications. *Algorithms for robotic motion and manipulation*, 1997. 2

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015. 1, 2, 3, 5, 7

[41] Bruce Merry, Patrick Marais, and James Gain. Animation space: A truly linear framework for character animation. *ACM Transactions on Graphics*, 2006. 2

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3d reconstruction in function space. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[43] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 3

[44] Marko Mihajlovic, Silvan Weder, Marc Pollefeys, and Martin R. Oswald. DeepSurfels: Learning online appearance fusion. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[45] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[46] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics*, 2013. 3

[47] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[49] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[50] Ralf Plankers and Pascal Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003. 2

[51] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[52] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black. Learning multi-human optical flow. *International Journal of Computer Vision*, 2020. 1

[53] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 2017. 2, 7

[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[55] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[56] Shunsuke Saito, Jinlong Yang, Qianli Ma1, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[57] Hanan Samet. *The design and analysis of spatial data structures*. Addison-Wesley Reading, MA, 1990. 2

[58] Christian Sigg. *Representation and rendering of implicit surfaces*. PhD thesis, ETH Zurich, 2006. 3

[59] Vincent Sitzmann, Julien Martel, Alexander Bergman, David

Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 3

[60] Frank Steinbrucker, Christian Kerl, and Daniel Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *Proc. International Conference on Computer Vision (ICCV)*, 2013. 3

[61] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 8, 2, 3, 4

[62] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 2016. 2, 5

[63] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[64] Xiaohuan Corina Wang and Cary Phillips. Multi-weight enveloping: least-squares approximation techniques for skin animation. In *Annual Conference of the European Association for Computer Graphics (Eurographics)*, 2002. 2

[65] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[66] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019. 3

[67] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 3

[68] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 3

[69] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[70] Ming Zeng, Fukai Zhao, Jiaxiang Zheng, and Xinguo Liu. A memory-efficient kinectfusion using octree. In *International Conference on Computational Visual Media*, 2012. 3

[71] Ming Zeng, Fukai Zhao, Jiaxiang Zheng, and Xinguo Liu. Octree-based fusion for realtime 3d reconstruction. *Graphical Models*, 2013. 3

[72] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3d environments. In *Proc. International Conference on 3D Vision (3DV)*, 2020. 1, 2, 3, 6, 8, 4, 5

[73] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 6

[74] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 3