Baby Mesh Recovery Enables Infant Behavior Identification

Charles Dardaman[1], Zifan Xu[1], Peter Stone[1,4], Karen Adolph[2], Danyang Han[3*], Georgios Pavlakos[1*]

Fig. 1. **Baby Mesh Recovery.** Given multiple calibrated and synchronized camera views, we reconstruct the 3D baby mesh. We use our 3D reconstruction to identify infant motor behavior, showing reliable body pose, body posture, and step identification.

***Abstract*— Natural infant motor behavior, such as sitting, crawling, and walking, is highly variable and idiosyncratic, and often deviates from standard, adult-like postures and locomotion. This variability makes it challenging for automated tools to support precise 3D video analysis of infant behavior. Here, we present a system for accurate 3D motion capture of infants' moment-to-moment body posture and step-to-step locomotion using synchronized multi-view RGB video. Our method builds on recent advances in parametric human modeling by adapting them to the unique challenges of infant behavior. We use a parametric baby body model—a "mesh baby"— to represent age-varying body shapes and develop a multi-view optimization approach that jointly recovers 3D body pose, shape, and motion. We demonstrate that our system provides accurate full-body 3D baby reconstruction and enables fine-grained analysis, including identification of posture, step counts, and gait timing. Our approach provides a scalable, non-intrusive, video-based tool for understanding infants' natural motor behavior and opens new avenues for early screening of developmental disorders and evaluation of clinical interventions. We release the code and results at `https://geopavlakos.github.io/babymesh/`.**

I. Introduction

Accurate video identification of infant natural motor behavior is critical for developmental and clinical science [1]. Specifically, tracking baby body posture and locomotion is critical to understand infant motor behavior and its development. Traditionally, researchers focus more on form than on function—using standardized tasks in which infants must sit or stand in a particular posture, or crawl or walk on a motorized treadmill or in a straight line over uniform open ground [2], [3], [4]. However, such standardized tasks do not capture the complexity, variability, and idiosyncrasies of infant movements during spontaneous everyday activity (see Figure 1), and thus cannot characterize natural motor behavior in infants [5] or in babies with atypical development or delay [6]. Without an adequate characterization of natural behavior, researchers cannot evaluate the efficacy of clinical interventions for everyday function [2].

In standardized tasks, the categories of infant motor behavior are clear-cut. But the prototypes for categories of sitting, crawling, standing, walking, and so on do not capture the variety and idiosyncrasies of infants' natural movements and the multiplicity of edge cases between categories from moment to moment. Infants, for example, often locomote in postures that defy the classic categories—by "inch-worming" or "snow-plowing" with their belly on the floor, "bum-shuffling" in a sitting posture, or "hitching" with one buttock and one knee or foot on the floor [7], [8]. They sometimes "walk" on their knees or "cruise" sideways while holding furniture for support. Even something as seemingly simple as identifying a walking "step" is not straightforward because infants often take steps in place, take backward or sideways steps, slide their feet along the floor, take single or double steps with the same foot, or partially lift their foot without displacement [5]. Such deviations are not rare anomalies. They are a central feature of real-world infant motor behavior, especially in early stages of skill acquisition and in

[1]Department of Computer Science, The University of Texas at Austin, {charlesdardaman, zfxu}@utexas.edu, {pstone, pavlakos}@cs.utexas.edu

[2]Departments of Psychology, Neuroscience, and Child & Adolescent Psychiatry, New York University, karen.adolph@nyu.edu

[3]Department of Psychology and Behavioral Sciences, Zhejiang University, danyang.han@zju.edu.cn

[4]Sony AI

*Correspondence concerning this article should be addressed to Georgios Pavlakos and Danyang Han.

the everyday environment with enticing objects, clutter in the path, and varied terrain that motivates varied movements [2].

Despite such challenges, human observers can easily identify the most variable, idiosyncratic infant postures and locomotion from video based on commonsense or rule-based criteria. However, manually labeling postures and locomotor steps is time-consuming and laborious. Moreover, human observers cannot easily quantify fine-grained spatial properties such as head orientation, limb angles, foot height from the floor, foot displacement, and exact location of the body or limbs in 3D. Accurate automated methods for analyzing infants' natural behavior from video could thus unlock new insights into motor development, provide a rigorous foundation for developmental science, and support clinical applications such as early detection of atypical development.

Simultaneously, the field of 3D human pose estimation has made tremendous progress in recent years, with algorithms now capable of reconstructing human pose from images or videos [9], [10], [11]. However, such methods are optimized for adults and do not generalize to infants, whose proportions and movements differ substantially from those of adults. Previous work addressed this gap by developing specialized tools to detect isolated features of infant posture—for example, by localizing infants' feet [12] or estimating head pose [13]. We argue that the recovery of infant motor behavior is most effective when performed holistically, through full-body 3D reconstruction—a "mesh baby"—that captures the entire baby pose and movements over time.

