AJAHR: Amputated Joint Aware 3D Human Mesh Recovery
Supplementary Material

Hyunjin Cho[1,3*] Giyun Choi[1*] Jongwon Choi[1,2†]

[1]Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Korea [2]Dept. of Artificial Intelligence, Chung-Ang University, Korea
[3]Korea Institute of Industrial Technology (KITECH), Korea
{jincho, cky}@vilab.cau.ac.kr, choijw@cau.ac.kr

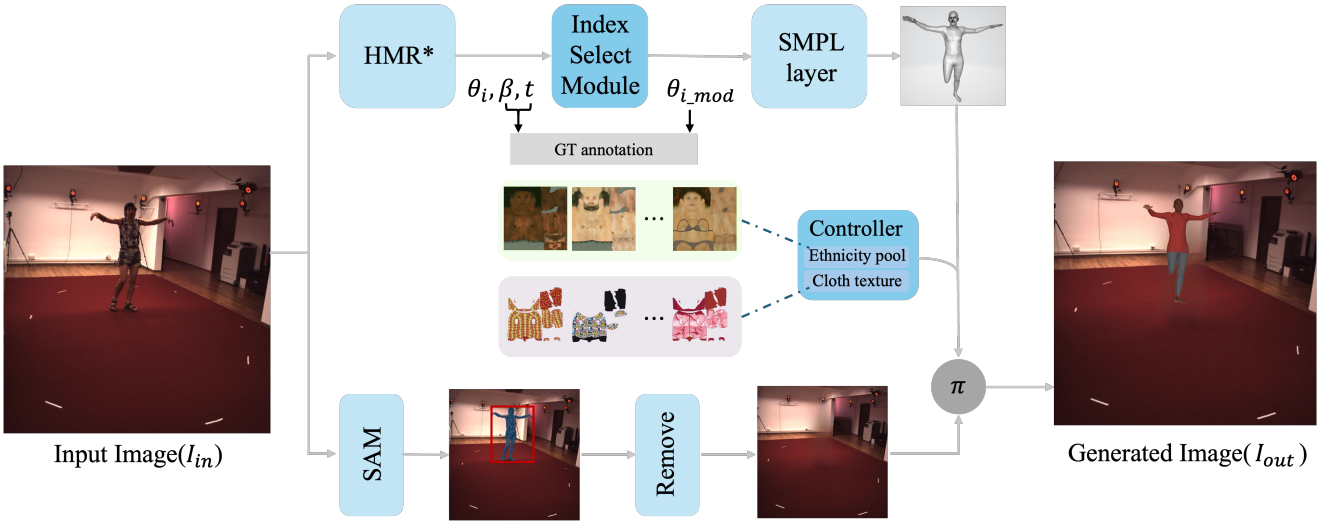

Figure A. **Pipeline for Synthesizing Images of Individuals with Amputations.** The input image (I_{in}) is sourced from benchmark datasets [1, 6, 11] and processed through ScoreHMR [16], an HMR-based model (*), which infers SMPL parameters from the image. We extracted the human region using SAM (Segment Anything Model) [10] for masking and then removed the region using LaMa [17] to generate the background. Finally, the generated human mesh was projected onto the background image to create the final image(I_{out}).

We provide additional details about AJAHR in the supplementary material. Sec. A introduces the Amputee 3D (A3D) dataset. Sec. B explains how we leveraged the SMPL prior to effectively represent amputations on the mesh and details the dataset synthesis pipeline for the A3D Dataset. Sec. B shows the visual effects of applying non-zero SMPL pose parameters. Sec. D and Sec. E evaluates A3D quality and shows that increasing its proportion improves in-the-wild performance without quality degradation. In Sec. F, Sec. G, Sec. H, introduces detail the architectures, training schemes, and hyperparameters of AJAHR, BPAC-Net, and the AJAHR-Tokenizer, respectively. Sec. I presents quantitative evaluations of the proposed AJAHR model's tuning strategy and variations in module configuration. Lastly, Sec. J presents ablation studies on the number of tokens and codebook size in the AJAHR-Tokenizer.

* Equal contribution. †Corresponding author.

