

# AJAHR: Amputated Joint Aware 3D Human Mesh Recovery

Hyunjin Cho<sup>1,3,\*</sup> Giyun Choi<sup>1,\*</sup> Jongwon Choi<sup>1,2,†</sup>

<sup>1</sup>Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Korea <sup>2</sup>Dept. of Artificial Intelligence, Chung-Ang University, Korea

<sup>3</sup>Korea Institute of Industrial Technology (KITECH), Korea

{jincho, cky}@vilab.cau.ac.kr, choijw@cau.ac.kr

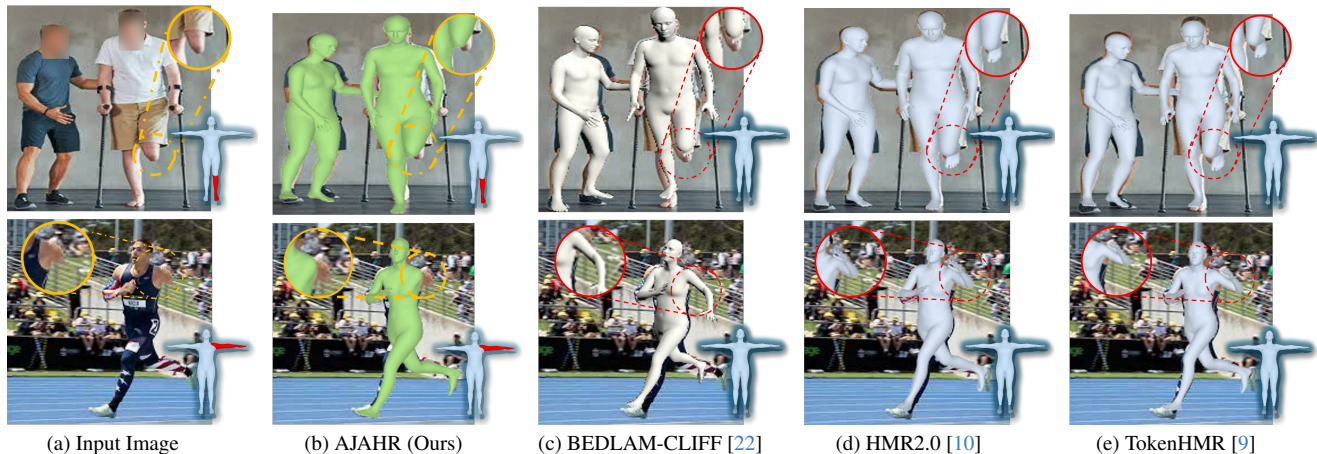


Figure 1. **Examples of Human Mesh Recovery for Amputee and Non-Amputee Individuals.** Column (a) shows input images: the top row includes a non-amputee (left) and an amputee (right), while the bottom row shows an amputee. Our method AJAHR (b) accurately handles both cases, whereas (c), (d), and (e) often misinterpret amputated limbs as intact and infer implausible poses in missing regions.

## Abstract

Existing human mesh recovery methods assume a standard human body structure, overlooking diverse anatomical conditions such as limb loss. This assumption introduces bias when applied to individuals with amputations—a limitation further exacerbated by the scarcity of suitable datasets. To address this gap, we propose *Amputated Joint Aware 3D Human Mesh Recovery (AJAHR)*, which is an adaptive pose estimation framework that improves mesh reconstruction for individuals with limb loss. Our model integrates a body-part amputation classifier, jointly trained with the mesh recovery network, to detect potential amputations. We also introduce *Amputee 3D (A3D)*, which is a synthetic dataset offering a wide range of amputee poses for robust training. While maintaining competitive performance on non-amputees, our approach achieves state-of-the-art results for amputated individuals. Additional materials can be found at: [https://chojinie.github.io/project\\_AJAHR/](https://chojinie.github.io/project_AJAHR/)

## 1. Introduction

Human pose information is essential for understanding human behavior and enabling effective human-computer interaction. Its importance is evident in applications such as sports [4, 40], AR/VR [2] and surveillance [7]. In computer vision, Human Mesh Recovery (HMR) from a single RGB image offers a cost-effective solution without requiring additional sensors. Although this approach is simple, it has demonstrated strong performance, driving continued research in the field. Our study follows this direction.

Despite recent advancements, a significant limitation remains: most existing methods are trained on datasets composed exclusively of non-amputees, implicitly assuming a standard human body structure. As a result, they tend to produce biased pose estimations when applied to anatomically diverse conditions—such as limb loss—often hallucinating unrealistic poses or body shapes for the missing limbs, rather than accurately reflecting the amputated regions. This highlights the need for more inclusive HMR models that generalize across a wider range of body types.

Prior work has explored model inclusivity by examining

\* Equal contribution. † Corresponding author.

the lack of representation in human-related motion datasets and addressing associated bias in pose estimation. Olugbade *et al.* [27] highlighted that among 704 publicly available datasets, none include individuals with disabilities performing activities such as sports or daily tasks. In parallel, studies such as Zhou *et al.* [48] and WheelPose [14] have identified biases in 2D human pose estimation models trained on general human movement datasets.

Inspired by recent findings and limitations, we present AJAHR—an adaptive framework improving mesh recovery in individuals with limb amputations—an area largely overlooked in prior HMR research. Addressing this scenario presents unique challenges: training data with real amputees is extremely scarce, making it difficult to learn body configurations that deviate from standard anatomy. As noted by [27], collecting such data—whether in controlled studio environments or in the wild—raises significant ethical and logistical concerns, including safety risks, privacy issues, and high acquisition costs. Moreover, amputation-induced joint absence can be easily confused with occlusions. In the latter case, the joint exists but is merely hidden from view—posing ambiguity for models relying solely on image cues. AJAHR integrates a body-part amputation classifier that is jointly trained with the mesh recovery network, allowing the model to distinguish between amputees and non-amputees and produce pose estimates tailored to each body condition.

As shown in Fig. 1, existing methods often hallucinate unrealistic body parts or fail to represent amputated poses accurately, revealing their limited generalization capability. To the best of our knowledge, no prior dataset or framework has been specifically designed to address this scenario. To fill this gap, we construct Amputee 3D (A3D), a synthetic dataset generated through a controlled data pipeline that offers diverse amputee pose samples for training. Additionally, we compile ITW-amputee, a real-world evaluation set consisting of in-the-wild images of individuals with limb loss collected from online sources. This dataset serves as a benchmark to assess generalization performance on real-world amputee cases, making our contributions a crucial step toward inclusive HMR.