In this work, we present a system for accurate 3D motion capture of infants using a small number of calibrated and synchronized cameras. We build on recent advances in 3D human reconstruction [9], [10], extending them to accommodate infant-specific shape patterns by incorporating a parametric baby body model, SMIL [14]. Our multi-view optimization framework jointly recovers body shape, pose, and movement, enabling precise full-body reconstructions that are robust to the complexities of infant motor behavior.

We demonstrate that our system achieves superior accuracy compared to existing methods, both for 3D pose recovery and in critical downstream tasks such as head pose estimation, joint localization, whole-body posture classification, and step identification and displacement in 3D. By enabling high-fidelity analysis of natural infant behavior—replete with infant variability and idiosyncrasies—our approach offers a new tool for developmental research and lays the foundation for data-driven understanding of early motor behavior. Our code, along with the anonymized data, *i.e.*, the reconstructed 3D meshes are available at `https://geopavlakos.github.io/babymesh/`.

II. Related work

3D human pose estimation. Although 3D human pose estimation has a long history [15], recent progress is largely driven by the success of deep learning and the incorporation of strong priors for human reconstruction. Early work relied on simplified representations, e.g., 3D keypoints [16], but recent approaches typically leverage parametric body models like SMPL [17], that provide a representation of the body surface. In this paradigm, reconstruction pipelines typically take as input a single image [18], [19] or video sequence [20] and estimate the corresponding SMPL parameters. This framework was extended successfully to more complex tasks, including multi-person tracking [21] and 4D human motion recovery [10], [11]. Here, we build on recent advances in 3D human motion recovery, and adapt them to the domain of infant behavior. We demonstrate that these tools, when properly adapted to account for age-specific differences in body shape and movement patterns, enable accurate and robust 3D baby motion reconstruction.

Baby pose estimation. In general, the specialized work on baby pose estimation is limited. For 2D keypoint detection, standard methods such as ViTPose [22] generalize reasonably well to images of infants [23]. However, for metric 3D pose estimation, performance on children, especially infants, degrades substantially because the most commonly used parametric models, such as SMPL [17] and SMPL-X [24], are trained to represent adult body shapes and proportions. To address this gap, SMIL [14] extends the SMPL framework with a body shape space specifically designed for infants, while remaining compatible with the SMPL family. Building on this extension, SMPL-A [25] offers a unified model that integrates the shape spaces of SMIL and SMPL, enabling consistent representation across the full age spectrum, from infants to adults. Despite this progress in modeling, few reconstruction approaches can robustly handle the wide variability in body shapes found in early development. BEV [26] is an example that uses the SMPL-A model and can reconstruct a broad range of human shapes from a single image. In this work, we adopt the SMPL-A model and leverage it to recover accurate 3D baby pose from multi-view RGB captures, enabling full-body reconstructions even under the variable, non-standard postures and locomotor patterns typical of infant behavior.

Body posture classification. In computer vision, body posture classification has been studied extensively, often as part of human action or activity recognition and serving as a precursor to higher-level behavioral analysis. Action classification datasets, such as Kinetics [27], include classes corresponding to specific postures (e.g., sitting, standing), leading to posture classification based on general action recognition methods. These methods typically process image sequences as input [28]. In some cases, 2D or 3D pose information is incorporated as an auxiliary input [29], [30]. There is also dedicated work on baby posture classification [31], [32], [33], where 2D and 3D pose information is used to decide posture labels. Similarly, in our experiments, we leverage features from our reconstructed 3D baby meshes, (e.g., 3D joint positions), as input to a posture classifier.

Step identification. Moment-to-moment step identification is necessary to quantify how much infants walk and to capture the spatio-temporal characteristics of their walking patterns. Recent advances in computer vision enable video-based gait analysis. However, existing methods are inadequate for infant