A. A3D Dataset Synthesis Pipeline Details

The detailed process of the data synthesis pipeline is presented in Fig. A. We employ ScoreHMR [16], which refines predictions by incorporating 2D image cues during inference, making it more effective than conventional regression models for estimating plausible human poses.

To effectively simulate amputated body parts in the SMPL [12] representation, we introduce the index select module. This module assigns an index from 0 to 11, representing different amputation types, to the 24 joints of the SMPL body model. The module selects all indices corresponding to the amputated region and its connected child joints, setting their SMPL pose parameters to a zero matrix. This modified SMPL pose is subsequently processed by the SMPL layer, generating a mesh representation of the amputated body.

In the controller module, we adopt BEDLAM [2]'s dataset generation approach, which includes incorporating

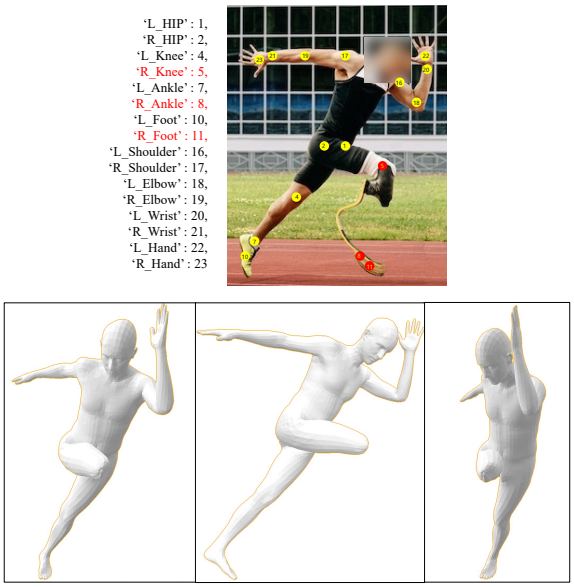

Figure B. The top image illustrates the visualization of SMPL [12] Skeleton Hierarchy and 3D Joint index mapping The bottom image presents multiple views of the scenario where the R_Knee (index 5) is the parent joint, while R_Ankle (index 8) and R_Foot (index 11) are its corresponding child joints. Setting the R_Knee to a zero matrix causes all vertex positions associated with these child joints to converge toward the parent joint, as visualized from different angles.

skin and clothing textures to ensure a balanced distribution across two genders (male and female) and seven ethnic groups. This process enhances the diversity of synthetic representations in terms of both ethnicity and appearance. These textures are synthesized onto mannequins in various poses, allowing us to generate images of individuals with diverse limb amputations.

Unlike other works in text-to-image or text-to-motion generation, our approach leverages widely used datasets in the human pose estimation field, such as H36M, MPII, and MSCOCO [1, 6, 11], ensuring a broad range of diverse poses. Furthermore, by leveraging our synthesis pipeline, we reduce the burden of manually generating diverse poses using motion generation models such as T2M-GPT [25], which are commonly used for text-to-motion synthesis.

The generated amputee meshes are subsequently overlaid onto various background images to improve environmental diversity. To achieve this, we incorporated both indoor lab data and in-the-wild pose images from diverse environments. We further utilized LaMa [17] for object removal, where human regions detected by a segmentation model [10] are masked and removed before projecting the amputee mesh onto the cleared area.

For removing existing humans from the background, we use the Segment Anything Model (SAM) [10] and LaMa [17] in sequence: SAM is used to segment the human region, and LaMa inpaints the masked area. Since SAM performs open-vocabulary segmentation based on various prompts (e.g., point clicks, boxes, masks), we prepend a pre-processing module that detects humans via bounding boxes and places point prompts at the top, bottom, left, and right edges of the detected box. This refinement ensures accurate segmentation of only the intended human region and enables clean removal before mesh overlay.

To improve person detection coverage in the input data, we replace Detectron2 [23] with a stronger detector: YOLOv11 [7], fine-tuned on the CrowdHuman dataset [15] for human-only detection. Compared to widely used detectors [23], this model detects more individuals across crowded scenes, thereby maximizing the usability of input data and boosting the diversity of generated amputee images.

Finally, we apply a post-processing filter to remove failed or low-quality syntheses. Specifically, we propose a quality checker module that evaluates the realism and visual fidelity of the background after human removal. First, we compute the Structural Similarity Index (SSIM) [21] between the original input image and the LaMa-inpainted background. Images with SSIM scores below 0.5 are excluded from the dataset, as they typically exhibit blurry or overly smoothed artifacts.