Our contributions can be summarized as follows:

1. We introduce and address the first-ever human mesh recovery problem for amputated individuals.
2. We propose AJAHR, an adaptive HMR model that detects amputated individuals while ensuring stable pose estimation for both amputees and non-amputees.
3. We construct a new dataset, A3D, for human mesh recovery of amputated individuals and introduce a method to synthesize datasets tailored for this task.
4. Our approach preserves competitive performance on non-amputees while achieving state-of-the-art results on amputee datasets.

## 2. Related Work

**Inclusiveness in Human Mesh Recovery.** Human mesh recovery research has primarily focused on improving pose accuracy in occluded conditions [20, 45, 47] or leveraging motion capture datasets like AMASS [25] and MOYO [33] to enhance pose priors across various postures [6, 9, 16]. While these methods improve accuracy, they largely focus on individuals with typical anatomical structures, overlooking anatomical diversity. To improve inclusivity, some studies have specifically explored individuals with physical disabilities [14, 48]. For instance, WheelPose [14] introduced a synthesis pipeline for wheelchair users, while Zhou *et al.* [48] reconstructed prosthetic limbs as intact limbs to enable robust pose estimation. However, human pose estimation or human mesh recovery for individuals with limb amputations remains unexplored. To address this gap, we introduce a synthesis pipeline for amputee pose data and analyze model adaptability to missing body parts.

**Monocular 3D Human Mesh Recovery.** 3D human mesh recovery from a single RGB image involves extracting visual features to reconstruct a parametric human body. Existing methods are broadly categorized into regression- and optimization-based approaches. Regression-based methods [10, 16, 20, 22, 32, 34] directly predict body model parameters [24, 28, 41] in a single forward pass, enabling real-time inference. Optimization-based methods [6, 21, 29, 30] estimate these parameters by fitting the SMPL model to 2D cues such as keypoints and silhouettes, and iteratively refine predictions using additional image information. Despite strong performance, most methods are trained on datasets without disabled individuals, resulting in pose priors that generalize poorly to amputees. To address this, we adopt TokenHMR [9] as our baseline, which reframes pose estimation as a token classification task to mitigate bias and incorporate structured priors. Building on this framework, we integrate pose priors for individuals with limb amputations, improving prediction accuracy for missing body regions.

## 3. Proposed Method

### 3.1. Preliminaries

**Body Model.** The SMPL [24] is a low-dimensional, differentiable parametric body model that represents the human body. The model takes as an input the pose parameters  $\theta \in \mathbb{R}^{24 \times 3 \times 3}$  and shape parameters  $\beta \in \mathbb{R}^{10}$ . As an output, the model generates the human body mesh,  $\mathcal{M}$ , which consists of  $V \in \mathbb{R}^{N \times 3}$ , where  $N=6890$  represents the number of vertices. The 3D joints,  $J_{3D}$ , are obtained by combining the pre-trained joint regressor with the vertices.

**Representation of Amputation under the SMPL.** The SMPL model has a kinematic tree structure, where 24 joints are organized in a parent-child relationship. As it is pre-

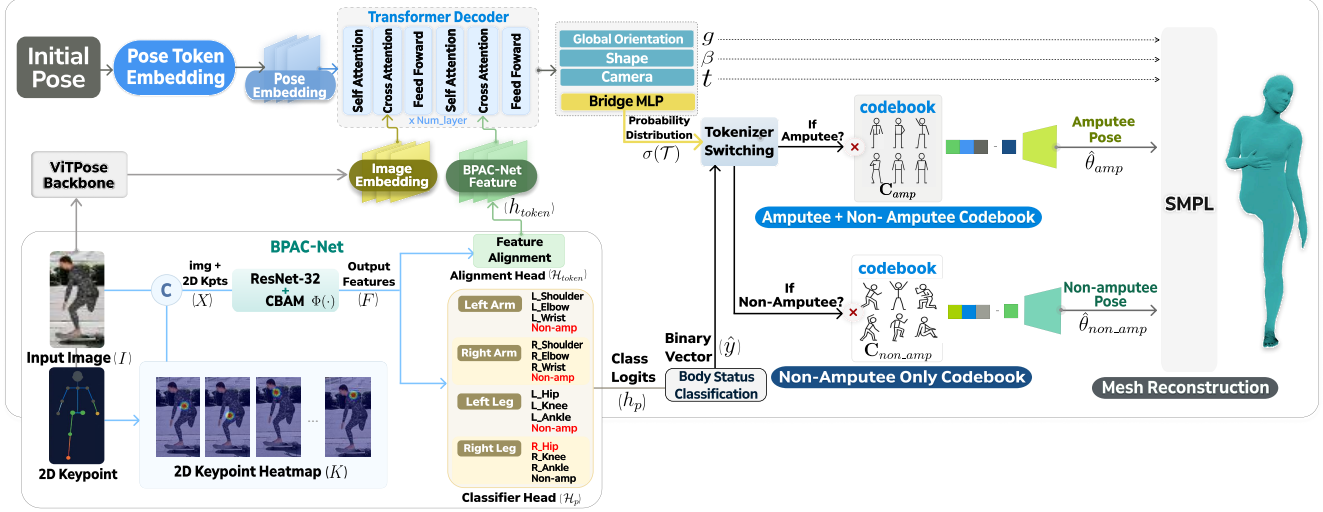


Figure 2. **Overview of AJAHR Architecture.** The proposed model takes typical human pose datasets and the A3D dataset—a synthetic dataset of amputated individuals (see Sec. 3.2)—as input. It employs BPAC-Net, a body-part amputation classifier (see Sec. 3.3.1), to detect limb absence and uses this information to guide mesh recovery for both amputees and non-amputees. Based on the predicted body-part status, the model connects to different pre-trained tokenizers (see Sec. 3.3.2), which remain frozen during training while the rest of the model is optimized accordingly.