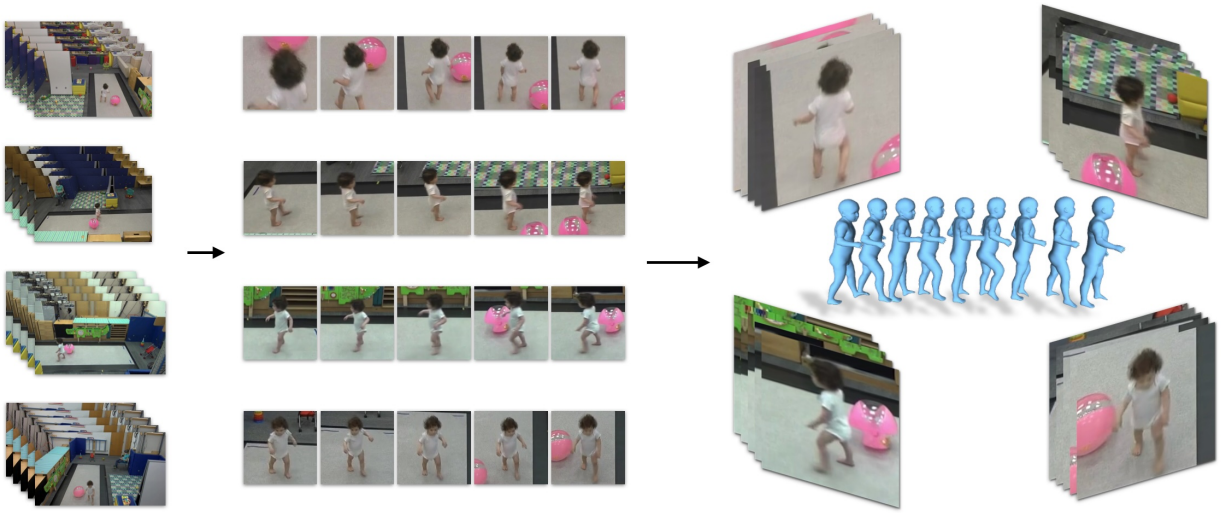

Fig. 2. **Overview of our 3D baby mesh reconstruction approach.** We start with synchronized and calibrated multi-view video. First, we track the people in the scene and we identify the tracklets that correspond to the baby (as described in "Preprocessing" of Section III-B). Then, given the tracklets that are detected for a specific temporal sequence, we reconstruct the baby using our multi-view optimization approach (as described in "Multi-view optimization" of Section III-B). We use these 3D mesh results for subsequent downstream tasks (i.e., posture and gait identification) and evaluation.

step identification, as they are trained on adult data [34], or are geared for highly controlled tasks such as treadmill stepping [35], and do not generalize to the variability and idiosyncrasies of infant everyday natural walking. Here, we use RGB video cameras to identify steps with high variability and complexity in freely mobile infants in a natural environment based on foot displacement from the whole-body pose.

III. Technical Approach

Our technical approach includes 3D baby mesh reconstruction from synchronized multi-view video data, baby posture classification, and our algorithm for baby step identification given the reconstructed locomotor sequences. To the best of our knowledge, our baby mesh reconstruction is the first to reliably estimate 3D infant body pose from multiple views.

A. Preliminaries

We represent the human body using the SMPL-A model [25], a parametric 3D human body model that extends SMPL (for Skinned Multi-Person Linear) [17] to support age-varying body shapes, including infants. SMPL-A defines a function: $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \rightarrow \mathbf{M} \in \mathbb{R}^{N \times 3}$, where $\boldsymbol{\theta} \in \mathbb{R}^{72}$ are the pose parameters (axis-angle rotations for 24 joints), $\boldsymbol{\beta} \in \mathbb{R}^{11}$ are the shape parameters, and $\mathbf{M}$ is a mesh of $N = 6890$ vertices. Unlike SMPL, which uses a 10-dimensional PCA shape space to model adult body shape variations, SMPL-A introduces an additional parameter $\beta_{11} \in [0, 1]$ that controls a linear interpolation between the adult and infant shape spaces. Specifically, $\beta_{11} = 0$ corresponds to adult body shape, while $\beta_{11} = 1$ corresponds to infant body shape, with intermediate values returning an interpolation between the two. The output mesh $\mathbf{M}$ shares the same topology as SMPL, enabling compatibility with standard processing tools.

B. Multi-view baby mesh reconstruction

Preprocessing. Let the input be a set of V synchronized and calibrated RGB video streams, $\{I_v^t\}_{v=1}^V$, each capturing T frames at time $t = 1, \ldots, T$. We begin by applying PHALP [21] to each video I_v^t, which returns a set of *tracklets*, *i.e.*, temporally consistent person detections across frames. For each view v, we denote the detected tracklets as: $\mathcal{T}_v = \{\tau_v^{(i)}\}_{i=1}^{N_v}$ where $\tau_v^{(i)}$ is the i-th tracklet and N_v is the number of people detected in that view. Afterwards, we run ViTPose [22] on each bounding box to estimate 2D body keypoints. For the i-th person in view v at time t, this gives: $\mathbf{x}_v^{t,(i)} \in \mathbb{R}^{J \times 2}$ where J is the number of keypoints. Since the baby's identity is not known a priori, we solve a *multi-view association problem*: for each timestep t, we consider all plausible combinations of per-view detections $\{\mathbf{x}_v^{t,(i)}\}_{v=1}^V$ and attempt to triangulate them into 3D joint locations $\mathbf{X}^t \in \mathbb{R}^{J \times 3}$. We define: $\mathbf{X}^t = \text{Triangulate}(\{\mathbf{x}_v^{t,(i_v)}\}_{v=1}^V)$ for a given selection $(i_1, \ldots, i_V)$ of tracklet indices, and estimate the limb lengths $\mathbf{L}^t$ from $\mathbf{X}^t$. Configurations whose proportions fall within expected limb lengths for infants are accepted as valid baby detections. We then backtrack to recover the corresponding 2D detections $\{\mathbf{x}_v^{t,(i_v^*)}\}$ and assign the tracklets $\{\tau_v^{(i_v^*)}\}$ as those belonging to the baby.

