COAP: Compositional Articulated Occupancy of People

Supplementary Material

In this supplementary document, we provide additional implementation details (Sec. A) and qualitative and quantitative results (Sec. B).

A. Implementation Details

Network Architectures. The PointNet encoder in Sec. 4 is implemented as an eight-layer perceptron network interleaved with ReLU activations and skip connections as in the previous work [28]. The shared MLP occupancy decoder is illustrated in Figure B.2.

Sampling Strategy in the Local Shape Decomposition (Sec. 4.1). Each local articulated part in Figure 2 is temporarily represented as a point cloud by sampling points on the mesh surface. Each point is sampled by first selecting a mesh face with probability proportional to the face area and then randomly sampling barycentric coordinates in order to calculate a point on the selected face. To further balance the overlap among local articulated body parts, the kth point cloud allocates one half of its capacity to encode the central component corresponding to bone G_k, whereas the other half covers the whole local articulated body part region. This design guides the neural networks to properly learn localized occupancy fields, where the largest part is reserved to represent the core bone component, while fewer samples for the non-central parts encourage smooth interpolation between connected occupancy fields.

In all experiments, we used a total of 1000 samples per body part which are encoded as local body codes z_k with 128 dimensions.

B. Additional Results and Experiment Details

Additional Results. We provide additional qualitative results for the generalization experiment (Sec. 5) in Figure B.1 and additional quantitative results of two more baselines (NASA [8], LEAP [28]) in Table B.1 for the single-subject experiment.

Resolving Self-intersections (Sec. 5.2). For the baseline [35, 49], we used default configuration parameters provided by the authors except for the collision weight, which we increased from 0.0001 to 0.005 for better performance. Our self-intersection procedure uses standard gradient-based optimization with a learning rate of 0.007 and a total of 1300 query points sampled (arbitrarily chosen) in the intersected volume of colliding bounding boxes.

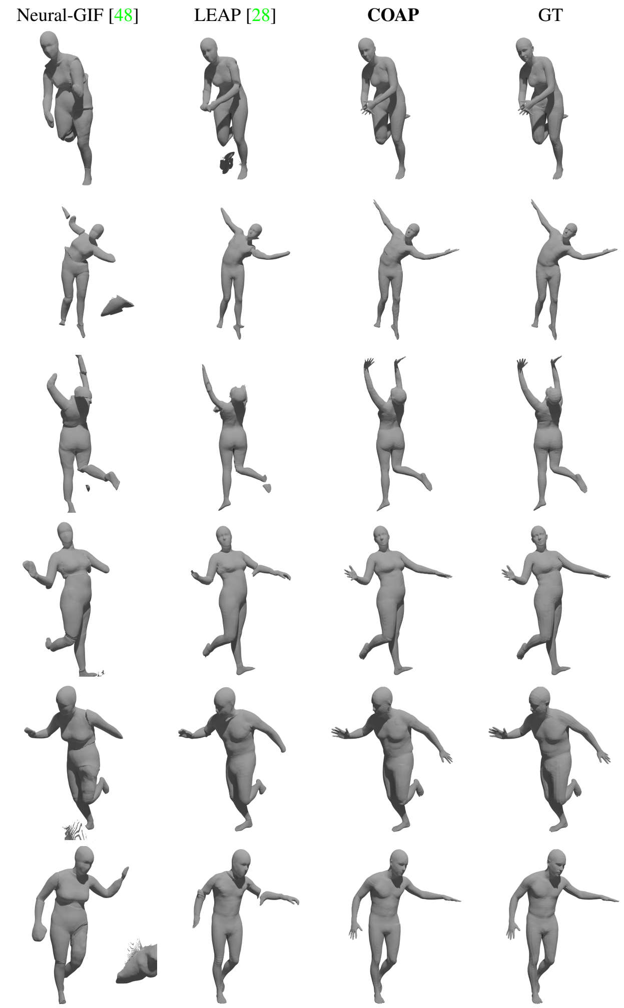

Figure B.1. **Generalization to unseen humans.** Comparison of our model with LEAP [28] and Neural-GIF [48] for the identities of the DFaust [5] and the PosePrior [1] datasets performing challenging novel poses from the PosePrior dataset. These qualitative results supplement results displayed in Figure 1 and Table 2.