Dataset	Year	Type	Annot.	SMPL GT	Amputee	# Images
WheelPose [14]	2024	Image	2D	×	×	-
BEDLAM [5]*	2023	Video	3D+2D	✓	×	18M
EMDB [18]*	2023	Video	3D+2D	✓	×	105K
3DPW [37]*	2018	Video	3D+2D	✓	×	53K
Human3.6M [15]*	2015	Video	3D+2D	×	×	3.6M
MSCOCO [23]	2014	Image	2D	×	×	200K
MPII [3]	2014	Image	2D	×	×	25K
<b>Ours (A3D)</b>	2025	Image	3D+2D	✓	✓	1.0M

Table 1. **Comparison of datasets.** \* indicates that the number of images refers to frames extracted from videos.

trained and not directly trainable, we leverage this structure to represent amputations without modifying the model. Our method encodes amputations by setting the pose parameter  $\theta$  of the amputated parent joint and all its descendants to a zero matrix. When these parameters are passed into the SMPL model, the corresponding vertices collapse into a single location near the amputated joint, effectively simulating limb absence. The generated vertices are then multiplied by a pre-trained joint regressor to obtain 3D joint positions, resulting in the child joints of the amputated region shifting toward the amputated parent due to the hierarchical structure of the body model. For details, please refer to Supplementary Sec. B.

### 3.2. Amputee Dataset Synthesis

**Overview.** Training a pose estimator for individuals with limb amputations requires annotated amputee images. However, collecting such data is costly and raises concerns regarding safety, accessibility, and diversity. In addition,

state-of-the-art generative models (e.g., GPT-4 [1], ControlNet [44]) still fall short in producing sufficient high-quality amputee images for our purposes. To address these limitations, we synthesize amputee data through a multi-stage pipeline that generates diverse amputee instances with various ethnic appearances and realistic backgrounds. This pipeline includes an index selection module to simulate amputated body parts in the SMPL representation, and a controller module that handles texture assignment for gender and ethnic diversity. See Supplementary Sec. A for architectural design and implementation details.

**Synthesizing Amputee Representations.** We leverage off-the-shelf models designed for recovering human mesh from a single RGB image using the SMPL body model to create synthetic representations of individuals with limb amputations, where amputated parts are represented using a zero pose. Additionally, we enhance visual diversity by incorporating skin and clothing textures from BEDLAM [5] assets, which reflect a range of human appearances and contribute to comprehensive representation.

**Generating Background Images.** To ensure generalization, we utilized both indoor and in-the-wild pose images from diverse environments. To obtain clean background images, we applied a two-stage removal process to benchmark datasets [3, 15, 23]: a segmentation model [19] was used to detect and mask human regions, which were then inpainted using LaMa [31], a model designed for object removal. This process generated approximately 3K backgrounds from MSCOCO [23] and 1K from MPII [3]. For Human3.6M [15], which consists of scenes captured from

four fixed camera viewpoints in a controlled multi-view environment, we extracted one background image per viewpoint, resulting in a total of four. Finally, we projected the synthesized amputee meshes onto these cleared backgrounds using weak perspective projection.

**A3D: Amputee 3D Pose Dataset.** Through our synthetic data pipeline, we generated over one million high-quality images simulating various amputation scenarios. These images reflect a diverse range of poses, clothing, and backgrounds, contributing to both generalization and robustness in training. Inspired by BEDLAM [5], we designed our dataset to reflect balanced demographic diversity: African, Asian, and Indian each account for 20% of the mannequin population, while Hispanic, Middle Eastern, Southeast Asian, and White each comprise 10%. To reflect a wide range of amputation types, we simulated limb loss across multiple body parts—including hand, forearm, full arm, ankle, knee, and full leg—based on poses commonly observed in benchmark datasets of non-amputee subjects. Alongside the images, we provide full annotations, including SMPL parameters, 3D and 2D joint coordinates, and detailed amputation region labels, all aligned with ground truth (GT). As shown in Tab. 1 shows that our dataset is sufficiently large and serves as the first dataset tailored for the amputee domain. During training, we leverage 2D keypoints, 3D keypoints, and SMPL parameters as supervision signals. For 2D keypoints, we follow standard practices by setting the coordinates of joints in non-visible regions—whether due to amputation or occlusion (i.e., the missing joint and its child joints)—to (0, 0), thereby excluding them from 2D supervision. In contrast, 3D keypoints and SMPL parameters corresponding to amputated regions are still used as supervision signals, allowing the model to learn representations for structurally absent but semantically defined joints. This distinction between modality-specific supervision handling for occlusion and amputation enables accurate learning based on the SMPL framework without requiring architectural modifications or additional retraining.

### 3.3. Architecture of AJAHR

**Overview.** As illustrated in Fig. 2, AJAHR adopts a Vision Transformer (ViT) [8] based architecture inspired by HMR2.0 [10] and TokenHMR [9]. Input images are encoded into embedding tokens via a ViT encoder and then refined through a Transformer decoder [36] featuring two cross-attention pathways. First, a predefined zero-pose parameter token attends to the image feature tokens to initialize the pose representation. Second, the classifier-generated token undergoes additional cross-attention within the decoder, allowing semantic cues from the classifier to guide the pose regression. Additionally, BPAC-Net classifies the amputation status of the body by taking the input image and its corresponding 2D keypoints as input. During training,

ground-truth 2D keypoints are used, while at inference time, predicted keypoints from the ViTPose [42] detector are employed.

The output tokens from the transformer decoder [36] are split into four branches, each corresponding to a major body region—left arm, right arm, left leg, and right leg. Each branch performs amputation classification by determining whether any child joints within its region, based on the SMPL kinematic tree, are amputated or non-amputated. In one branch, a bridge MLP transforms the decoder output to match the dimensionality of the codebook, producing logits for each codebook entry. Applying the softmax function yields a probability distribution that serves as soft weights over the pre-trained codebook, each entry representing a latent pose component. The pre-trained codebook consists of two types: one trained on both amputee and non-amputee datasets, and another trained only on non-amputee data. Based on the amputation status predicted by BPAC-Net, a tokenizer-switching strategy selects the appropriate codebook. The selected codebook is then aggregated via multiplication using the token distribution as weights, resulting in the final predicted pose parameter,  $\hat{\theta}$ . The other three branches independently regress the global rotation ( $g$ ), body shape parameters ( $\beta$ ), and camera translation ( $t$ ) through separate regression heads. Finally, the predicted parameters are passed to SMPL for mesh reconstruction.

