Age-Inclusive 3D Human Mesh Recovery for Action-Preserving Data Anonymization

Supplementary Material

This supplementary material provides additional technical details and results concerning our proposed method, AionHMR, and the newly introduced 3D-BabyRobot Dataset. The document is organized into four main sections: Section A details the optimization pipeline used for the AionHMR-a; Section B provides a deeper technical dive into AionHMR-b, specifically outlining its architectural configuration, a comprehensive breakdown of all utilized loss functions, and the composition of its training datasets; Section C describes the Action Preservation Test conducted while Section D provides additional information about the 3D-BabyRobot Dataset; and Section E discusses the data anonymization capabilities of AionHMR. Finally, Section F presents extensive qualitative results, including a large number of example images from the 3D-BabyRobot Dataset, along with visual comparisons illustrating the performance of AionHMR against two baselines: HMR2.0 [7] and BEV [22].

A. AionHMR-a Details

A.1. Optimization Stages

For the optimization phase of AionHMR-a, we adapt parts of the optimization process from SLAHMR [27], which we present more formally here.
For a video of T frames containing N people, each person i at time step t is represented as:

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}$$

where $\Phi_t^i \in \mathbb{R}^3$ is the global orientation, $\Theta_t^i \in \mathbb{R}^{22\times3}$ the body pose with 22 joint angles, $\beta^i \in \mathbb{R}^{11}$ the shape over all time steps t, where the 11^{th} value is the α interpolation weight, and $\Gamma_t^i \in \mathbb{R}^3$ the root translation.

The first step is to estimate each person's per-frame pose $\hat{\mathbf{P}}_t^i$ and compute their unique identity track associations over all frames using a 3D tracking system, 4DHumans [7].

In a video, the net motion, *i.e.*, a person's motion in the camera coordinates, depends both on the human's and camera's motion in the world frame. Therefore, the camera motion should be modeled in a correct way. Let ${}^c\mathbf{P}_t^i = \{{}^c\Phi_t^i, \Theta_t^i, \beta^i, {}^c\Gamma_t^i\}$ the pose in the camera frame and ${}^w\mathbf{P}_t^i = \{{}^w\Phi_t^i, \Theta_t^i, \beta^i, {}^w\Gamma_t^i\}$ the pose in the world with the same local pose $\hat{\Theta}_t^i$ and shape $\hat{\beta}^i$ parameters.

AionHMR-a uses DROID-SLAM [23], a SLAM system, to estimate the world-to-camera transform at each time t, $\{\hat{R}_t, \hat{T}_t\}$. This is essential to compute the relative camera motion between video frames. A human motion in the world prior is used to determine the camera scale α_c and people's global trajectories. The camera scale α_c is important to be estimated correctly to place the people in the world, so the human bodies and motion are plausible.

First, the global orientation and root translation in the world coordinate frame using the estimated camera transforms and camera-frame parameters are initialized. Camera scale is initialized at the value of $\alpha_c = 1$.

$$\begin{aligned} {}^w\Phi_t^i &= R_t^{-1c}\hat{\Phi}_t^i, & {}^w\Gamma_t^i &= R_t^{-1c}\hat{\Gamma}_t^i - \alpha_c R_t^{-1}T_t, \\ \beta_i &= \hat{\beta}_i, & \Theta_t^i &= \hat{\Theta}_t^i, \end{aligned}$$

The world frame joints are expressed as:

$${}^w\mathbf{J}_t^i = \mathcal{M}({}^w\Phi_t^i, \Theta_t^i, \beta^i) + {}^w\,\Gamma_t^i$$

where $\mathcal{M}$ is the differentiable function that the SMPL [15] model uses to generate the mesh vertices and joints.

SLAHMR defines a 2D joint reprojection loss to align the projected 3D to 2D joints with the detected from ViTPose 2D keypoints x_t^i that AionHMR-a also uses:

$$E_{\text{data}} = \sum_{i=1}^{N}\sum_{t=1}^{T} \psi_t^i \rho(\Pi_K(R_t \cdot {}^w\,\mathbf{J_t^i} + \alpha\mathbf{T_t}) - \mathbf{x_t^i})$$

where $\Pi_K(\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T) = K\begin{bmatrix} \frac{x_1}{x_3} & \frac{x_2}{x_3} & 1 \end{bmatrix}^T$ is perspective camera projection with camera intrinsics matrix $K \in \mathbb{R}^{2\times3}$, ρ is the robust Geman-McClure function [3] and ψ_t^i are the confidence scores of the detected 2D keypoints.

