H4MER: Human 4D Modeling by Learning Neural Compositional Representation With Transformer

Boyan Jiang, Yinda Zhang, Jingyang Huo, *Graduate Student Member, IEEE*, Xiangyang Xue, *Member, IEEE*, and Yanwei Fu, *Member, IEEE*

***Abstract*—Despite the impressive results achieved by deep learning based 3D reconstruction, the techniques of directly learning to model 4D human captures with detailed geometry have been less studied. This work presents a novel neural compositional representation for Human 4D Modeling with transformER (H4MER). Specifically, our H4MER is a compact and compositional representation for dynamic human by exploiting the human body prior from the widely used SMPL parametric model. Thus, H4MER can represent a dynamic 3D human over a temporal span with the codes of shape, initial pose, motion and auxiliaries. A simple yet effective linear motion model is proposed to provide a rough and regularized motion estimation, followed by per-frame compensation for pose and geometry details with the residual encoded in the auxiliary codes. We present a novel Transformer-based feature extractor and conditional GRU decoder to facilitate learning and improve the representation capability. Extensive experiments demonstrate our method is not only effective in recovering dynamic human with accurate motion and detailed geometry, but also amenable to various 4D human related tasks, including monocular video fitting, motion retargeting, 4D completion, and future prediction.**

***Index Terms*—4D representation, compositional representation, human modeling, transformer.**

I. Introduction

THE vanilla SMPL based parametric representations have been extensively studied and widely utilized for modeling 3D human shapes. These representations have critical impacts to many human-centric tasks, such as pose estimation [1], [2], [3] and body shape fitting [4], [5], [6], [7], [8]. Unfortunately,

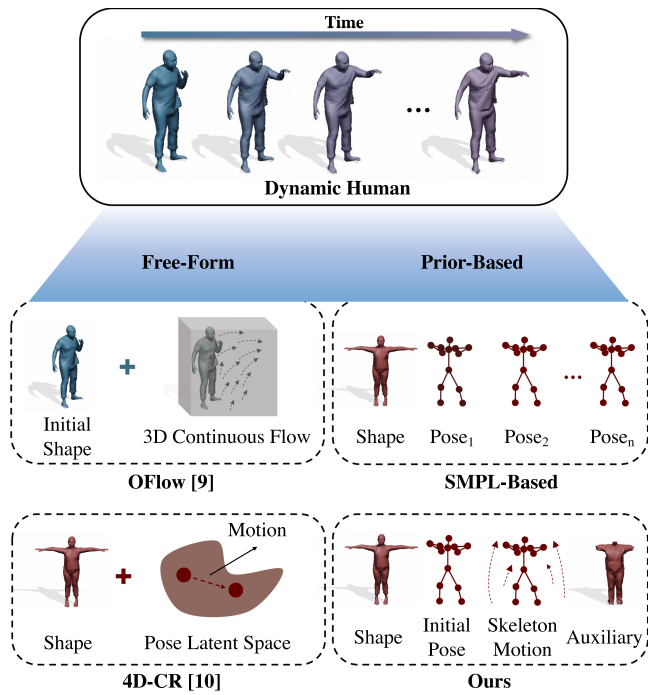

Fig. 1. *Comparison with existing 4D human representations.* Our representation supports faster inference and more complete reconstructions compared with free-form methods (Fig. 3). And it provides the long-range temporal context and additional fine-grained geometry controlled by low-dimensional codes, which is more compact compared with previous SMPL-based methods.

these vanilla 3D representations are arguably insufficient for the applications involving dynamic/temporal signals concerned in this paper such as 3D moving humans (Fig. 1 top), as the temporal information is not captured.

There are only a few works on the representations of 4D human modeling. These works are roughly categorized into free-form [9], [10] and prior-based methods [11], [12], [13] depending on the 3D representation of the output shape (Fig. 1). The free-form methods leveraging Neural ODE [14] and deep implicit function [9], [10] often rely on computationally expensive architectures to learn the compact latent spaces and reconstruct 4D sequences. Unfortunately, since the human body prior is not explicitly modeled, the reconstruction results of these methods may contain obvious geometry artifacts such as missing hands, and their modeling errors accumulate rapidly over time. On the other hand, prior-based methods [11], [12], [13] are mostly derived from the SMPL parametric model [15].