Multi-view optimization. For each frame t, camera $v \in \{1, \ldots, V\}$, and body joint $j \in \{1, \ldots, J\}$, we denote the 2D keypoint detection as $\mathbf{x}_v^{t,(j)} \in \mathbb{R}^2$ with confidence score $c_v^{t,(j)} \in [0, 1]$. Because the cameras are calibrated, we can project 3D points into each view using the standard pinhole model: $\Pi_v(\mathbf{X}) = \pi(\mathbf{K}_v[\mathbf{R}_v \,|\, \mathbf{t}_v]\,\mathbf{X})$ where π denotes perspective projection, $\mathbf{K}_v$ is the intrinsic matrix, and $\mathbf{R}_v, \mathbf{t}_v$ are the extrinsic rotation and translation for camera v.

We adopt an optimization-based approach to estimate the SMPL-A parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$, and global root translation $\boldsymbol{\tau}$ for each frame by fusing information from all views. The core

objective is a multi-view reprojection loss:

$$\mathcal{E}_{\text{proj}} = \sum_{v=1}^{V}\sum_{t=1}^{T}\sum_{j=1}^{J} c_v^{t,(j)}\rho\left(\Pi_v\left(\mathbf{X}^{t,(j)}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\tau})\right) - \mathbf{x}_v^{t,(j)}\right), \tag{1}$$

where ρ is a robust penalty [36] and $\mathbf{X}^{t,(j)}$ is the 3D location of joint j at time t, computed from the SMPL-A mesh.

To regularize the pose, we leverage the VPoser prior [24]. Rather than optimizing directly over $\boldsymbol{\theta}$, we optimize over the VPoser latent code $\mathbf{z}$ and decode to pose parameters as needed. To encourage plausible body configurations, we penalize the magnitude of $\mathbf{z}$: $\mathcal{E}_{\text{pose}}(\mathbf{z}) = \|\mathbf{z}\|^2$.

For shape regularization, we apply an ℓ_2 prior on the shape coefficients and a soft constraint to keep the age parameter within the valid infant range:

$$\mathcal{E}_{\text{shape}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}_{1:10}\|^2 + \exp\left(\lambda_{\beta_{11}}\left(\beta_{11} - 1\right)\right), \tag{2}$$

where $\lambda_{\beta_{11}}$ controls the weight of the age constraint.

Finally, we impose a temporal smoothness term: $\mathcal{E}_{\text{smooth}} = \sum_{t=1}^{T-1}\|\mathbf{X}^{t+1} - \mathbf{X}^{t}\|^2$. The total loss combines all objectives:

$$\mathcal{E}_{\text{total}} = \lambda_{\text{proj}}\mathcal{E}_{\text{proj}} + \lambda_{\text{pose}}\mathcal{E}_{\text{pose}} + \lambda_{\text{shape}}\mathcal{E}_{\text{shape}} + \lambda_{\text{smooth}}\mathcal{E}_{\text{smooth}} \tag{3}$$

We optimize in two stages. In the first stage, we estimate only the global root translation $\boldsymbol{\tau}$ and body orientation while keeping the pose fixed. In the second stage, we jointly refine the full body pose (via the latent code $\mathbf{z}$) and shape $\boldsymbol{\beta}$.

C. Posture Classification

A posture classifier can be formulated as a function $\phi\colon \mathcal{X} \mapsto \mathcal{Y}$ from an input feature space $\mathcal{X}$ to a discrete set of posture classes $\mathcal{Y}$. The input features are the 3D positions of $P = 25$ body keypoints extracted from the reconstructed SMPL-A mesh. These keypoints include a combination of SMPL-A skeletal joints and additional mesh-derived vertex locations, and they are expressed relative to the root joint. Formally, each frame is represented as: $(\hat{\mathbf{p}}^1, \hat{\mathbf{p}}^2, \ldots, \hat{\mathbf{p}}^P) \in \mathbb{R}^{P\times 3}$ where each keypoint position is centered at the root: $\hat{\mathbf{p}}^k := \mathbf{p}^k - \mathbf{p}^0$, with $\mathbf{p}^k$ denoting the 3D position of the k-th keypoint in the world frame. In addition to the joint positions, we augment the input into the model with 28 geometric features which are computed from the raw joint positions: 6 joints angles (bilateral knee, hip flexion, and ankle angles), 6 relative heights of key joints with respect to the lowest foot contact point, 12 normalized limb orientation vectors (bilateral thigh skin directions), and 4 summary features including bilateral hip-to-ankle distances, hip-to-ground height, and torso inclination angle.