At this stage of the optimization, due to the underconstrained reprojection loss, the optimization is being held only to the global orientation and root translation ${}^w\Phi_t^i, {}^w\,\Gamma_t^i$ of the human pose parameters:

$$\min_{\{\{{}^w\Phi_t^i, {}^w\Gamma_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}}$$

The optimization lasts 30 iterations with $\lambda_{\text{data}} = 0.001$.

For the camera scale α_c, human shape β_i and body pose Θ_t^i optimization, additional priors about human movement in the world are used. This optimization stage smooths the transitions between poses in the world trajectories so that the displacements of the people are plausible. The prior of joint smoothness is defined as:

$$E_{\text{smooth}} = \sum_{i}^{N} \sum_{t}^{T} \|\mathbf{J_t^i} - \mathbf{J_{t+1}^i}\|^2$$

The other priors concern the pose $E_{\text{pose}} = \sum_{i=2}^{N} \sum_{t=1}^{T} \|\zeta_t^i\|^2$ and the shape $E_\beta = \sum_i^N \|\beta^i\|^2$, where $\zeta_t^i \in \mathbb{R}^{32}$ represent the body pose parameters Θ_t^i in the latent space of the VPoser [19] model. The updated objective function to be minimized is the following:

$$\min_{\alpha, \{\{{}^w\mathbf{P_t^i}\}_{\mathbf{t=1}}^{\mathbf{T}}\}_{\mathbf{i=1}}^{\mathbf{N}}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}$$

The optimization is performed over 60 iterations using $\lambda_{\text{smooth}} = 5, \lambda_\beta = 0.05$ and $\lambda_{\text{pose}} = 0.04$.

B. AionHMR-b Details

B.1. Losses

AionHMR-b uses different losses during the training, based on the ground-truth (or pseudo-ground-truth) annotations of the training datasets.

If ground-truth SMPL-A [18] shape parameters β^* and pose parameters θ^* are available, an MSE loss is used for the model predictions:

$$\mathcal{L}_{\texttt{smpl}} = \|\theta - \theta^*\|_2^2 + \|\beta - \beta^*\|_2^2$$

When the dataset provides accurate ground-truth 3D keypoints X^*, a L1 loss is added to penalize the distance from the predicted 3D keypoints X:

$$\mathcal{L}_{\texttt{kp3D}} = \|X - X^*\|_1$$

Similarly, if there are 2D keypoints annotations x^*, an L1 loss is used to penalize the projection of the predicted 3D keypoints $\pi(X)$:

$$\mathcal{L}_{\texttt{kp2D}} = \|\pi(X) - x^*\|_1$$

Finally, to get plausible 3D poses, a discriminator D_k is trained for each factor of the body model, *i.e.*, the body pose parameters θ_b, the shape parameters β and the per-part relative rotations θ_i with the generator loss expressed as:

$$\mathcal{L}_{\texttt{adv}} = \sum_k (D_k(\theta_b, \beta) - 1)^2$$

B.2. Architecture Details

The AionHMR-b model that we train consists of one Vision Transformer [5] image encoder and a transformer decoder [24]. The ViT encoder is taken from the ViTPose model [26], which was pre-trained for the 2D joint detection task. It takes as input a 256×192 image and consists of 50 transformer layers. The encoder outputs 16×12 image tokens, each of dimension 1280. These tokens serve as the encoded representation of the input image for the decoder.

The transformer decoder has 6 layers, each with multi-head self-attention, multi-head cross-attention, and feed-forward blocks with layer normalization. It has a hidden dimension of 2048. Both the self-attention and cross-attention blocks use 8 heads, each with a dimension of 64. The feed-forward MLP has a hidden dimension of 1024.

For the SMPL-A parameters prediction, a 2048-dimensional learnable SMPL-A query token is fed into the transformer decoder. The decoder uses cross-attention to attend to the 16×12 image tokens from the ViT encoder. The output of the decoder is then passed through a linear layer to predict the final parameters. The output of the network consists of the pose (θ), the shape (β), and the camera (π) parameters.

Table 1 shows the number of trainable parameters for the backbone, the SMPL-A head and the discriminator, for a total of 671M parameters.