We follow the TokenHMR [9] paradigm by pretraining the tokenizer separately. Specifically, the codebook and decoder are trained in advance, and then frozen during pose estimation training.

#### 3.3.1. Body Part Amputation Classifier (BPAC-Net)

The proposed BPAC-Net serves three key roles: (1) **Loss Adjustment:** Enhancing learning on amputee data by increasing loss for amputated regions. (2) **Implicit Learning Assistance:** BPAC-Net features undergo cross-attention with the transformer decoder [36] to improve pose estimation for amputated parts. (3) **Visualization:** Enforcing zero values in SMPL pose parameters for amputated joints to ensure accurate visual representation.

To address the ambiguity of missing or occluded limbs in RGB images, we incorporate 2D keypoint heatmaps as additional visual cues. For each amputee and non-amputee image, we incorporate its corresponding keypoint heatmap as an additional visual cue by concatenating the RGB input  $I \in \mathbb{R}^{H \times W \times 3}$  with the keypoints  $K \in \mathbb{R}^{H \times W \times J}$  along the channel dimension to form  $X = \text{concat}(I, K)$ . These combined inputs are fed into  $\Phi(\cdot)$ , which integrates ResNet-32 [13] and CBAM [38] to extract spatial and semantical feature maps  $F = \Phi(X) \in \mathbb{R}^{h \times w \times c}$ .

The extracted features are passed to four parallel classification heads  $\mathcal{H}_p$  and one feature alignment head  $\mathcal{H}_{token}$ . Four classification heads, denoted as  $\mathcal{H}_p \in \{\mathcal{H}_{Larm}, \mathcal{H}_{Rarm}, \mathcal{H}_{Lleg}, \mathcal{H}_{Rleg}\}$ , are responsible for pre-

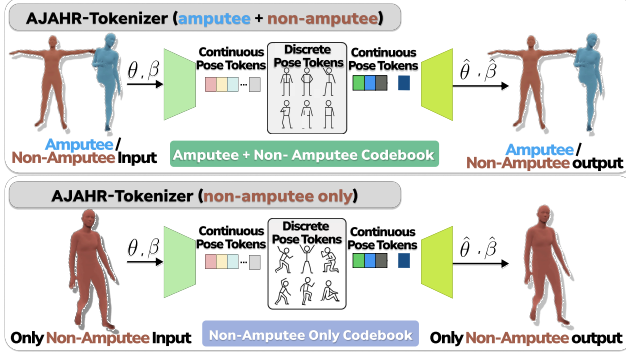


Figure 3. **Tokenizing for diverse pose prior.** The input  $\theta$  consists of SMPL pose parameters from amputees in the A3D dataset and non-amputees in AMASS [25] and MOYO [33]. This allows the tokenizer to encode amputee pose information into the codebook, enabling the trained decoder to incorporate pose priors for both amputees and non-amputees, thereby improving its ability to reconstruct diverse human poses.

dicting the amputation status of each limb. For each body part  $p \in \{L_{arm}, R_{arm}, L_{leg}, R_{leg}\}$ , the corresponding classification head outputs a part-specific logit feature vector  $h_p = \mathcal{H}_p(F) \in \mathbb{R}^4$ . Here,  $h_p \in \mathbb{R}^4$  represents the logits for four classes corresponding to each limb, as shown in the head components of BPAC-Net in Fig. 2. These predictions are then compared to the ground-truth amputation or non-amputation labels  $lb$  using a cross-entropy loss function.

To enable the tokenizer switching strategy, we determine whether any of the predicted classes correspond to an amputated state. Each classification head  $\mathcal{H}_p$  outputs a logit vector  $h_p \in \mathbb{R}^4$ , where class 0 corresponds to a non-amputated limb, and classes 1, 2, and 3 represent amputation types specific to the body part that the head is responsible for. A binary decision  $\hat{y}_p \in \{0, 1\}$  is obtained as:

$$\hat{y}_p = \begin{cases} 0 & \text{if } \arg \max(h_p) = 0, \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

The values obtained for the four parts are concatenated to form the 4-dimensional vector  $\hat{y} = [\hat{y}_{L_{arm}}, \hat{y}_{R_{arm}}, \hat{y}_{L_{leg}}, \hat{y}_{R_{leg}}]^T$ , representing the predicted amputation status for each limb.

The feature alignment head, denoted as  $\mathcal{H}_{\text{token}}$ , produces a global feature vector  $\mathcal{H}_{\text{token}}(F) \in \mathbb{R}^{1280}$ , which is directly used as the cross-attention mechanism of the transformer decoder [36]. The BPAC-Net classification loss is defined as a sum over four limb-specific predictions:

$$\mathcal{L}_{cls} = \sum_{p \in \{L_{arm}, R_{arm}, L_{leg}, R_{leg}\}} CE(h_p, lb), \quad (2)$$

where  $CE$  denotes the cross-entropy loss function. Details of the amputation labels can be found in the BPAC-Net head outputs in Fig. 2 and the Supplementary Sec. G.

### 3.3.2. AJAHR-Tokenizer

**Tokenizer Switching Strategy.** In this study, we present two versions of the tokenizer as depicted in Fig. 3: one trained on a combined dataset of amputee and non-amputee poses, and the other trained exclusively on non-amputee poses. Once each tokenizer has been trained, the codebook  $\mathbf{C}_{amp}$ , which can reconstruct both amputee and non-amputee poses, and the codebook  $\mathbf{C}_{non\_amp}$ , specialized for non-amputee poses, along with their corresponding decoders, are kept frozen. Then, following the tokenizer switching strategy, the predicted binary indicator  $\hat{y}$  from BPAC-Net is applied to the output tokens of the transformer decoder to infer the pose parameters as follows:

$$\hat{\theta} = \begin{cases} \sigma(\mathcal{T}) \times \mathbf{C}_{amp} & \text{if } \|\hat{y}\|_1 > 0, \\ \sigma(\mathcal{T}) \times \mathbf{C}_{non\_amp} & \text{otherwise,} \end{cases} \quad (3)$$

where  $\hat{\theta}$  is composed of  $\hat{\theta}_{amp}$ , obtained from  $\mathbf{C}_{amp}$ , and  $\hat{\theta}_{non\_amp}$ , obtained from  $\mathbf{C}_{non\_amp}$ , and is used for final mesh reconstruction.

