TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation

Supplementary Material

1. Introduction

In this supplemental document, we provide more implementation details and discuss limitations of TokenHMR. Please refer to the **supplemental video** for a brief review of the paper and more qualitative results.

2. More Implementation Details

2.1. Data Preparation for Tokenizer

For pose tokenization, we use 21 body pose parameters following Vposer [7]. As shown in Tab. 3 of main paper, we evaluate our tokenization in two settings: in-distribution and out-of-distribution. For in-distribution, we train on the training set of AMASS [6] and evaluate on the test set of AMASS. To show the efficacy of tokenization, we also evaluate on an out-of-distribution yoga dataset, MOYO [9]. For training, we use the following datasets: {`CMU`, `KIT`, `BMLrub`, `DanceDB`, `BMLmovi`, `EyesJapan`, `BMLhandball`, `TotalCapture`, `EKUT`, `ACCAD`, `TCDHands`, `MPI-Limits`} with a weighting of {0.14, 0.14, 0.14, 0.06, 0.06, 0.06, 0.06, 0.06, 0.04, 0.04, 0.04, 0.16}, respectively.

2.2. Joint-wise Thresholds for TALS

To establish effective joint-wise thresholds for TALS (Sec. 4.2), we conducted a detailed statistical analysis on the 20221018_3-8_250_batch01hand_6fps validation subset of the BEDLAM [2] dataset, encompassing over 34k samples of diverse human 3D pose, shape, and camera perspectives. Table 1 presents the threshold distances for each joint used by TALS.

2D Joints	Threshold	SMPL Joints	Threshold
OP Nose	0.00850	Pelvis	0.46
OP Neck	0.00649	LHip	0.22
OP RShoulder	0.00748	RHip	0.21
OP RElbow	0.01103	Spine	0.15
OP RWrist	0.01356	LKnee	0.33
OP LShoulder	0.00742	RKnee	0.30
OP LElbow	0.01097	Thorax	0.17
OP LWrist	0.01414	LAnkle	0.20
OP MidHip	0.00974	RAnkle	0.27
OP RHip	0.01127	Thorax	0.12
OP RKnee	0.01663	LToe	0.29
OP RAnkle	0.00565	RToe	0.28
OP LHip	0.01126	Neck	0.24
OP LKnee	0.01616	LCollar	0.26
OP LAnkle	0.00533	RCollar	0.26
OP REye	0.00830	Jaw	0.28
OP LEye	0.00831	LShoulder	0.29
OP REar	0.00737	RShoulder	0.32
OP LEar	0.00743	LElbow	0.35
OP LBigToe	0.00544	RElbow	0.35
OP LSmallToe	0.00551	LWrist	0.62
OP LHeel	0.00536	RWrist	0.59
OP RBigToe	0.00565	LHand	0.20
OP RSmallToe	0.00582	RHand	0.20
OP RHeel	0.00573		
LSP RAnkle	0.00554		
LSP RKnee	0.01515		
LSP RHip	0.00986		
LSP LHip	0.00998		
LSP LKnee	0.01520		
LSP LAnkle	0.00511		
LSP RWrist	0.01288		
LSP RElbow	0.01106		
LSP RShoulder	0.00711		
LSP LShoulder	0.00710		
LSP LElbow	0.01092		
LSP LWrist	0.01388		
LSP Neck	0.00648		
LSP Head Top	0.00766		
MPII Pelvis	0.00931		
MPII Thorax	0.00647		
H36M Spine	0.00677		
H36M Jaw	0.00744		
H36M Head	0.00752		

Table 1. Thresholds for 44 2D joints and 24 SMPL joints. 2D joint names start with the skeleton origin, where OP stands for OpenPose [3]. LSP [5], MPII [1], and H36M [4] are the datasets.

2.3. Augmentations

Data augmentation plays a pivotal role in enhancing the robustness and generalization capabilities of HPS regressors. Hence, following HMR2.0, we perform various augmentations. These include random translations in both x and y directions with a factor of 0.02, scaling with a factor of 0.3 and rotations with 30 degrees. Other augmentations include horizontal flipping and color rescaling. We observe that extreme cropping i.e. removing part of the human body limb in random also improves the robustness to occlusion.

3. Discussion

3.1. Pose Space Analysis

We analyse the pose space by evaluating reconstruction of OOD poses that are not present in the training set. We do this by training on AMASS and testing on MOYO. The qualitative result is shown in Fig. 1 which shows good generalization to the out-of-distribution yoga poses from MOYO [9]. In contrast, we find that noisy test poses are not well recovered.