Second, we verify whether the human has been successfully removed by applying a 2D human pose detector to the inpainted image. Only images in which no human keypoints are detected are retained. This two-stage filtering process ensures that the background remains both visually natural and free of human remnants, thereby enhancing the overall quality of the synthesized dataset.

B. Representation of Amputation under the SMPL

In Fig. B, our study demonstrates that when a specific joint index in the SMPL skeleton hierarchy, such as the R_Knee joint, is set to a zero matrix and passed through the SMPL Layer, all vertex positions associated with its child joints converge toward the R_Knee. This behavior effectively illustrates the hierarchical influence of the parent joint on its corresponding child joints within the SMPL model structure.

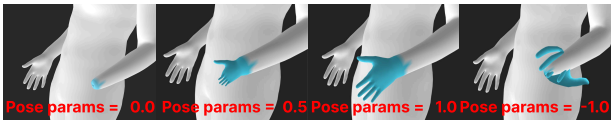

Figure C. **Effect of varying the pose parameters on the left wrist from** -1.0 **to** 1.0**.**

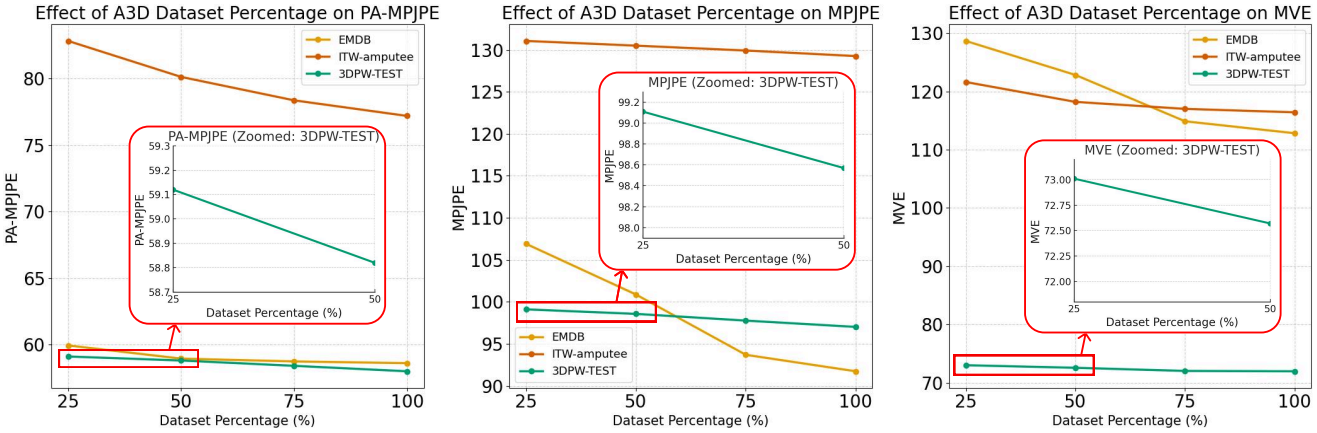

Figure D. **Impact of A3D Data Proportion on Performance.** Comparison of model performance when varying the ratio of A3D amputee data within each training batch, evaluated on in-the-wild datasets.