**Training Objective for AJAHR-Tokenizer.** As described in Sec. 3.1, the AJAHR-Tokenizer predicts full-body pose parameters  $\theta$  to reconstruct both amputee and non-amputee poses. Directly assigning zero values to amputated joints in 3D rotation representations can induce numerical instability (e.g., NaNs). To mitigate this, original pose parameters are preserved during training, and zero masking is applied in post-processing according to the predicted body-part status  $\hat{y}_p$  from BPAC-Net, where  $\hat{y}_p \in \{0, 1\}^4$  is a binary indicator vector over 4 predefined body parts (1 denotes an amputated part).

The AJAHR-Tokenizer is trained on a mix of large-scale non-amputee pose data (AMASS [25], MOYO [33]) and amputee-specific SMPL poses from the A3D dataset. To ensure stable 3D rotation representation, each joint is encoded using the continuous 6D representation from [49].

The process encodes the pose parameters via an encoder  $\mathbf{E}$  into latent features  $Z = \mathbf{E}(\theta) = [z_1, z_2, \dots, z_S]$ , where  $z_i \in \mathbb{R}^d$  and  $S$  is the number of tokens. A learnable codebook  $\mathbf{C} = \{c_m\}_{m=1}^M$  is maintained, with each code  $c_m \in \mathbb{R}^d$  and  $d$  denoting the dimensionality of the code vectors. Each latent vector  $z_i$  is quantized to its nearest codebook entry:

$$\tilde{z}_i = \arg \min_{c_m \in \mathbf{C}} \|z_i - c_m\|_2, \quad (4)$$

producing the quantized set  $\tilde{Z} = [\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_S]$ .

Following VQ-VAE [35] and TokenHMR [9], the total loss combines three components: mixed reconstruction loss  $\mathcal{L}_{\text{mix}}$ , codebook (embedding) loss, and commitment loss.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mix}} \mathcal{L}_{\text{mix}} + \lambda_{\text{cb}} \|\text{sg}[Z] - \tilde{Z}\|_2^2 + \lambda_{\text{com}} \|Z - \text{sg}[\tilde{Z}]\|_2^2, \quad (5)$$

Method	A3D			ITW-amputee		
	MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓
HMR2.0 [10]	89.35	96.75	86.14	<b>110.33</b>	154.43	121.83
BEDLAM-CLIFF [5, 22]	83.38	88.12	56.45	128.09	150.12	117.74
TokenHMR [9]	76.01	74.70	49.94	136.52	146.12	91.00
AJAHR (Ours)	<b>73.42</b>	<b>73.19</b>	<b>49.42</b>	116.42	<b>129.25</b>	<b>77.18</b>

Table 2. Results on Amputee Data.

Method	EMDB [18]			3DPW [37]		
	MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓
HMR2.0 [10]	141.41	117.66	75.89	95.29	81.64	53.95
BEDLAM-CLIFF [5, 22]	129.00	97.88	62.40	99.32	76.45	51.21
TokenHMR [9]	113.26	93.77	58.98	<b>90.23</b>	72.87	47.17
AJAHR (Ours)	<b>112.83</b>	<b>91.74</b>	<b>58.62</b>	95.26	<b>71.77</b>	<b>44.94</b>

Table 3. Results on Non-Amputee Data.

Experiments	Use Classifier	EMDB [18]			3DPW [37]			A3D			ITW-amputee		
		MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓
(a) Noise Ratio : 100%	✓	117.71	96.22	60.97	99.03	75.64	49.31	91.30	91.21	71.31	144.08	147.41	88.08
Noise Ratio : 75%	✓	115.77	94.78	59.31	97.91	73.31	46.88	89.12	89.32	69.74	142.21	145.99	86.51
Noise Ratio : 50%	✓	115.31	94.12	59.22	97.43	72.77	45.87	88.76	88.98	69.32	141.78	145.01	86.17
Noise Ratio : 25%	✓	114.82	94.03	58.88	97.31	72.08	45.08	87.98	88.37	68.71	140.09	144.24	85.21
(b) Image only	✓	131.81	109.98	74.21	113.71	87.09	59.54	105.88	103.12	85.44	152.21	154.55	92.71
Keypoint only	✓	118.21	96.12	61.71	100.87	74.87	46.91	90.12	89.21	70.77	141.64	146.21	87.88
(c) HMR2.0 [10]	✓	149.31	125.69	80.74	100.21	89.74	56.91	104.72	104.75	94.32	134.71	176.46	132.27
BEDLAM-CLIFF [5, 22]	✓	133.75	100.29	73.24	103.98	83.21	54.28	92.77	96.48	75.87	147.51	166.07	126.90
(d) 160 Tokens	✓	117.38	98.12	61.94	101.56	75.83	47.21	90.47	90.28	71.04	144.78	147.91	89.63
640 Tokens	✓	127.92	107.43	64.75	106.67	77.69	50.36	93.81	96.92	75.08	149.35	151.80	94.12
<b>Ours</b>	✓	<b>114.52</b>	<b>93.73</b>	<b>58.01</b>	<b>97.02</b>	<b>71.97</b>	<b>44.98</b>	<b>87.11</b>	<b>87.91</b>	<b>68.01</b>	<b>139.64</b>	<b>143.74</b>	<b>84.91</b>
(e) Amputation Only (Single)		115.70	93.75	59.08	96.32	72.76	45.92	74.71	74.51	49.93	118.09	131.12	78.08
Non Amputation Only (Single)		113.09	92.07	57.97	95.34	72.02	45.02	76.01	76.31	50.99	120.81	134.71	81.82
<b>Ours (Unified)</b>		<b>112.83</b>	<b>91.74</b>	<b>58.62</b>	<b>95.26</b>	<b>71.77</b>	<b>44.94</b>	<b>73.42</b>	<b>73.19</b>	<b>49.42</b>	<b>116.42</b>	<b>129.25</b>	<b>77.18</b>