Manuscript received 19 October 2022; revised 3 June 2023; accepted 28 August 2023. Date of publication 11 September 2023; date of current version 3 November 2023. This work was supported in part by NSFC Project under Grant 62076067. Recommended for acceptance by K.G. Derpanis. *(Corresponding authors: Yanwei Fu; Xiangyang Xue; Yinda Zhang.)*

Boyan Jiang and Xiangyang Xue are with the School of Computer Science, Fudan University, Shanghai 200437, China (e-mail: jiangboyan96@gmail.com; xyxue@fudan.edu.cn).

Yinda Zhang is with the Google, Mountain View, CA 94043 USA (e-mail: zhangyinda@gmail.com).

Jingyang Huo is with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200437, China (e-mail: jyhuo22@m.fudan.edu.cn).

Yanwei Fu is with the School of Data Science and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200437, China, and also with Fudan ISTBI–ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, Zhejiang 321017, China (e-mail: yanweifu@fudan.edu.cn).

The video demos are in the project homepage: https://boyanjiang.github.io/H4MER/.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3313311, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3313311

[73] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3170–3184, Jun. 2022.

[74] S. Saito, J. Yang, Q. Ma, and M. J. Black, "Scanimate: Weakly supervised learning of skinned clothed avatar networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2886–2897.

[75] S. Wang, A. Geiger, and S. Tang, "Locally aware piecewise transformation fields for 3D human mesh registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7639–7648.

[76] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1175–1186.

[77] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *Proc. IEEE Int. Conf. 3D Vis.*, 2018, pp. 98–109.

[78] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman, "Photo wake-up: 3D character animation from a single photo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5908–5917.

[79] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2293–2303.

[80] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *Proc. Int. Conf. 3D Vis.*, 2019, pp. 643–653.

[81] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao, "BCNet: Learning body and cloth shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 18–35.

[82] D. Mehta et al., "XNect: Real-time multi-person 3D human pose estimation with a single RGB camera," 2019, *arXiv: 1907.00837*.

[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[84] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[85] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "Humor: 3D human motion model for robust pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11 488–11 499.

[86] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3D human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3395–3404.

[87] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 601–617.

[88] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 399–405, 2004.

[89] P. Palafox, A. Božič, J. Thies, M. Nießner, and A. Dai, "NPMs: Neural parametric models for 3D deformable shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12 695–12 705.

[90] A. Jaegle et al., "Perceiver IO: A general architecture for structured inputs & outputs," 2021, *arXiv:2107.14795*.

Boyan Jiang received the BE degree in information security from Hangzhou Dianzi University, Hangzhou, China, in 2018, and the PhD degree in computer science & technology from Fudan University, Shanghai, China, in 2023. His research interests include human modeling, neural rendering, 3D/4D representation and reconstruction.

Yinda Zhang received the bachelor's degree from Department of Automation in Tsinghua University, the master's degree from Department of ECE in National University of Singapore co-supervised by Prof. Ping Tan and Prof. Shuicheng Yan, and the PhD degree in computer science from Princeton University, advised by Professor Thomas Funkhouser. He is a research scientist and manager with Google. His research interests lie at the intersection of computer vision, computer graphics, and machine learning. Recently, he focuses on empowering 3D vision and perception via machine learning, including dense depth estimation, 3D shape analysis, 3D scene understanding, and neural rendering.

Jingyang Huo (Graduate Student Member, IEEE) received the BS degree in mathematics from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. She is currently working toward the PhD degree in mathematics from Fudan University with supervisor Dr. Yanwei Fu. Her research interests include human modeling and multimodal learning.

Xiangyang Xue (Member, IEEE) received the BS, MS, and PhD degrees in communication engineering from Xidian University, Xian, China, in 1989, 1992, and 1995, respectively. He is currently a professor of computer science with Fudan University, Shanghai, China. His research interests include multimedia information processing and machine learning.

Yanwei Fu (Member, IEEE) received the MEng degree from the Department of Computer Science and Technology, Nanjing University, China, in 2011, and the PhD degree from the Queen Mary University of London, in 2014. He held a post-doctoral position with Disney Research, Pittsburgh, PA, from 2015 to 2016. He is currently a tenure-track professor with Fudan University. He was appointed as the professor of Special Appointment (Eastern Scholar) with Shanghai Institutions of Higher Learning. His work has led to many awards, including the IEEE ICME 2019 best paper. He published more than 100 journal/conference papers including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Multimedia*, ECCV, and CVPR. His research interests are one-shot learning, and learning-based 3D reconstruction.