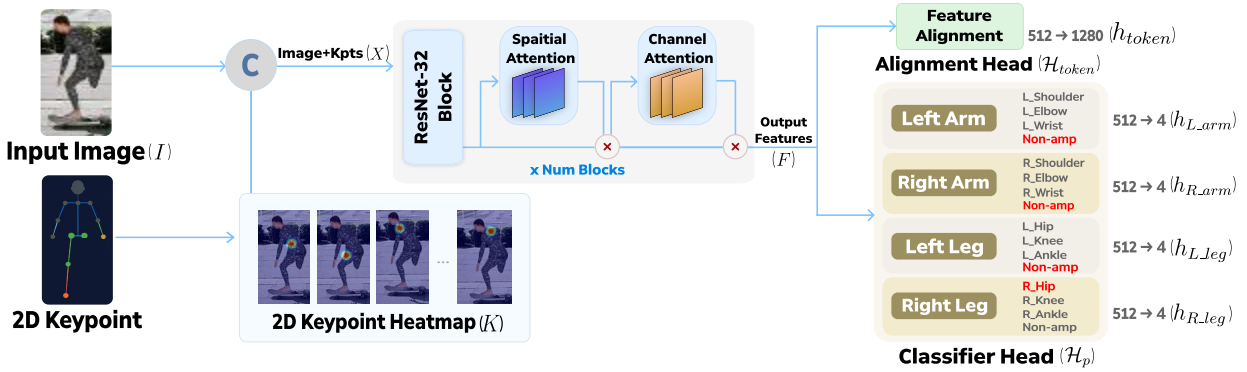

Figure E. **Overview of the Body Part Amputation Classifier (BPAC-Net).** The input image I and 2D keypoints K (converted to heatmaps) are concatenated and processed through a ResNet-32 [4] enhanced with Convolutional Block Attention Module (CBAM) [22], which applies spatial and channel attention. The extracted features F are fed into a feature alignment head to produce h_{token}, which is later used in the Transformer Decoder [19] via cross-attention. Four classifier heads $\mathcal{H}_p \in \{\mathcal{H}_{L_{arm}}, \mathcal{H}_{R_{arm}}, \mathcal{H}_{L_{leg}}, \mathcal{H}_{R_{leg}}\}$, predict amputation status for each corresponding body part.

C. Visualization with Varying Left Wrist Pose Parameters

As shown in Fig. C, applying a zero pose to the left wrist results in a clean amputation effect, with no distortion in surrounding body parts. In contrast, non-zero values cause visible artifacts or unnatural deformations near the joint. This empirically supports our choice of using zero pose parameters as a reliable proxy for amputated regions while maintaining local shape integrity.

D. A3D Quality

Dataset	A3D(MPII [1])	A3D(MSCOCO [11])	A3D(H3.6M [6])	Avg.
LPIPS [26],↓	0.0735	0.0421	0.16186	0.155

Table A. **LPIPS scores for A3D across datasets.**

As shown in Tab. A, our A3D dataset, synthesized as described in Sec. A, exhibits high perceptual realism with an average LPIPS [26] of 0.155. Since lower LPIPS corresponds to a smaller perceptual distance between image pairs, this supports that our synthesis pipeline faithfully preserves background, lighting, and texture details at a level comparable to real images. Furthermore, the ablation results in Fig. D show that simply augmenting the training set with A3D consistently reduces PA-MPJPE and MVE on various in-the-wild benchmarks, including 3DPW [20] and ITW-amputee. This suggests that A3D sufficiently incorporates outdoor visual factors (such as lighting variation and background complexity) commonly encountered in in-the-wild environments, thereby enabling the model to improve its generalization when trained with it. The high perceptual quality of A3D thus plays a critical role in boosting human mesh recovery performance on natural images even without real amputee data.

	Full Fine tuning	Partially Fine tuning	Cross Attention	Use Classifier	EMDB [9]			3DPW [20]			A3D			ITW-amputee		
					MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓	PVE↓	MPJPE↓	PA-MPJPE↓	PVE↓	MPJPE↓	PA-MPJPE↓
(a)	✓				114.53	94.37	58.68	97.16	72.68	46.00	74.98	74.45	**48.80**	121.62	133.73	77.94
		✓			112.99	92.05	58.84	95.73	**71.33**	47.72	73.98	74.30	49.71	122.34	132.95	81.23
	✓		✓		116.78	95.84	**57.29**	98.66	72.01	44.95	75.99	73.29	49.54	126.46	139.05	83.88
		✓	✓		**112.83**	**91.74**	58.62	**95.26**	71.77	**44.94**	**73.42**	**73.19**	49.42	**116.42**	**129.25**	**77.18**
(b)	✓			✓	**113.42**	97.33	59.87	97.98	73.68	47.76	89.34	89.96	69.87	143.49	154.83	85.17
		✓		✓	113.93	94.12	58.87	98.51	73.88	48.90	89.12	88.12	68.17	145.01	153.17	90.12
	✓		✓	✓	115.65	97.98	59.08	99.34	72.23	45.16	88.51	88.32	68.19	148.28	157.51	93.11
		✓	✓	✓	114.52	**93.73**	**58.01**	**97.02**	**71.97**	**44.98**	**87.11**	**87.91**	**68.01**	**139.64**	**143.74**	**84.91**