Table 4. Ablation Experiments on the Components of BPAC-Net and AJAHR Tokenizer. We compare the performance across (a) ablation of BPAC-Net components, (b) evaluation of AJAHR with Gaussian noise-injected keypoints as input BPAC-Net, (c) joint training of baseline methods with BPAC-Net, (d) comparison of AJAHR performance with 160 and 640 tokens and (e) a comparison between single-tokenizer and tokenizer-switching strategies.

where  $\text{sg}[\cdot]$  is the stop-gradient operator. The mixed loss  $\mathcal{L}_{\text{mix}}$  measures  $\ell_2$  distance between predicted and ground-truth mesh vertices  $V$ , 3D joints  $J_{3D}$ , and pose parameters  $\theta$ . The loss weights are set to  $\lambda_{\text{mix}} = 100.0$ ,  $\lambda_{\text{cb}} = 1.0$ , and  $\lambda_{\text{com}} = 1.0$ . To prevent codebook collapse, we apply codebook reset and exponential moving average (EMA) updates as in prior work [9, 43]. Full training details are provided in Supplementary Sec. H.

### 3.3.3. AJAHR Losses

The overall loss of AJAHR combines SMPL-related regression terms with the amputation-aware classification loss from BPAC-Net. Let  $\theta, \beta$  and  $\hat{\theta}, \hat{\beta}$  be the ground-truth and predicted SMPL pose and shape parameters, respectively. Predicted 3D joint locations  $\hat{J}_{3D}$  are obtained via a pre-trained joint regressor from  $\hat{\theta}, \hat{\beta}$ , and projected to 2D as  $\hat{J}_{2D}$ , with ground-truth targets  $J_{3D}, J_{2D}$ . The pose loss  $\mathcal{L}_{\theta}(\theta, \hat{\theta})$  measures the  $\ell_2$  distance in a stable (e.g., 6D) rotation representation, and  $\mathcal{L}_{\beta}(\beta, \hat{\beta})$  is the  $\ell_2$  distance between shape coefficients. These are complemented by the 3D joint loss  $\mathcal{L}_{3D}(J_{3D}, \hat{J}_{3D})$ , 2D projection loss  $\mathcal{L}_{2D}(J_{2D}, \hat{J}_{2D})$ , and BPAC-Net’s classification loss  $\mathcal{L}_{\text{cls}}$ , all computed as  $\ell_2$  or cross-entropy losses. The overall loss  $\mathcal{L}_{\text{overall}}$  is defined as:

$$\begin{aligned} \mathcal{L}_{\text{overall}} = & \lambda_{\theta} \mathcal{L}_{\theta}(\theta, \hat{\theta}) + \lambda_{\beta} \mathcal{L}_{\beta}(\beta, \hat{\beta}) \\ & + \lambda_{2D} \mathcal{L}_{2D}(J_{2D}, \hat{J}_{2D}) + \lambda_{3D} \mathcal{L}_{3D}(J_{3D}, \hat{J}_{3D}) \\ & + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}. \end{aligned} \quad (6)$$

We set the hyperparameter as follows:  $\lambda_{\theta} = 10^{-3}$ ,  $\lambda_{\beta} = 5 \cdot 10^{-4}$ ,  $\lambda_{3D} = 5 \cdot 10^{-2}$ ,  $\lambda_{2D} = 10^{-2}$ , and  $\lambda_{\text{cls}} = 10^{-2}$ .

## 4. Experiments

**Training Datasets.** To train the AJAHR-Tokenizer, we used the training split of AMASS [25], MOYO [33], and the pose data from our A3D dataset. Following the TokenHMR [9] training protocol, the classifier and pose-estimation modules were jointly optimized. For AJAHR training, we employed BEDLAM [5], a synthetic dataset with accurate ground-truth 3D annotations, alongside standard datasets used in prior works [10, 20, 21], including Human3.6M [15], MPI-INF-3DHP [26], COCO [23], and MPII [3]. Additionally, similar to HMR2.0 [10], we incorporated in-the-wild 2D datasets, such as InstaVariety [17], AVA [12], and AI Challenger [39]. Furthermore, we included our A3D dataset to enhance training diversity. To ensure a fair comparison in our experiments, we finetuned open-source models, including CLIFF [22], HMR2.0b [10], and TokenHMR [9], using the same training data recipe and evaluated their performance against ours. Additional implementation details of AJAHR are provided in the Supplementary Sec. F.

**Evaluation Dataset and Metrics.** For tokenizer evaluation, we use Mean Vertex Error (MVE) and Mean Per Joint Position Error (MPJPE). Final pose accuracy is assessed using MVE, MPJPE, and Procrustes-Aligned MPJPE (PA-MPJPE), which calculates the average 3D joint distance (in millimeters) after aligning the estimated and ground-truth joint sets via Procrustes analysis [11]. Classifier performance is evaluated via confusion matrices, from which we derive metrics such as accuracy, precision, recall, and F1 score. Experiments are conducted on the test splits of