Resolving Collisions with 3D Environments (Sec. 5.3). For the human-scene reconstruction pipeline, we use the optimization schedule from the PROX pipeline [13] with the original weighting terms. Our proposed collision term is added to the final optimization loss and weighted by 100. Please see the supplementary video for qualitative results. The optimization algorithm is sensitive to estimated joint locations and cannot resolve deep collisions with the environment (Figure B.3).

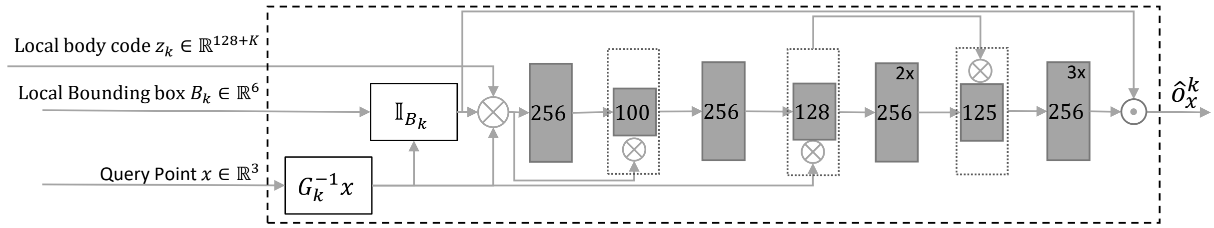

Figure B.2. **Architecture of the local MLP decoder.** The input parameters correspond to the kth articulated body part. Deterministic differentiable blocks are shown in black, fully-connected layers are shown in gray. the number inside each block denotes the dimensionality of the input feature vector, the number in the top-right corner denotes layer repetition, the operator $\otimes$ denotes feature concatenation, the operator $\odot$ denotes multiplication, $\mathbb{I}_{B_k}$ is an indicator function returning the value 1 if the local query is inside the bounding box B_k or 0 otherwise. All fully-connected layers are activated by Softplus with beta of 100 and a threshold of 20. The final output of the decoder is the occupancy prediction $\hat{o}_x^k$ for the kth articulated part.

	Female Subjects					Male Subjects				
Method	50004	50020	50021	50022	50025	50002	50007	50009	50026	50027
NASA [6]	77.75/77.68	55.93/80.20	90.99/78.13	90.87/77.86	71.20/78.64	68.14/74.82	67.57/71.82	44.84/74.32	87.44/77.47	48.84/79.30
LEAP [28]	88.53/67.05	90.42/77.84	89.84/76.15	88.18/64.79	91.33/77.09	74.67/35.31	83.65/53.83	84.04/65.81	88.78/68.29	90.76/77.35
SNARF [6]	95.75/84.32	95.42/86.32	95.43/86.07	**96.08**/85.47	95.57/85.01	96.05/82.50	**95.69/82.11**	94.44/83.41	95.35/83.41	95.22/84.91
COAP	**95.97/85.35**	**95.84/87.62**	**95.57/86.82**	95.98/**85.65**	**95.84/86.28**	**96.61/82.96**	95.27/81.90	**94.91/84.90**	**96.07/85.89**	**95.78/86.90**

Table B.1. **Single-subject neural implicit models.** Comparison with NASA, LEAP and SNARF [6] on per-subject training. There results supplement results displayed in Table 1.

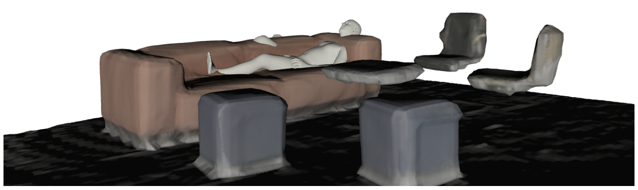

Figure B.3. **Limitation - resolving collisions with 3D environments.** The optimization algorithm has difficulties resolving deep collisions with the environment as demonstrated here for an example from the PROX dataset [13].

Steps:	1%	5%	10%	15%	25%	30%	40%	50%	60%	70%
IOU:	96.86	96.93	96.96	96.96	96.97	96.98	96.98	96.98	96.97	96.96

Table B.2. **Ablation of the bounding box size.**

Ablation of the bounding box size. We further study the impact of the size of the bounding boxes B_k on model performance. We compute the uniform IoU in Table B.2 for a varying number of up-sampling steps for the generalization experiment on the PosePrior dataset (Tab. 2 in the paper). Very tight boxes (less than 10% of the original size) slightly degrade the representation quality, while the performance saturates at 15%. We decided to use a tight box in this range for the experiments simply because these bounding boxes are used to detect an initial set of potentially collided body parts for resolving self-intersections. If the boxes are too large, the initial set of candidates would be larger and slow down the optimization.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 2, 4, 6, 8