Table B. **Comparison of AJAHR Module Ablation Studies.** (a) Partially Fine-Tuning method freezes the AJAHR parameters while selectively updating a limited set of trainable parameters within each module. In contrast, Full Fine-Tuning updates all parameters of AJAHR during training. (b) After BPAC-Net infers the body part status, the corresponding label is used to force the inferred SMPL body pose parameters to zero. The evaluation is then conducted using 3D keypoints obtained from the reconstructed mesh.

E. Effect of A3D Proportion on In-the-Wild Performance

In Fig. D, the effect of varying the proportion of the amputee dataset A3D within the training batch is examined under in-the-wild evaluation settings. The ratio of A3D was gradually increased from 25% to 100% of each training batch, and the model was evaluated at each stage. As the proportion of A3D increased, consistent improvements were observed across all evaluation metrics, including MPJPE, PA-MPJPE, and MVE.

The most notable improvement in PA-MPJPE was observed on the ITW-amputee dataset, whereas EMDB [9] exhibited the most significant gains in MPJPE and MVE. These results suggest that the model effectively learns from the A3D dataset and benefits from increased exposure, resulting in enhanced reconstruction performance for both amputee and non-amputee subjects.

F. AJAHR Architecture Implementation Detail

AJAHR architecture utilizes ViTPose [24] as the backbone network to embed input images, and adopts a Transformer decoder [19] following HMR2.0 [8]. Among the output tokens from the Transformer decoder, the global orientation, body shape, and camera translation are each regressed through separate linear layers. For the body pose, in order to match the distribution of AJAHR-Tokenizer, the 1024-dimensional features are passed through six sequential blocks, each composed of two multilayer perceptrons (MLPs) and a GELU [5] activation function. We adopt a partially fine-tuned strategy, where only the last four blocks of the ViTPose [24] backbone, the patch embedding layer, the pose embedding layer, and the final two blocks of the Transformer decoder are updated during training.

AJAHR is trained in parallel on two NVIDIA A100 GPUs using the AdamW [13] optimizer, with a batch size of 64, a learning rate of $5e^{-6}$, and a weight decay of $1e^{-4}$. To ensure balanced learning, we sample the amputee and non-amputee datasets with equal probability (0.5 each). Training is conducted for a total of 150,000 iterations.

G. BPAC-Net Architecture Implementation Detail

The architecture of Body Part Amputation Classifier(BPAC-Net) is presented in Fig. E, where the input and output feature dimensions of each learnable block are indicated below the respective blocks. BPAC-Net takes batch images along with their corresponding keypoint information as input, which are transformed into 2D keypoint heatmaps before being fed into the ResNet-32 [4] backbone. ResNet-32 consists of 16 Basic Blocks, with each block enhanced by a Convolutional Block Attention Module (CBAM) [22] at its endpoint to perform spatial and channel attention. The final feature vector output from the ResNet-32+CBAM module is 512-dimensional, which is then processed by dedicated classifier heads. Each head predicts one of three amputation types or the non-amputated state, resulting in a 4-dimensional output vector for classification. During inference, we replace the Ground Truth (GT) keypoints with 2D keypoints predicted from the image using the ViTPose [24].

H. AJAHR-Tokenizer Implementation Detail

The AJAHR-Tokenizer architecture is inspired by TokenHMR [3]. Both the encoder and decoder consist of four 1D convolutional layers and a single ResNet [4] block. It includes a codebook of size 256×2048 and 320 pose tokens. This configuration was selected based on the lowest reconstruction errors in MPJPE and MVE metrics across the AMASS [14], MOYO [18], and A3D datasets. Training is conducted for 200,000 iterations with a batch size of 256, a learning rate of 2×10^{-4}, a gamma value of 5×10^{-2}, and a weight decay of 1×10^{-5}. To avoid data imbalance, amputee and non-amputee samples are drawn with equal probability (0.5 each) during training.