The classifier ϕ is implemented as a part-aware multi-branch MLP. The 25 joints are partitioned into three anatomically motivated groups: a *core* branch (7 joints: nose, neck, left/right shoulders, mid-hip, and left/right hips), an *upper-limb* branch (8 joints: left/right elbows, left/right wrists, left/right eyes, and left/right ears), and a *lower-limb* branch (10 joints: left/right knees, left/right ankles, left/right big toes, left/right small toes, and left/right heels). Each branch is encoded by a dedicated two-layer MLP with GELU [37] activations; a parallel encoder processes the 28 geometric features. The four branch embeddings are concatenated and passed through a shared three-layer classification head. Model parameters are optimized with AdamW [38] using cosine learning-rate annealing. Class imbalance is handled by weighted random sampling combined with class-weighted cross-entropy loss. The final checkpoint is then selected based on the minimum per-class validation accuracy.

D. Step Identification

Given the reconstructed full-body poses $\boldsymbol{\theta}_t$, we extract the left and right ankle positions $\mathbf{p}_t^{\text{lAnkle}}$ and $\mathbf{p}_t^{\text{rAnkle}}$ at each timestamp t and compute the frame-to-frame displacement:

$$d_t^l = \|\mathbf{p}_{t+1}^{\text{lAnkle}} - \mathbf{p}_t^{\text{lAnkle}}\|_2, \quad d_t^r = \|\mathbf{p}_{t+1}^{\text{rAnkle}} - \mathbf{p}_t^{\text{rAnkle}}\|_2. \tag{4}$$

Steps are characterized by local peaks in these displacement signals. For the left ankle, a candidate step at time t requires $d_t^l > d_{t-1}^l$ and $d_t^l > d_{t+1}^l$. This is further parameterized by a prominence threshold p, which requires a peak to exceed its surrounding local baseline by at least p, filtering out small oscillations caused by pose-estimation noise that do not correspond to a genuine foot lift, and a minimum inter-peak distance $t_{\min}$, which prevents the rising and falling edges of a single step's displacement pulse from being double-counted as two distinct steps. This criterion yields a raw step count N_l^{raw} and a corresponding set of timestamps $\mathcal{T}_l^{\text{raw}} = \{t_n\}_{n=1}^{N_l^{\text{raw}}}$ indicating when left steps occur. The same process is then applied to the right side.

To eliminate spurious or duplicate detections, we apply the following post-processing steps to $(\mathcal{S}_l^{\text{raw}}, \mathcal{S}_r^{\text{raw}})$: (i) shift each detected peak by a small temporal offset to align with step completion, $\hat{\tau} = \min(\tau + t_{\text{offset}}, T-1)$ where $t_{\text{offset}} = 2$ frames; (ii) enforce plausible left-right alternation by inserting missing opposite-foot events when supported by local movement magnitude; (iii) resolve near-duplicate left/right events within a short window using a movement-ratio rule; and (iv) apply a dense-cadence rescue pass with a reduced minimum inter-peak distance. If no step is detected near the end of a bout, a final end-of-bout completion is added. This yields refined sets $\mathcal{S}_l^{\text{final}}$ and $\mathcal{S}_r^{\text{final}}$, with counts N_l^{final} and N_r^{final}. In our implementation, we use $p = 0.0065\,\text{m}$, $t_{\min} = 14.65$, a near-duplicate window of 4 frames with movement-ratio threshold 0.7, and a dense-cadence rescue pass with $t_{\min}^{\text{rescue}} = 11.0$. End-of-bout handling uses a 7-frame terminal window. These hyperparameters are selected via a search on a small held-out validation set.

IV. Experiments

Our experiments involve a video dataset of infant natural behavior. The first goal of our evaluation is to test whether our unified baby mesh recovery pipeline yields higher 3D pose accuracy than the prior work [13], which relied on separate algorithms for different body parts. The second goal is to test if the reconstructed 3D meshes can be used directly for downstream analysis, such as posture classification and step identification. We compare "gold-standard" reliable,

Method	RWrist	LWrist	Head	Head Pose
Ours	**4.7cm**	**4.0cm**	**5.5cm**	**24.3°**
Han *et al.* [13]	7.4cm	5.7cm	6.8cm	60.8°

TABLE I

3D POSE ACCURACY EVALUATION. WE COMPARE OUR APPROACH WITH [13]. WE REPORT MEAN POSITION ERROR FOR THE WRISTS AND THE HEAD IN CM, AND THE MEAN HEAD POSE ERROR IN DEGREES.

human annotations of joint location, head pose, body posture, and steps to the output of our system.

Dataset. We use the dataset introduced by Han *et al.* [13], which contains multi-view RGB video capture of infants during natural, unconstrained play in a large laboratory play room with abundant toys and furniture. Each session includes 8 synchronized and calibrated fixed camera views capturing a wide field of the play environment. We conduct our analysis on 47 10-min recording sessions (M = 10.33, SD = .48).