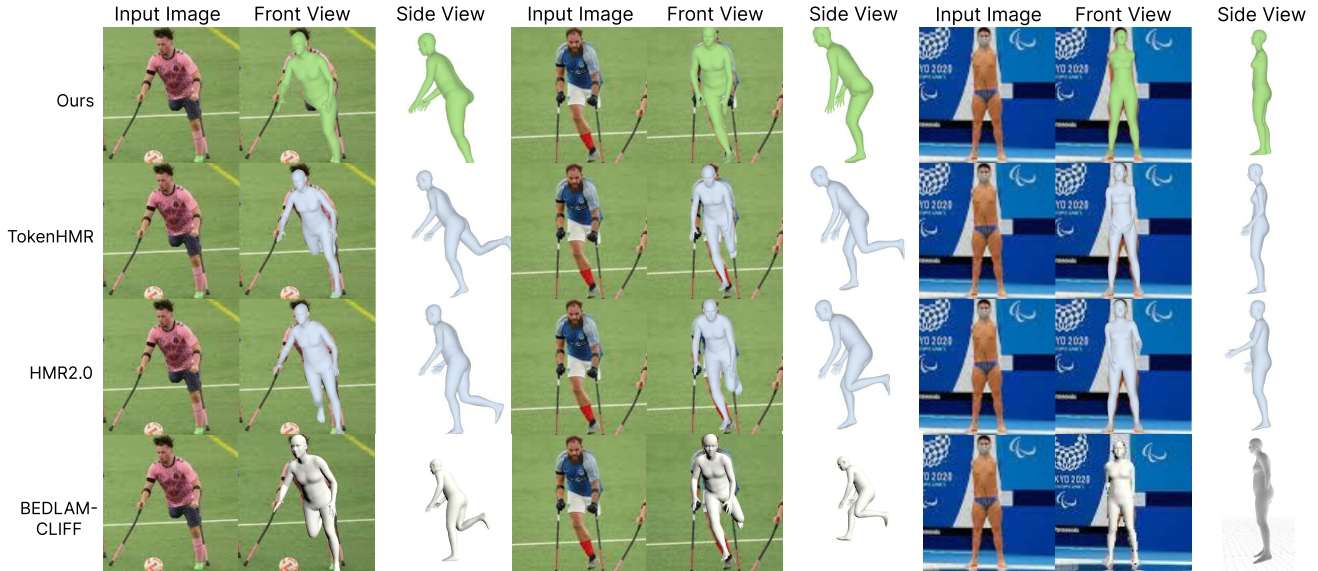


Figure 4. **A qualitative comparison with other Human Mesh Recovery methods trained on the A3D Dataset.** Unlike TokenHMR [9], HMR2.0 [10], and BEDLAM-CLIFF [5, 22], which do not employ BPAC-Net to identify amputated regions, AJAHR leverages BPAC-Net to explicitly represent these regions in the input images.

3DPW [37], EMDB [18], and our A3D dataset. To enable fair model comparisons and verify real-world applicability, we additionally evaluated the models on the In-the-Wild Amputee (ITW-amputee) dataset. This dataset was built by partially synthesizing web-crawled images using the A3D pipeline and manually annotating them. It includes both single- and multi-limb amputation cases captured in daily-life, rehabilitation, and sports scenes.

#### 4.1. Robust Generalization to Amputee and Non-Amputee Subjects

In Tab. 2, all evaluation models utilize Ground Truth (GT) labels for amputation regions to remove the corresponding body parts from the estimated mesh before evaluation. The integration of features extracted from BPAC-Net with cross attention results in overall superior performance compared to other human mesh recovery models. Notably, on the ITW-amputee dataset, the proposed method outperforms TokenHMR [9]. This indicates that training on the amputee dataset using the AJAHR-Tokenizer enhances the model’s ability to reconstruct body poses specific to individuals with limb differences. In AJAHR, human mesh recovery is performed in the same manner for both amputee and non-amputee subjects. In Tab. 3 presents the evaluation of human mesh recovery on non-amputee datasets (3DPW [37], EMDB [18]). The results demonstrate that the AJAHR model exhibits comparable performance in reconstructing human meshes for non-amputee subjects. This improvement can be attributed to the application of cross attention, which enhances the model’s ability to reconstruct

Method	A3D (amputation)				3DOH50K [46] (occlusion)			
	Accuracy↑	Precision↑	Recall↑	F1↑	Accuracy↑	Precision↑	Recall↑	F1↑
Ours	0.881	0.756	0.922	0.820	0.956	0.956	1.000	0.977

Table 5. **Amputation Classification Performance on A3D and 3DOH50K.** BPAC-Net accurately distinguishes amputation from occlusion, showing consistent classification performance across both synthetic and real-world occlusion scenarios.

body poses with greater precision, even for non-amputee data. Furthermore, when considered alongside the results in Tab. 2, these findings quantitatively confirm that the AJAHR model is effective not only for amputee subjects but also for non-amputee human mesh.

#### 4.2. Qualitative Experiments

We conducted a qualitative evaluation on 640 test images from the ITW-amputee dataset, comparing our model against existing models fine-tuned on A3D. While prior methods often failed to reconstruct amputated regions—resulting in distorted overall poses—our model, as shown in Fig. 4, successfully identified amputation regions even in side-view images and aligned the reconstructed mesh accurately with the person’s location in the image. Notably, in front-view images, existing models frequently misinterpreted amputated legs as folded limbs. In contrast, our model accurately distinguished between amputated and intact limbs, enabling anatomically consistent mesh reconstruction. These results demonstrate its ability to reconstruct amputee-specific body shapes and poses.

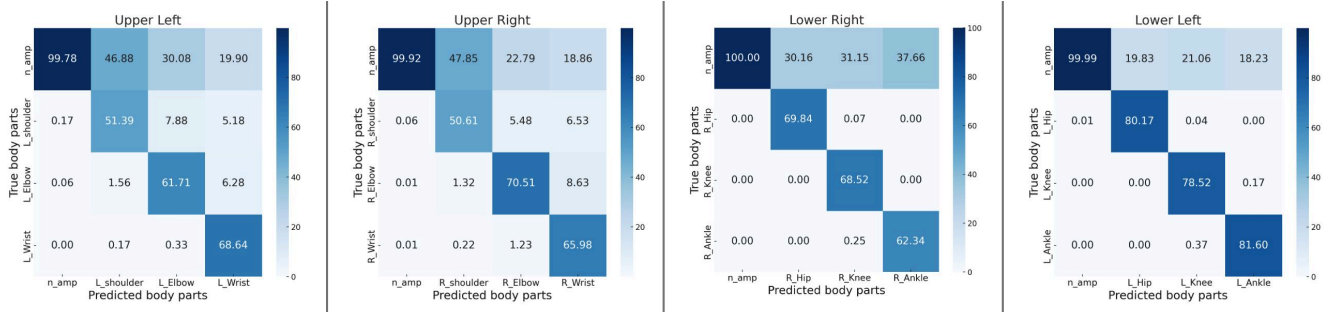


Figure 5. **Confusion Matrix of A3D Dataset Classification Results using BPAC-Net.**  $n_{amp}$  refers to non-amputee, and each label represents the body part predicted by each head. The values are expressed as percentages based on the predicted label columns.