Method		A3D		AMASS [14]		MOYO [18]	
		MPJPE↓	MVE↓	MPJPE↓	MVE↓	MPJPE↓	MVE↓
CodeBook (320 tokens)	128x2048	1.72	8.05	2.24	8.62	7.49	16.50
	256x1024	1.92	8.50	2.50	8.80	7.15	16.18
Tokens (Codebook: 256x2048)	20	5.51	12.60	7.62	15.00	23.10	38.87
	40	3.49	9.76	4.51	10.67	14.50	26.04
	80	2.39	8.56	3.07	8.97	10.05	19.44
	160	2.07	8.38	2.74	8.67	6.68	14.60
	640	2.59	9.03	**2.60**	9.03	8.02	16.60
Ours (256x2048, 320 tokens)		**1.56**	**8.01**	**1.90**	**8.08**	**5.52**	**13.47**

Table C. **Ablation Studies on AJAHR Tokenizer Configuration.** The MOYO [18] validation dataset was used for evaluation. In the codebook size comparison experiment, the total number of tokens was set to 320. In the token count comparison, the codebook size was fixed to 256×2048.

I. Ablation Study of AJAHR Training

Tab. B compares the performance of the AJAHR model across different training strategies. Here, **Full Fine-Tuning** refers to updating all model parameters except for the frozen AJAHR-Tokenizer, while **Partially Fine-Tuning** follows the methodology outlined in Sec. F, where only a subset of parameters is updated. In all experiments, the AJAHR-Tokenizer remains pre-trained and frozen, not updated during training.

In Tab. B(a), the evaluations utilize Ground Truth (GT) labels of amputation regions, removing the corresponding body parts from the predicted mesh prior to assessment. On the non-amputee dataset, performance differences between Partially Fine-Tuning and Full Fine-Tuning were minimal. However, on the EMDB [9], Full Fine-Tuning demonstrated superior performance. Notably, the Partially Fine-Tuning approach combined with cross-attention achieved the best performance on the ITW-amputee dataset. This indicates that the partially fine-tuning strategy enabled by cross-attention effectively enhances mesh reconstruction performance irrespective of amputation status.

Meanwhile, Tab. B(b) extends the experiments by using amputation states predicted by BPAC-Net instead of GT labels. The predicted labels were used to modify SMPL pose parameters, and evaluations were conducted based on meshes reflecting these amputation states.

On amputee datasets (A3D, ITW-amputee), our method recorded the lowest errors in both MPJPE and PA-MPJPE metrics, demonstrating that BPAC-Net reliably identifies amputation sites and significantly improves reconstruction accuracy. Although some cases involved misclassification, incorrectly removing existing body parts or generating non-existent limbs, the overall improvement in performance was clearly evident. These results validate the proposed modular architecture and illustrate the practical capability of BPAC-Net to reliably predict amputation status in real-world scenarios. Furthermore, there was no noticeable performance degradation on non-amputee datasets after applying BPAC-Net, confirming its stability and reliability in general human mesh reconstruction tasks.

J. Ablation Study on AJAHR-Tokenizer Settings

As shown in Tab. C, performance improved consistently with increased codebook size, highlighting the importance of adequately large codebooks for representing diverse and complex human structures and poses. However, excessively increasing the number of tokens posed a risk of overfitting without further performance gains. Considering these results, we selected a final tokenizer configuration of a 256×2048 codebook with 320 tokens for training the AJAHR model, achieving optimal performance.

Acknowledgements. This research was partly supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project Name : Development of high-freedom large-scale user interaction technology using multiple projection spaces to overcome low-light lighting environments, Project Number : RS-2023-00222280, Contribution Rate : 50%) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [IITP-2025-RS-2024-00437102, ITRC(Information Technology Research Center) support program; RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 2, 3

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[3] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[5] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 1, 2, 3

[7] Glenn Jocher, Ji Qiu, and Ayush Chaurasia. Ultralytics YOLO (Version 8.0.0). https://github.com/

ultralytics/ultralytics, 2023. Computer software. 2

[8] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 4

[9] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 4, 5

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 2

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1, 2

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[14] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 4, 5

[15] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2

[16] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 906–915, 2024. 1

[17] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1, 2

[18] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4713–4725, 2023. 4, 5

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[20] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 3, 4

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 4

[23] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[24] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 4

[25] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3