Human Annotations. (1) Posture. We classify baby body posture frame-by-frame for 12 sessions into 7 distinct classes: *supine* (on back or side), *prone* (belly on floor, elbows extended/flexed, chest propped/not propped, pivoting), *quadruped* (belly off floor, hands-knees crawling, hands-feet crawling, crawl with one knee and one foot), *sitting* (hands free or used to prop body, butt on ground or heels), *kneeling* (one or both knees on floor, butt not on heels), *squatting* (not supporting weight on hands on floor, butt below knees), and *upright* (cruising, standing, walking, legs straight). (2) Walking steps. We manually annotate infant locomotor bouts (e.g., crawling, walking) for every session and we randomly select subsets of walking bouts to provide fine-grained step annotations. We select 202 bouts (based on at least 5 walking bouts per session) to identify the number of steps and 102 bouts (at least 2 walking bouts per session) to identify left versus right steps and the timing of individual steps. (3) Joint location and head pose. We evaluate our 3D reconstructions using the 3D positions of the wrist joints and the head (midpoint between the eyes), as well as the orientation of the head, using the annotations provided by Han *et al.* [13].

Human Annotation Reliability. Human annotators achieve high agreement in annotating body postures, number of steps, and timing of steps. The percent agreement for the frame-wise body posture identification between human annotators was 97.87%, Cohen's κ = .97. The correlations between annotators was nearly perfect for the total, left, and right step counts, rs = 1.0, ps $< .001$. For 93.40%, 94.12%, and 96.08% of bouts, the total, left, and right step counts were the same between the two annotators. For step timing, 96.24% and 96.60% of one annotator's steps were identified by the other annotator within 100 ms (3 frames). Reliability for wrist location, head position and pose are presented in [13].

Baselines. For our evaluation, the primary baseline is the prior work of Han *et al.* [13] which proposed the dataset we use for our experiments. They propose separate approaches for 3D wrist location, head pose, and position estimation. We compare their approaches to the 3D estimates from our 3D mesh reconstruction. Moreover, for baseline comparisons, we evaluate our multi-view optimization approach against

Method	Supine	Prone	Quadruped	Sit	Kneel	Squat	Upright	Overall
Ours	**95.0**	**82.8**	**82.6**	**78.6**	**80.6**	**85.9**	**95.6**	**89.8**
Triangulation	6.5	0.1	66.3	76.5	64.1	3.1	93.7	65.8

TABLE II

POSTURE CLASSIFICATION ACCURACY (%). WE REPORT PER-CLASS AND OVERALL ACCURACY ACROSS 7 POSTURE CLASSES. WE COMPARE AN MLP CLASSIFIER USING AS INPUT OPTIMIZED SMPL-A JOINTS (OURS) AND TRIANGULATED 3D JOINTS (BASELINE). ACROSS THE THREE TEST SESSIONS, ACCURACY WAS $89.8\% \pm 4.2\%$ FOR "OURS", COMPARED TO $65.8\% \pm 41.9\%$ FOR THE TRIANGULATION BASELINE, REFLECTING SUBSTANTIALLY MORE CONSISTENT PERFORMANCE.

Method	Relative error	w/ τ_{th}=30ms	w/ τ_{th}=90ms	w/ τ_{th}=150ms
Ours	**-0.6%/-0.1%**	**58.2%/59.2%**	**89.6%/88.6%**	**95.5%/95.6%**
Triangulation	-8.6%/-10.9%	48.4%/48.7%	77.9%/79.3%	86.4%/88.7%

TABLE III

STEP IDENTIFICATION RESULTS. WE REPORT THE RELATIVE ERROR IN THE STEP COUNTING RESULTS (POSITIVE FOR OVERCOUNTING, NEGATIVE FOR UNDERCOUNTING) AND THE PERCENTAGE OF STEPS IDENTIFIED TO BE WITHIN A SPECIFIC TEMPORAL THRESHOLD τ_{TH}. WE REPORT RESULTS SEPARATELY FOR THE LEFT AND THE RIGHT LEG (IN L/R FORMAT). WE INCLUDE RESULTS WHEN USING 3D JOINTS FROM OUR OPTIMIZATION VS THE STRAIGHTFORWARD TRIANGULATION.

a simple triangulation-based method that also aggregates multi-view information but does not leverage a parametric baby body model [14]. As a result, it lacks the strong shape and proportion priors provided by the SMIL/SMPL-A model, which are crucial for reconstructing anatomically plausible infant poses. We assess the performance of both methods on the downstream task of posture and step identification, where precise pose estimation is critical to accurately classify the body posture and to detect subtle locomotor events.

