

# Supplementary Material

Jingyao Cai<sup>1</sup>, Boyuan Cheng<sup>1</sup>, Yingjie Xi<sup>1</sup>, and Xiaosong Yang<sup>1</sup>

National Centre for Computer Animation, Bournemouth University, Poole Dorset  
BH12 5BB, United Kingdom

{s5604009, bcheng, yxi, xyang}@bournemouth.ac.uk

## 1 Preliminaries

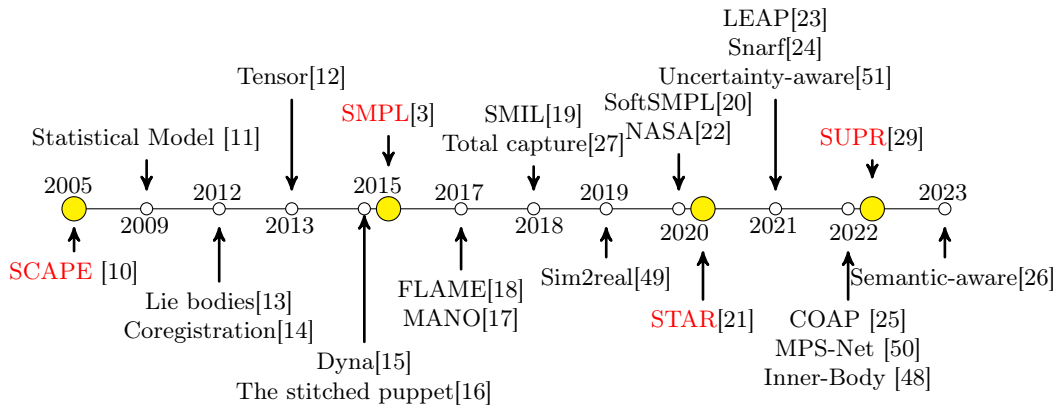
Reconstructing multiple 3D human bodies from monocular input is a quite challenging task. This is primarily due to the need to predict uncertain information such as global positions in 3D space from 2D images, compounded by the complexity arising from indistinct boundaries between the human body and the background. Consequently, most current research on multi-person reconstruction is based on single-person reconstruction algorithms, attempting to adapt and improve them for multi-person scenarios.

As a foundational and preliminary work for multi-person reconstruction, single-person reconstruction algorithms focus on extracting and outputting the 3D human body mesh of an individual from a single image or a monocular video. At first, geometric primitives such as planar rectangles[4], cylinders[5], ellipsoids[6], superquadric ellipsoids[7], metaballs[8] and customized graphical model[9] are usually used to approximate and model human body shape in order to represent human in 3D space. Subsequently, in order to achieve a more detailed representation of human body, later research shifts its focus to learning a statistical human model from an extensive collection of 3D human body scanning data. Therefore, recent research of reconstructing single 3D human model can be categorized into two approaches. The first and well-established method is to define a parametric model, where a set of parameters is used to characterize a human body shape and pose. The second method involves reconstructing the human model with comprehensive details, including clothing, hair and facial features, usually referred to as “clothed human reconstruction”. This method primarily utilizes implicit function to get a more detailed and precise clothed human model.

### 1.1 Parametric human model

A parametric model serves as an approximation of 3D human body shape and pose to be reconstructed.

SCAPE[10] is the first publications introducing parametric models to generate a complete human mesh. It employs triangle deformations to reshape the human body with shape deformation and pose deformation separately. There are many models such as [11], [12], [13], [14], [15] and [16] built upon SCAPE as its pioneering contributions.



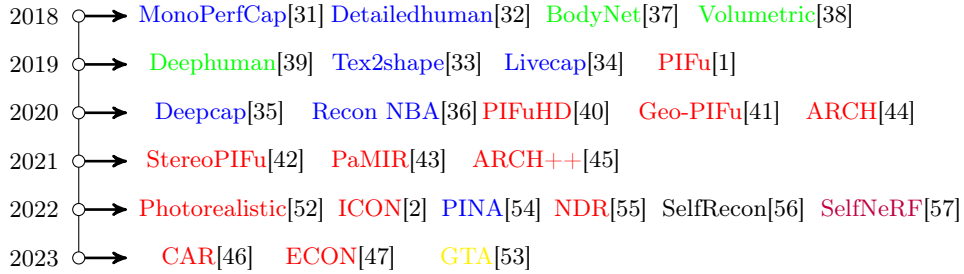
**Fig. 1.** Chronological overview of essential parametric human models of single-person 3D reconstruction.