Name	Type	Number of Trainable Parameters
backbone	ViT	630 M
smpla_head	SMPL-A Transformer Decoder Head	39.5 M
discriminator	Discriminator	1.8 M

Table 1. Trainable Parameters for Model Components

B.3. Training Dataset

For the initial training phase, we set the weights for the datasets as follows: the largest weight was assigned to our dataset (from SyRIP [10] and Relative Human [22])(0.50), followed by AVA [8], AIC [25], and INSTA [12] datasets, each weighted at 0.10. The remaining datasets (Human3.6m [11], MPII [1], COCO [14], and MPI-INF-3DHP [16]) were each weighted at 0.05. We decided to assign the largest weight to our dataset in order for the model to learn the child and infant shape. For validation, we split the weight equally between ours (validation subset) and COCO-VAL (both at 0.50). For the subsequent fine-tuning phase, the weights were adjusted: our dataset weight was reduced to 0.3000, since we now focus on 3D pose training. The remaining eight training datasets were weighted equally at 0.0875. The validation dataset weights for fine-tuning remained unchanged; our validation subset and COCO-VAL were both set to 0.50.

B.4. Architectures Comparison

We trained different architectures before selecting training from scratch the HMR2.0 [7] model. In Figure 1, we show qualitative examples of the different architectures and training schemes. These results show that the model struggles to learn the 3D shape of children when the HMR2.0 and TokenHMR [6] architectures are used. When we apply

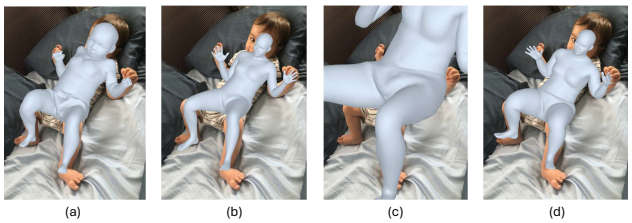

Figure 1. **Results of different training schemes and architectures.** (a) Training HMR2.0 from scratch, (b) Fine-tuning HMR2.0, (c) Fine-tuning HMR2.0 with LoRA, (d) Fine-tuning TokenHMR. While (b) and (d) provide accurate 3D Pose, their 3D shape estimation is inaccurate. Case (c) provides an inaccurate estimation of both shape and pose. Training from scratch HMR2.0 (a) has the best results for both 3D shape and pose.

LoRA [9] during HMR2.0 fine-tuning, the 3D reconstruction becomes inaccurate. However, training the HMR2.0 architecture from scratch produces high-quality reconstructions as the model learns the 3D shape of children.

B.5. Real-time Performance

To evaluate the real-time capabilities of AionHMR, we measured the inference latency on an NVIDIA RTX 3090 GPU. Under these conditions, AionHMR achieves an inference rate of approximately 20 FPS. The

C. Action Preservation

C.1. Description Generation

To obtain detailed, consistent, and action-focused descriptions for both the original and reconstructed videos, we employed a Large Video Language Model (LVLM) [2]. The following specific prompt was used to guide the LVLM to focus exclusively on human movement and ignore irrelevant details:

Prompt: `Act as a kinematic analyst. Describe the human's movement in this video by following this exact template. Ignore the background, clothing, and facial expressions.`
`1. Initial Pose: Describe the starting stance (e.g., feet width, arm placement) and orientation relative to the camera.`
`2. Trunk & Locomotion: Does the person stay in one spot, or is there a change in location? Describe any walking, stepping, or weight shifting.`
`3. Upper Body Sequence: Detail the movement of the right and left arms independently. Specify the path (e.g., circular, linear, lateral, or forward) and the height reached (e.g., waist level, overhead).`
`4. Extremities: Describe the state of the hands (open, fists, fingers spread/curled) and any specific finger actions.`
`5. Head & Gaze: Describe the rotation or tilt of the head and whether it follows the limb movements.`
`Constraint: Be clinical and objective. Instead of 'they move naturally,' use 'the arm swings in a sagittal arc.' If a limb does not move, explicitly state 'remains stationary.`

C.2. Descriptions Similarity

The core metric for action preservation is the semantic similarity between the two generated descriptions (original vs. reconstructed). We utilized a separate Large Language Model (LLM) [4] specifically for semantic comparison. This model was prompted to output a percentage score based purely on the actions. The prompt used for the similarity comparison is provided below, emphasizing the exclusion of non-action-related details:

Prompt: `Compare the two descriptions and give a percentage of similarity. Focus exclusively on the semantic similarity of the actions, movements, and implied behaviors of the human figure, while completely ignoring all details related to the setting (e.g., green mat, furniture, room description), clothing, and other appearance details.`

C.2.1. Example Video Descriptions

For illustration, we present two representative examples of the description pairs generated for the original video and its reconstructed counterpart, along with the similarity score assigned by the comparative LLM. The first exemplifies a case where, while the original video and the reconstructed video perform the same action from the same child, the description similarity is low. On the other hand, the second example presents two descriptions where the LLM assigned a semantic similarity score of 100%.

Similarity: 45%

- **Original**: `1.Initial Pose: The individual stands with feet shoulder-width apart, arms relaxed at the sides, and facing slightly towards the camera. The body is upright, and the gaze is directed forward.`
  `2.Trunk & Locomotion: The person`

remains in one spot, with no significant changes in location. There is minimal weight shifting, and the individual appears to be standing still.
3.Upper Body Sequence: Both arms remain stationary throughout the sequence. The right and left arms do not swing or move in any direction.
4.Extremities: The hands remain open and relaxed, with no specific finger actions observed.
5. Head & Gaze: The head remains upright and stationary, with no noticeable rotation or tilt. The gaze continues to be directed forward, maintaining a consistent orientation.

- **AionHMR**: 1.Initial Pose: The subject stands with feet shoulder-width apart, arms relaxed at the sides, and facing forward. The orientation is straight ahead, perpendicular to the camera.
2.Trunk & Locomotion: The subject remains in one spot, with no noticeable walking, stepping, or weight shifting. The body appears to be floating or suspended in place.
3.Upper Body Sequence: Both arms swing in a sagittal arc. The right arm swings forward and then back, reaching approximately waist level. The left arm mirrors the motion but lags slightly behind. The arms do not reach overhead.
4.Extremities: The hands remain open throughout the sequence. There are no specific finger actions observed.
5.Head & Gaze: The head remains stationary, facing forward. It does not follow the limb movements; instead, it maintains a fixed position.

Similarity: 100%

- **Original** 1.Initial Pose: The individual stands with feet shoulder-width apart, arms relaxed at the sides, and facing slightly towards the right side of the frame. The body is upright, and the gaze is directed forward.
2. Trunk & Locomotion: The person remains in one spot throughout the sequence. There is no noticeable walking, stepping, or weight shifting.
3.Upper Body Sequence: The right arm swings in a sagittal arc from the side to the front, reaching approximately waist level before returning to the initial position. The left arm remains stationary, hanging by the side.
4.Extremities: The hands remain open and relaxed. No specific finger actions are observed.
5.Head & Gaze: The head remains upright and stationary, maintaining a forward gaze that does not follow the limb movements.
- **AionHMR**: 1.Initial Pose: The subject stands with feet shoulder-width apart, arms hanging loosely at the sides. The subject faces slightly towards the right side of the frame.
2.Trunk & Locomotion: The subject remains in one spot, with no significant changes in location. There is minimal weight shifting, but the subject appears to be standing still.
3.Upper Body Sequence: The right arm swings in a sagittal arc from the side to the front, reaching approximately waist level. The left arm remains stationary, hanging at the side.
4.Extremities: The hands remain open throughout the sequence. No specific finger actions are observed.
5.Head & Gaze: The head remains stationary, facing forward. It does not follow the limb movements.

D. 3D-BabyRobot Dataset

D.1. 3D Children Reconstruction Pipeline

During interactions with robots, children exhibit a wide range of expressive gestures and intricate finger movements. Since AionHMR does not capture this information, we implement a pipeline to accurately reconstruct both the body and hands from a single image. Inspired by the multi-stage approach of HSfM [17], our pipeline first extracts the bounding boxes of each person in the image [21]. We then estimate 2D poses using ViTPose [26], the 3D reconstruction of the body using AionHMR and the hand pose using WiLoR [20]. To ensure anatomical consistency, we perform a joint optimization over the elbow and wrist pose parameters to align the global body pose with the local hand recon-