Tasks. We conduct three experiments to evaluate the effectiveness of our system in reconstructing infant body pose and enabling downstream behavioral analysis:

- **3D Pose Accuracy:** We first evaluate the raw accuracy of the recovered 3D joint positions. We measure the Euclidean distance between the predicted and annotated 3D joint positions (in centimeters). We evaluate 3D head orientation by reporting the angular error (in degrees) between the predicted and ground-truth head pose vectors.
- **Posture Classification:** We evaluate whether the recovered 3D mesh can be used for frame-level posture recognition. We report per-class and overall classification accuracy.
- **Step Identification:** We use the reconstructed 3D baby skeletal sequence to detect and count walking steps. We report the relative error in left and right step identification, and the percentage of correctly temporally localized steps for different timing thresholds.

3D Pose Accuracy. For 3D pose accuracy, we compare the accuracy of our approach, with the main baseline of Han *et al.* [13]. Specifically, we report the average position error for the wrist and the head, and the average rotation error for the head pose. The complete results are presented in Table I, where we consistently outperform the previous baseline. We highlight that previous work [13] developed separate techniques to track the wrists and the head from multi-view data.

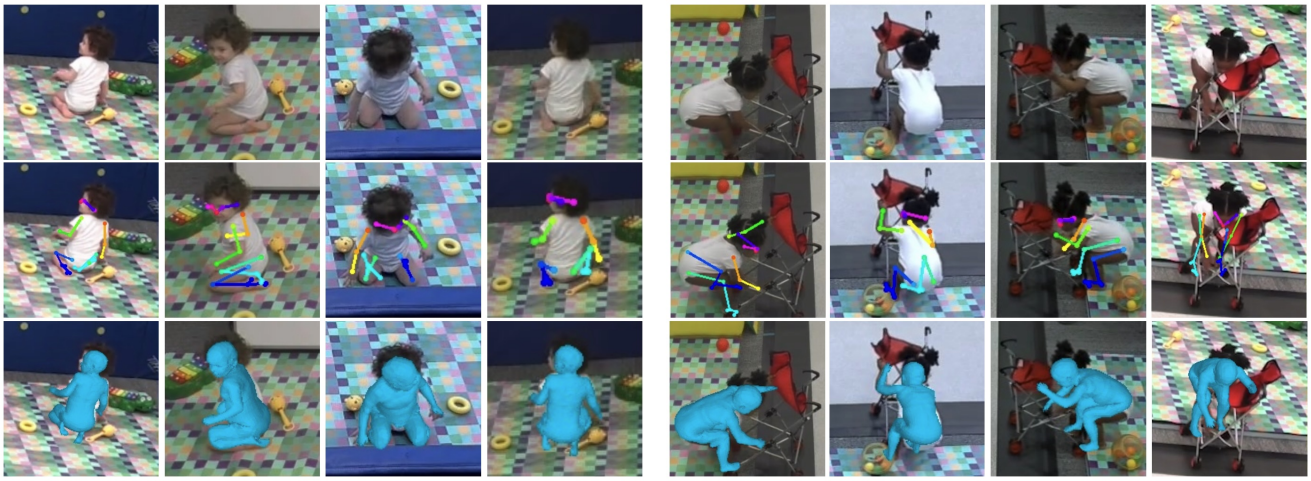

Fig. 3. **Multi-view optimization results.** In the first row, we show the input to our approach. Typically, we have access to image observations for the baby from 2-5 views. In the second row, we show keypoint detections from ViTPose [22]. We use these keypoints along with their confidence for the multi-view optimization. In the third row, we show results of our multi-view optimization approach, visualized and rendered on all input viewpoints.

Instead, our approach consolidates all information in our 3D mesh representation, where we get a complete 3D baby reconstruction and we can read out specific measurements, *e.g.*, the wrist location, or the head orientation.

Posture Classification. Our annotated dataset consists of 12 sessions; 6 are used for training, 3 for validation and 3 for testing. Classification accuracy is reported on the held-out test sessions. We use the 3D joint positions from our reconstructed SMPL-A meshes, centered at the root joint, together with 28 geometric features derived from the raw joints as input to a part-aware MLP classifier (as described in Section III). For training, we use AdamW with learning rate 3×10^{-4}, weight decay 10^{-5}, label smoothing (0.05), and cosine learning-rate annealing for 50 epochs. To address class imbalance, we use weighted random sampling and class-weighted loss. We select the final checkpoint based on validation minimum per-class accuracy.

Table II reports per-class and aggregate classification accuracy across 7 posture classes for two input sources: (1) our optimized SMPL-A joint trajectories and (2) triangulated 3D joints, using the same architecture and training protocol. Our classifier using optimized SMPL-A joints achieves an overall accuracy of 89.8%. In contrast, the triangulated baseline drops to 65.8% overall, with severe failure on prone (0.1%), supine (6.5%), and squatting (3.1%).