In 2015, SMPL [3] was published with a realistic, naturally deformable and efficiently animated 3D human parametric model which was widely acknowledged. It currently stands as the most commonly used human body model in the research community. SMPL also utilizes two sets of parameters for shape and pose to generate the corresponding deformations. It provides a differentiable function for generating final 3D model. Building upon SMPL, numerous methods that are more precise and effective have been proposed in recent years. MANO [17] is designed for modelling human hands, FLAME [18] for modelling human heads and SMIL [19] for modelling infant bodies, each are proposed for specific body regions and shapes. In pursuit of enhanced descriptive capability, SoftSMPL [20] and STAR [21] achieved superior performance. Moreover, more different learning-based methods are proposed such as implicit models [22], [23], [24], [25] and explicit model [26]. For whole-body modeling including hands or heads, SMPL+H [17] combines MANO and SMPL to create a parametric human body model with hands. The Frankenstein Model [27] incorporates an artist-designed hand rig and the FaceWarehouse face model [28] into a simplified version of SMPL. SMPL-X [58] truly achieves a comprehensive model that utilizes more detailed parameters to represent the human model with body, hands and facial expressions. Recently, SUPR [29] is proposed for generating whole-body model more precisely and vividly and GHUMGHUML [30] generate the parametric model with variational auto encoders.

Figure 1 presents core methods for single-person reconstruction using parametric models in chronological order.

## 1.2 Clothed human model

Clothed human model reconstruction refers to generating a photorealistic 3D representation of a clothed body shape based on 2D monocular input. It aims to



**Fig. 2.** Chronological overview of primary methods for clothed human model reconstruction in single-person 3D modeling. In this figure, blue represents reconstruction methods based on deformation, green represents voxel-based reconstruction, red represents methods based on pixel-aligned of implicit functions, yellow is based on transformer, purple is based on NeRF, and black utilizes multiple methods for reconstruction.

model a body mesh that closely and accurately resembles the real state of the human depicted in the input including hair, clothing and facial expressions.

In 2018, MonoPerfCap[31] applied deformation to SMPL for modeling clothed human. Paper[32] used vertex displacement divided from SMPL model on monocular videos and increased details based on it to create a realistic model. Subsequently, Tex2shape[33], Livecap[34], Deepcap[35] and Reconstructing NBA players[36] also utilized similar methods to generate photorealistic human models with detailed hair and clothing.

Voxel, regarded as a fundamental representation of 3D mesh is widely used to generate fully detailed 3D clothed human models. Building on voxel, BodyNet[37] predicted the human body shape with the details of volumetric representation from a single image in 2018. Gilbert et al.[38] proposed a method to transform the coarse human model into a more detailed 3D human representation. Then, Deephuman[39] was purposed to refine the voxelized SMPL model to clothed model by transerring the image features into the voxel space. However, voxel usually has limitation for taking up too much memory when restoring high geometry.

To acquire more accurate model with less memory-consuming, implicit function has been purposed for its strength in modeling details. It can generate high-resolution 3D human mesh even for some loose-fitting clothing model and distinctive poses. PIFu[1] as the pioneering and representative work aligns image pixels with 3D global shape and texture of the realistic input. It utilizes implicit function to detect whether the points in 3D space corresponding to the pixels in the 2D image are on the surface of human body. As a result, it allows for more detailed effects, including hair, loose clothing and exaggerated poses. Subsequently, many methods for improving performance based on PIFu have also been proposed, such as PIFuHD[40], Geo-PIFu[41], StereoPIFu[42] and PaMIR[43]. And there are also many related works combined with other methods purposed to

achieve better results. For example, ARCH[44], ARCH++[45] and CAR[46] use SMPL to unpose the pixel-aligned query points from a posed space to a canonical space. ICON[2] and ECON[47] regress shapes from 2D normal maps and depth maps.