### 4.3. Evaluating the Effectiveness of BPAC-Net

We assess the role of BPAC-Net in enabling amputation-aware mesh recovery through comprehensive experiments. First, we investigate the sensitivity of the overall pipeline to 2D keypoint quality by injecting Gaussian noise at varying levels. As shown in Tab. 4(a), performance degrades with increasing noise, indicating that 2D detector quality—e.g., ViTPose [42]—affects both classification and reconstruction. However, the drop remains moderate, suggesting robustness to real-world keypoint imperfections. We then compare single- and multi-modality inputs. Tab. 4(b) further shows that combining image and keypoint inputs yields superior classification performance over single-modality variants, supporting the effectiveness of our multi-modal design where spatial and appearance cues complement each other. Next, we evaluate BPAC-Net’s generalization to occlusion-heavy scenarios using the 3DOH50K [46]. As shown in Tab. 5, BPAC-Net distinguishes amputation from occlusion with a high F1 score of 0.977, demonstrating robustness even when limb visibility is affected by occlusion rather than absence. Finally, Fig. 5 shows the confusion matrices for each BPAC-Net classifier head. Classification is stable across all body parts, with the  $n_{amp}$  class achieving near-perfect accuracy. The Lower Left and Lower Right heads show fewer misclassifications than their upper-body counterparts, likely due to lower motion variability. These results highlight the effectiveness and reliability of BPAC-Net in conditioning the downstream pose recovery.

### 4.4. Necessity of complex model

As shown in Tab. 4(c), BPAC-Net built on a weak baseline does not yield performance gains, confirming that reliable pose estimates are essential for effective amputation classification. Furthermore, as presented in Tab. 4(e), using separate tokenizers for amputees and non-amputees leads to improved performance compared to a unified tokenizer. While full-body tokenization may limit joint-level interpretability, our dual-tokenizer strategy effectively compensates for this limitation by leveraging amputation-aware

cues derived from image and keypoint inputs. These results suggest that incorporating structural priors into tokenizer selection is beneficial when modeling subjects with amputation characteristics.

### 4.5. Effect of Token Count on AJAHR Performance

Tab. 4(d) presents the results of training the AJAHR model using AJAHR-Tokenizers with 160 and 640 tokens. The results demonstrate that the configuration with 320 tokens (Ours) yields the highest performance. When the number of tokens is insufficient, the model fails to capture pose variations across amputee and non-amputee individuals sufficiently. In contrast, using a large number of tokens introduces redundant information and increases token interference, which leads to degradation in performance.

## 5. Conclusion

Existing 3D human mesh recovery models [9, 10, 22] are not designed to handle limb amputations, often hallucinating missing limbs instead of recognizing actual absence. To address this, we present AJAHR—the first framework explicitly built for amputee mesh recovery. Our synthetic data pipeline enables ethical training without real amputee data, and our BPAC-Net-based architecture models amputation explicitly through regional classification and a tokenizer switching mechanism. Extensive experiments show that AJAHR significantly outperforms prior methods on amputee data, while maintaining competitive performance on standard benchmarks.

**Limitations and future directions.** AJAHR currently supports only joint-level amputations aligned with the SMPL kinematic tree, and the A3D dataset models only actual amputations, excluding prosthetics. Future extensions will target prosthetic limbs and irregular patterns beyond joint boundaries. From an application standpoint, the framework can support Paralympic sports analysis and inclusive AR/VR systems, enhancing accessibility for individuals with diverse limb differences, such as partial amputations or missing fingers.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] K Amrutha, P Prabu, et al. Human body pose estimation and applications. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–6. IEEE, 2021. 1
- [3] 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. 3, 6
- [4] Tobias Baumgartner and Stefanie Klatt. Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5109–5118, 2023. 1
- [5] Michael J Black, Priyanka Patel, Joachim Tesch, and Jintong 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. 3, 4, 6, 7
- [6] 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 *European conference on computer vision*, pages 561–578. Springer, 2016. 2
- [7] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 591–601, 2022. 1
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4
- [9] 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. 1, 2, 4, 5, 6, 7, 8
- [10] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 1, 2, 4, 6, 7, 8
- [11] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6
- [12] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 6
- [13] 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. 4
- [14] William Huang, Sam Ghahremani, Siyou Pei, and Yang Zhang. Wheelpose: Data synthesis techniques to improve pose estimation performance on wheelchair users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024. 2, 3
- [15] 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. 3, 6
- [16] 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. 2
- [17] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 6
- [18] 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. 3, 6, 7
- [19] 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. 3
- [20] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11127–11137, 2021. 2, 6
- [21] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 6
- [22] Zhihao Li, Jianzhuang Liu, Zhen Song Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 1, 2, 6, 7, 8
- [23] 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. 3, 6
- [24] 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. 2
- [25] 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. 2, 5, 6
- [26] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 6
- [27] Temitayo Olugbade, Marta Bienkiewicz, Giulia Barbareschi, Vincenzo D’amato, Luca Oneto, Antonio Camurri, Catherine Holloway, Mårten Björkman, Peter Keller, Martin Clayton, et al. Human movement datasets: An interdisciplinary scoping review. *ACM Computing Surveys*, 55(6):1–29, 2022. 2
- [28] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020. 2
- [29] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2
- [30] 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. 2
- [31] 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. 3
- [32] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017. 2
- [33] 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. 2, 5, 6
- [34] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in neural information processing systems*, 30, 2017. 2
- [35] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5
- [36] 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. 4, 5
- [37] 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, 6, 7
- [38] 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. 4
- [39] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shiwei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019. 6
- [40] Xinyao Xi, Chen Zhang, Wen Jia, and Ruxue Jiang. Enhancing human pose estimation in sports training: Integrating spatiotemporal transformer for improved accuracy and real-time performance. *Alexandria Engineering Journal*, 109: 144–156, 2024. 1
- [41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2
- [42] 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, 8
- [43] 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. 6
- [44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3
- [45] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 2
- [46] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 7, 8
- [47] Yi Zhang, Pengliang Ji, Adam Kortylewski, Angtian Wang, Jieru Mei, and Alan L Yuille. 3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation. In *The IEEE/CVF International Conference on Computer Vision*, 2023. 2
- [48] Tianxun Zhou, Muhammad Nur Shahril Iskandar, and Keng-Hee Chiam. Diffusion models enable zero-shot pose es-

timation for lower-limb prosthetic users. *PLOS Digital Health*, 4(3):e0000745, 2025. [2](#)

- [49] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. [5](#)