[2] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Int. Conf. Comput. Vis.*, 2021. 3

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Eur. Conf. Comput. Vis.*, 2020. 2

[4] 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 *Eur. Conf. Comput. Vis.*, 2016. 2, 3

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2, 6, 1

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 5, 6, 8

[7] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[8] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: Neural Articulated Shape Approximation. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3, 4, 5, 1

[9] Christer Ericson. *Real-time collision detection*. Crc Press, 2004. 7

[10] 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. 6, 7

[11] Peng Guan. *Virtual human bodies with clothing and hair: From images to animation*. PhD thesis, Brown University Providence, RI, USA, 2012. 2, 3

[12] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Int. Conf. Comput. Vis.*, 2009. 3

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Int. Conf. Comput. Vis.*, 2019. 2, 3, 5, 7, 8, 1

[14] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and J. P. Lewis. Skinning: Real-time shape deformation (full text not available). In *ACM SIGGRAPH 2014 Courses*. ACM, 2014. 1

[15] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *ACM Trans. Graph.*, 2013. 3

[16] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3

[17] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Trans. Graph.*, 2008. 2

[18] Ladislav Kavan and Olga Sorkine. Elasticity-inspired deformers for character articulation. *ACM Trans. Graph.*, 2012. 2

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

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 5

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 8

[22] 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 *Int. Conf. Comput. Vis.*, 2019. 8

[23] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Computer graphics and interactive techniques*, 2000. 2

[24] Yijing Li and Jernej Barbič. Immersion of self-intersecting solids and surfaces. *ACM Trans. Graph.*, 2018. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1, 2, 3, 4, 5

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Int. Conf. Comput. Vis.*, 2019. 5

[27] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[28] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 3, 4, 5, 6, 8

[29] Neil Molino, Robert Bridson, and Ronald Fedkiw. Tetrahedral mesh generation for deformable bodies. In *Symposium on Computer Animation*, 2003. 2

[30] Matthieu Nesme, Paul G Kry, Lenka Jeřábková, and François Faure. Preserving topology and elasticity for embedded deformable models. In *ACM SIGGRAPH*. ACM, 2009. 2

[31] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, 2006. 7

[32] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Eur. Conf. Comput. Vis.*, 2020. 2

[33] Pablo Palafox, Aljaz Bozic, Justus Thies, Matthias Nießner, and Angela Dai. Neural parametric models for 3d deformable shapes. In *Int. Conf. Comput. Vis.*, 2021. 3

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[35] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 3, 7, 8

[36] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis.*, 2020. 3

[37] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon. Metric regression forests for correspondence estimation. *Int. J. Comput. Vis.*, 2015. 2, 3

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 4, 5

[39] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Int. Conf. Comput. Vis.*, 2021. 2, 7

[40] Damien Rohmer, Stefanie Hahmann, and Marie-Paule Cani. Exact volume preserving skinning with shape control. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2009. 2

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017. 2

[42] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 3

[43] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *CVPR*, 2021. 3

[44] Eftychios Sifakis, Kevin G Der, and Ronald Fedkiw. Arbitrary cutting of deformable tetrahedralized objects. In *ACM SIGGRAPH*, 2007. 2

[45] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Int. Conf. Comput. Vis.*, 2021. 2, 8

[46] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, M-P Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, et al. Collision detection for deformable objects. In *Computer graphics forum*, 2005. 3

[47] Jean-Marc Thiery, Émilie Guy, and Tamy Boubekeur. Sphere-meshes: Shape approximation using spherical quadric error metrics. *ACM Trans. Graph.*, 2013. 7

[48] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 5, 6, 8

[49] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.*, 2016. 3, 7, 8, 1

[50] Rodolphe Vaillant, Loïc Barthe, Gaël Guennebaud, Marie-Paule Cani, Damien Rohmer, Brian Wyvill, Olivier Gourmel, and Mathias Paulin. Implicit skinning: Real-time skin deformation with contact modeling. *ACM Trans. Graph.*, 2013. 2

[51] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[52] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Adv. Neural Inform. Process. Syst.*, 2021. 1, 3

[53] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *arXiv preprint arXiv:2111.11426*. 3

[54] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[55] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Int. Conf. Comput. Vis.*, 2021. 2, 3, 7

[56] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020. 3

[57] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3