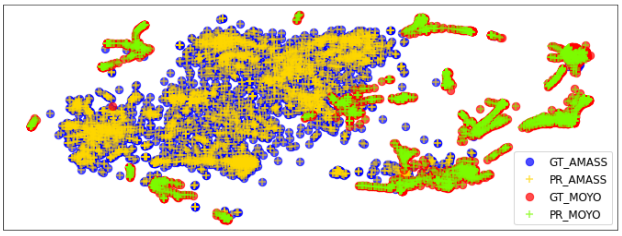


Figure 1. t-SNE visualization of *unseen poses* (3D body joints) reconstructed by our tokenizer trained on AMASS only. We are able to reconstruct the out-of-distribution Yoga poses from MOYO. GT is ground-truth poses and PR is predicted poses.

3.2. TALS loss vs Filtering Strategy

Similar to HMR2.0, we employ filtering strategies to ensure high-quality 2D image alignment of the p-GT. Filtering strategies, however, are "all or nothing"; i.e. data samples are either rejected or considered. Our TALS loss is different in that it uses all the filtered pseudo-ground-truth samples up to a threshold, after which the supervision is scaled down. This goes beyond standard filtering and data cleaning pipelines.

4. Limitation Discussion

4.1. Poor 2D Alignment under Weak-perspective Camera Model

The experimental analysis in Sec. 3 shows that using existing flawed camera projection models results in overfitting to 2D keypoints and that this leads to learning biased poses. To avoid this issue, we design a lenient TALS supervision training strategy and incorporate prior knowledge through our token-based pose representation. As shown in Fig. 2 a), with the combination of loose 2D supervision using TALS and built-in prior in representation, TokenHMR is able to estimate reasonable 3D poses but these do not always align well in 2D image when there is foreshortening. As expected under the weak-perspective camera model, the more obvious the perspective distortion, the worse the 2D alignment.

4.2. Failure Cases

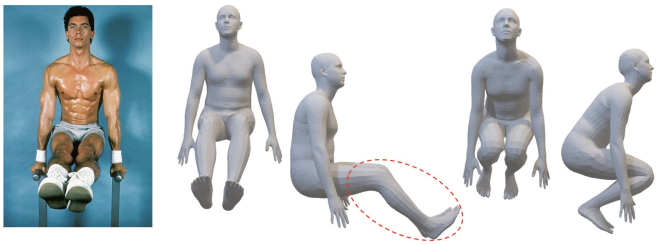
In this work, we introduce TokenHMR to reduce camera/pose bias and alleviate the ambiguity with a tokenized pose prior. However, TokenHMR still has some limitations that could be further explored in future work.

As shown in Fig. 2 b), foreshortening remains challenging without a better camera model. In cases like Fig. 2 c), the global orientation is ambiguous when only considering body cues. We may need to exploit more cues from the face and the feet to determine the correct global orientation. Future work could try to extend TokenHMR to full-body pose estimation (i.e. SMPL-X) to address this issue.

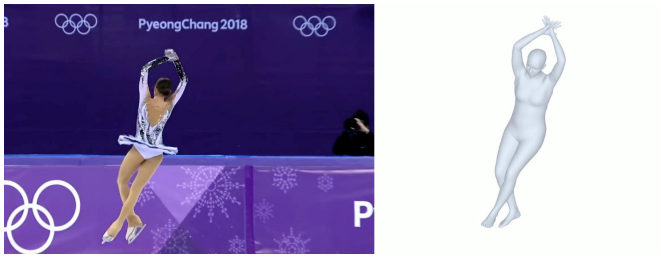
5. Future Work

Future work should, obviously, address the camera projection problem directly by recovering more accurate camera estimates.

a) Due to the loose supervision of TALS, our prediction does not align well in 2D under weak-perspective camera.



b) Depth-wise ambiguity is still very challenging.



c) Global orientation estimation sometimes fails because facial and foot cues are not thoroughly explored.

Figure 2. 2D alignment problem and failure cases.

Even with such improvements, we anticipate that the token representation retains value as it consistently improves performance across varied test scenarios. A promising next step is to extend the tokenization over time. Recent work on generating human motion from text exploits tokenized representations of human motions [8]. Looking further ahead, an intriguing direction for future research involves exploring the application of our token-based pose representation with Large Language Models (LLMs). The discrete, robust nature of our pose tokens, designed for 3D human pose estimation, presents an opportunity to bridge the gap between computer vision and natural language processing.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern*

Recognition (CVPR), 2014. 1

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 1

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 1

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 1

[5] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[6] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[8] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[9] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1