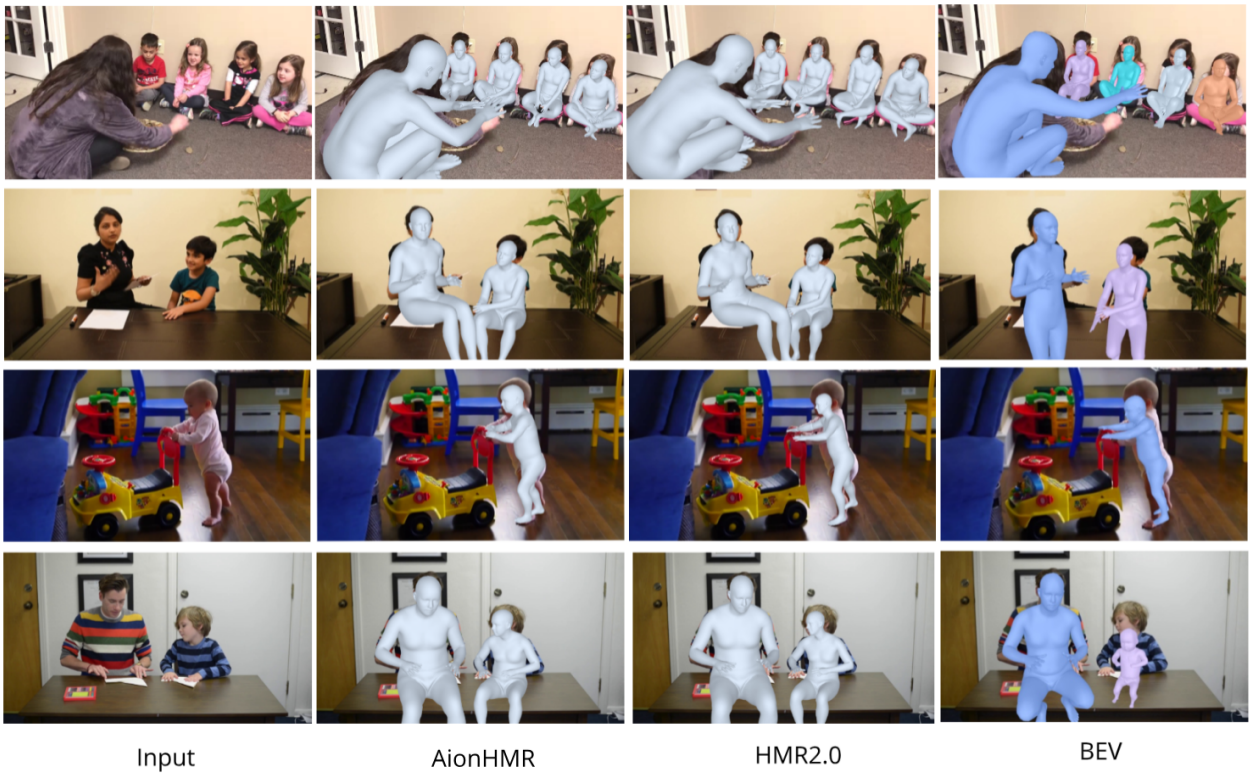

Figure 2. **AionHMR and Baselines Comparisons Examples.** From left to right: Original Image, AionHMR, HMR2.0 and BEV. Compared to HMR2.0 and BEV, AionHMR provides the most accurate estimation of the child's shape while effectively estimating the pose.

structions. Our total objective function is defined as:

$$\mathcal{L}_{total} = \lambda_w \mathcal{L}_{wrist} + \lambda_h \mathcal{L}_{hand} + \lambda_p \mathcal{L}_{vposer} + \lambda_g \mathcal{L}_{hinge} \tag{1}$$

To ensure consistency with the detected 2D keypoints, we minimize the reprojection error between the projected 3D wrist joints and the ViTPose wrist detections:

$$\mathcal{L}_{wrist} = \sum_{i \in \{L,R\}} \|\Pi(J_{wrist,i}) - \hat{x}_{wrist,i}\|_2^2 \tag{2}$$

where $\Pi(\cdot)$ denotes the camera projection operator, $J_{wrist,i}$ is the reconstructed 3D wrist joint, and $\hat{x}_{wrist,i}$ are the corresponding 2D keypoints predicted by ViTPose.