Figure 2 presents main methods for clothed human reconstruction in chronological order.

## References

1. Saito S, Huang Z, Natsume R, et al. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2304-2314.
2. Xiu Y, Yang J, Tzionas D, et al. Icon: Implicit clothed humans obtained from normals[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 13286-13296.
3. Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 851-866.
4. Ju S X, Black M J, Yacoob Y. Cardboard people: A parameterized model of articulated image motion[C]//Proceedings of the Second International Conference on Automatic Face and Gesture Recognition. IEEE, 1996: 38-44.
5. Marr D, Nishihara H K. Representation and recognition of the spatial organization of three-dimensional shapes[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1978, 200(1140): 269-294.
6. Wang M, Qiu F, Liu W, et al. Monocular human pose and shape reconstruction using part differentiable rendering[C]//Computer Graphics Forum. 2020, 39(7): 351-362.
7. Sminchisescu C, Triggs B. Estimating articulated human motion with covariance scaled sampling[J]. The International Journal of Robotics Research, 2003, 22(6): 371-391.
8. Plänkers R, Fua P. Tracking and modeling people in video sequences[J]. Computer Vision and Image Understanding, 2001, 81(3): 285-302.
9. Kakadiaris L, Metaxas D. Model-based estimation of 3D human motion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1453-1459.
10. Anguelov D, Srinivasan P, Koller D, et al. Scape: shape completion and animation of people[M]//ACM SIGGRAPH 2005 Papers. 2005: 408-416.
11. Hasler N, Stoll C, Sunkel M, et al. A statistical model of human pose and body shape[C]//Computer graphics forum. Oxford, UK: Blackwell Publishing Ltd, 2009, 28(2): 337-346.
12. Chen Y, Liu Z, Zhang Z. Tensor-based human body modeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 105-112.
13. Freifeld O, Black M J. Lie bodies: A manifold representation of 3D human shape[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12. Springer Berlin Heidelberg, 2012: 1-14.
14. Hirshberg D A, Loper M, Rachlin E, et al. Coregistration: Simultaneous alignment and modeling of articulated 3D shape[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12. Springer Berlin Heidelberg, 2012: 242-255.

15. Pons-Moll G, Romero J, Mahmood N, et al. Dyna: A model of dynamic human shape in motion[J]. *ACM Transactions on Graphics (TOG)*, 2015, 34(4): 1-14.
16. Zuffi S, Black M J. The stitched puppet: A graphical model of 3d human shape and pose[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 3537-3546.
17. Romero J, Tzionas D, Black M J. Embodied hands: Modeling and capturing hands and bodies together[J]. *arXiv preprint arXiv:2201.02610*, 2022.
18. Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. *ACM Trans. Graph.*, 2017, 36(6): 194:1-194:17.
19. Hesse N, Pujades S, Romero J, et al. Learning an infant body model from RGB-D data for accurate full body motion analysis[C]//*Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer International Publishing, 2018: 792-800.
20. Santesteban I, Garces E, Otaduy M A, et al. SoftSMPL: Data-driven Modeling of Nonlinear Soft-tissue Dynamics for Parametric Humans[C]//*Computer Graphics Forum*. 2020, 39(2): 65-75.
21. Osman A A A, Bolkart T, Black M J. Star: Sparse trained articulated human body regressor[C]//*Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI* 16. Springer International Publishing, 2020: 598-613.
22. Deng B, Lewis J P, Jeruzalski T, et al. Nasa neural articulated shape approximation[C]//*Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII* 16. Springer International Publishing, 2020: 612-628.
23. Mihajlovic M, Zhang Y, Black M J, et al. LEAP: Learning articulated occupancy of people[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021: 10461-10471.
24. Chen X, Zheng Y, Black M J, et al. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 11594-11604.
25. Mihajlovic M, Saito S, Bansal A, et al. COAP: Compositional articulated occupancy of people[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 13201-13210.
26. Sun X, Feng Q, Li X, et al. Learning semantic-aware disentangled representation for flexible 3D human body editing[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 16985-16994.
27. Joo H, Simon T, Sheikh Y. Total capture: A 3d deformation model for tracking faces, hands, and bodies[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 8320-8329.
28. Cao C, Weng Y, Zhou S, et al. Facewarehouse: A 3d facial expression database for visual computing[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 20(3): 413-425.
29. Osman A A A, Bolkart T, Tzionas D, et al. Supr: A sparse unified part-based human representation[C]//*European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022: 568-585.
30. Xu H, Bazavan E G, Zanfir A, et al. Ghum & ghuml: Generative 3d human shape and articulated pose models[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 6184-6193.
31. Xu W, Chatterjee A, Zollhöfer M, et al. Monoperfcap: Human performance capture from monocular video[J]. *ACM Transactions on Graphics (ToG)*, 2018, 37(2): 1-15.

32. Alldieck T, Magnor M, Xu W, et al. Detailed human avatars from monocular video[C]//2018 International Conference on 3D Vision (3DV). IEEE, 2018: 98-109.
33. Alldieck T, Pons-Moll G, Theobalt C, et al. Tex2shape: Detailed full human body geometry from a single image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2293-2303.
34. Habermann M, Xu W, Zollhoefer M, et al. Livecap: Real-time human performance capture from monocular video[J]. ACM Transactions On Graphics (TOG), 2019, 38(2): 1-17.
35. Habermann M, Xu W, Zollhofer M, et al. Deepcap: Monocular human performance capture using weak supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5052-5063.
36. Zhu L, Rematas K, Curless B, et al. Reconstructing nba players[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 177-194.
37. Varol G, Ceylan D, Russell B, et al. Bodynet: Volumetric inference of 3d human body shapes[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 20-36.
38. Gilbert A, Volino M, Collomosse J, et al. Volumetric performance capture from minimal camera viewpoints[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 566-581.
39. Zheng Z, Yu T, Wei Y, et al. Deephuman: 3d human reconstruction from a single image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7739-7749.
40. Saito S, Simon T, Saragih J, et al. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 84-93.
41. He T, Collomosse J, Jin H, et al. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction[J]. Advances in Neural Information Processing Systems, 2020, 33: 9276-9287.
42. Hong Y, Zhang J, Jiang B, et al. Stereopifu: Depth aware clothed human digitization via stereo vision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 535-545.
43. Zheng Z, Yu T, Liu Y, et al. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(6): 3170-3184.
44. Huang Z, Xu Y, Lassner C, et al. Arch: Animatable reconstruction of clothed humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3093-3102.
45. He T, Xu Y, Saito S, et al. Arch++: Animation-ready clothed human reconstruction revisited[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11046-11056.
46. Liao T, Zhang X, Xiu Y, et al. High-fidelity clothed avatar reconstruction from a single image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8662-8672.
47. Xiu Y, Yang J, Cao X, et al. Econ: Explicit clothed humans optimized via normal integration[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 512-523.
48. Li X, Huang J, Zhang J, et al. Learning to infer inner-body under clothing from monocular video[J]. IEEE Transactions on Visualization and Computer Graphics, 2022.

49. Doersch C, Zisserman A. Sim2real transfer learning for 3d human pose estimation: motion to the rescue[J]. *Advances in Neural Information Processing Systems*, 2019, 32.
50. Wei W L, Lin J C, Liu T L, et al. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 13211-13220.
51. Lee G H, Lee S W. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 12375-12384.
52. Alldieck T, Zanfir M, Sminchisescu C. Photorealistic monocular 3d reconstruction of humans wearing clothing[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 1506-1515.
53. Zhang Z, Sun L, Yang Z, et al. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction[J]. *Advances in Neural Information Processing Systems*, 2024, 36.
54. Dong Z, Guo C, Song J, et al. PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 20470-20480.
55. Cai H, Feng W, Feng X, et al. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 967-981.
56. Jiang B, Hong Y, Bao H, et al. Selfrecon: Self reconstruction your digital avatar from monocular video[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 5605-5615.
57. Peng B, Hu J, Zhou J, et al. Selfnerf: Fast training nerf for human from monocular self-rotating video[J]. *arXiv preprint arXiv:2210.01651*, 2022.
58. Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019: 10975-10985.