Hand Alignment Loss To align the predicted hands with the body skeleton, we enforce consistency between WiLoR's relative hand keypoints and the reconstructed wrist joints:

$$\mathcal{L}_{hand} = \sum_{i \in \{L,R\}} \|(J_{fingers,i} - J_{wrist,i}) - \mathcal{X}_{hand,i}\|_2^2 \tag{3}$$

where $J_{fingers,i}$ denotes the 3D finger joints predicted by WiLoR and $\mathcal{X}_{hand,i}$ represents the relative hand pose with respect to the wrist.

Pose Prior To maintain anatomically plausible body configurations, we regularize the latent body pose representation within the VPoser manifold [19]:

$$\mathcal{L}_{vposer} = \|z\|_2^2 \tag{4}$$

where z is the latent pose embedding predicted by the body model.

Elbow Hinge Constraint Finally, we restrict elbow rotations to behave approximately as hinge joints by penalizing rotations outside the primary flexion axis:

$$\mathcal{L}_{hinge} = \sum_{j \in \{elbows\}} \left(\theta_{j,x}^2 + \theta_{j,y}^2\right) \tag{5}$$

where $\theta_{j,x}$ and $\theta_{j,y}$ denote the elbow rotations around the non-flexion axes.

Following empirical tuning, we set the weights to $\lambda_w = 0.02$, $\lambda_h = 800.0$, $\lambda_p = 0.01$, and $\lambda_g = 50.0$. The optimization is performed using Adam [13] for 200 epochs with a learning rate of 0.04.

Figure 3. **Qualitative results from the 3D-BabyRobot Dataset**. The figure showcases frames from our newly released 3D-BabyRobot Dataset (over 4M frames of child-robot interaction). While the full dataset includes hand reconstructions via WiLoR, this figure specifically highlights AionHMR's body-level 3D shape and pose estimation to demonstrate the accuracy of our proposed method. These examples serve as qualitative results for the primary body reconstruction task.

D.2. Ethics and Data Collection

Data collection was approved by the Institutional Ethical Board with written parental informed consent. We strictly adhere to GDPR standards, and the dataset will be released under a CC BY-NC 4.0 license for non-commercial research with all compliance documentation.

E. Anonymization

A primary application of AionHMR is the anonymization of sensitive human data, particularly child and infant imagery where legal and ethical protections are most stringent. By generating a 3D mesh for every detected person, our framework provides a high-fidelity substitute for the original subject.

The resulting 3D meshes encode only body shape and pose; they do not retain primary identifiers such as facial features, skin texture, iris patterns, or clothing appearance. This decoupling allows for the visualization of human movement and geometric interaction without exposing the identifiable visual attributes of the minor. Unlike traditional blurring or pixelation, which obscure motion and joint relationships, AionHMR maintains the scientific integrity of the kinematic data.

We define our approach as Action-Preserving Anonymization. The core validation provided in Section 6 (demonstrating an 83.1% semantic similarity between original and reconstructed behaviors) shows that the research utility of the child's actions is preserved. While we acknowledge that distinctive body morphology or specific gait signatures could theoretically act as "soft biometrics," AionHMR significantly raises the threshold for re-identification compared to raw RGB imagery.

The most tangible evidence of this section's contribution is the release of the 3D-BabyRobot Dataset. By utilizing AionHMR to anonymize over 5.8 million frames, we facilitate large-scale child-centric research that would otherwise be prohibited by privacy regulations. This practice establishes a strong practical baseline for the ethical dissemination of sensitive datasets where zero-risk anonymization is technically unattainable.

F. Qualitative Results

In Figure 2 we present some comparisons of AionHMR and two baselines, HMR2.0 and BEV. Figure 3 provides an extensive sample from the 3D-BabyRobot Dataset, which also serves as a qualitative assessment of AionHMR.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Jonathan T. Barron. A general and adaptive robust loss function. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[5] 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 (ICLR)*, 2021. 2

[6] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[8] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022. 3

[10] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021. 2

[11] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36:1325–1339, 2013. 2

[12] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 5

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34:248:1–248:16, 2015. 1

[16] 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 *International Conference on 3D Vision (3DV)*, 2017. 2

[17] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21948–21958, 2025. 4

[18] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[19] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[20] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 4

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 4

[22] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[23] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[25] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. Large-scale datasets for going deeper in image understanding. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2019. 2

[26] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 4

[27] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1