Step Identification. For our final experiment, we evaluate the use of our reconstructed 3D baby mesh sequences for the tasks of step counting, left-right step discrimination, and temporal precision of step detection. We use the step detection algorithm described in Section III-D and we compare two different versions of 3D joint positions—one that uses the results of our optimization and another that uses the triangulated 3D joints. The comparison is shown in Table III. We first report the relative error between detected and annotated steps. Positive values indicate overcounting, and negative values correspond to undercounting. The closer the relative error is to zero, the more reliable the counting algorithm. With our 3D reconstructions, the relative error remains around -0.6%, indicating only slight undercounting. In contrast, with triangulated keypoints the error exceeds -10.9%, reflecting severe undercounting. Step counts were strongly correlated with ground-truth annotations across all locomotion bouts (r = 0.9997, $p < .001$), the triangulation baseline yielded a weaker correlation (r = 0.9558, $p < .001$). For a more detailed analysis, we examine the temporal offset between each detected step and the corresponding ground-truth annotation. We report the percentage of correctly localized steps for several thresholds τ_{th}, approximately 30 ms, 90 ms, and 150 ms (corresponding to 1, 3, and 5 frames, respectively). This finer-grained evaluation again demonstrates the superiority of our 3D estimates for step identification. Moreover, we observe that most detected steps are localized well temporally with only short temporal offsets.

Qualitative evaluation. Complementing the quantitative evaluation, we also provide qualitative results of our approach. In Figure 3, we present reconstructions of babies from multiple camera views. Even in challenging poses, our method produces consistent and reliable 3D meshes. In Figure 4, we compare our results against state-of-the-art human mesh recovery approaches. Specifically, we evaluate HMR2.0 [9] a top-performing baseline, and BEV [26], which can also handle baby body shapes. Although these baselines operate on single frames and thus are not directly comparable to our method, it is helpful to identify the limitations of existing approaches. HMR2.0 captures body pose reasonably well but fails to model baby-specific proportions—in particular, note the discrepant head size, while BEV is less robust overall and struggles with accurate pose estimation.

Visual attention. One of the attributes of the dataset we use for our experiments is that it provides a digital twin of the recording environment. Given our 3D baby reconstruction, this allows us to reason about the baby's interaction with the environment. For each frame, we leverage the estimated head orientation to define a *visibility cone*, i.e., a geometric approximation of the infant's instantaneous field of view, centered on the head and aligned with the estimated head/gaze direction. Figure 6 illustrates this for three instances of the mesh baby in different postures: lying, squatting, and upright.

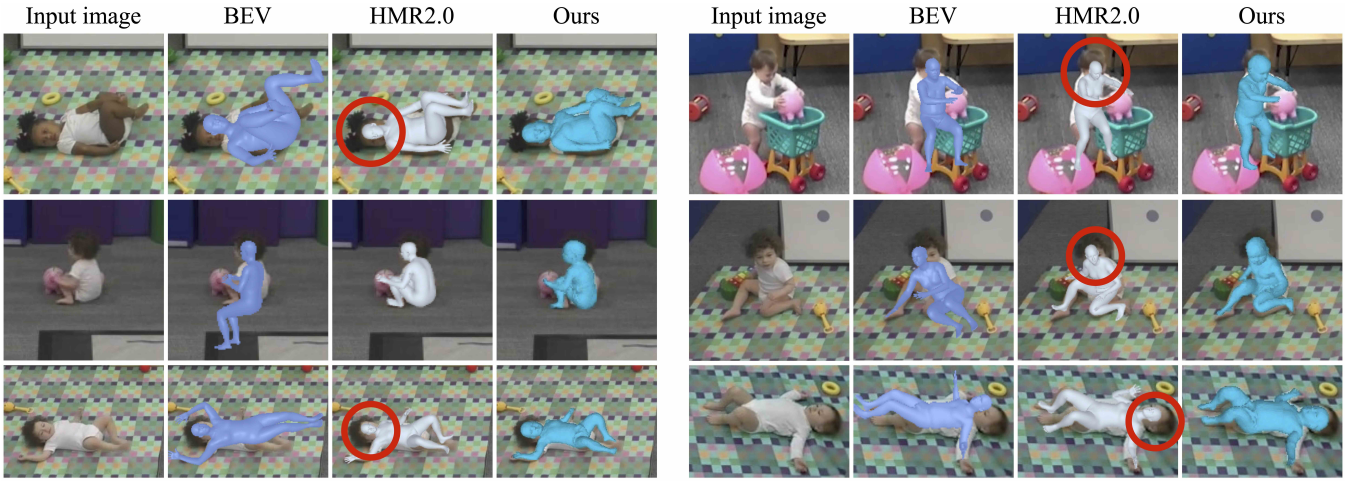

Fig. 4. **Qualitative comparison.** We visualize the results of various approaches on the Han *et al.* [13] dataset. Specifically, we show results and limitations from popular state-of-the-art methods for human mesh recovery, BEV [26] and HMR2.0 [9]. BEV can also reconstruct baby body shapes but fails in most cases to recover accurate body pose or shape. HMR2.0 captures the body pose, but can only model adult body types, which corresponds to adult body proportions. *Please note the discrepancy specifically to the head size*. Our approach recovers faithful body pose and shape for the baby in various postures.

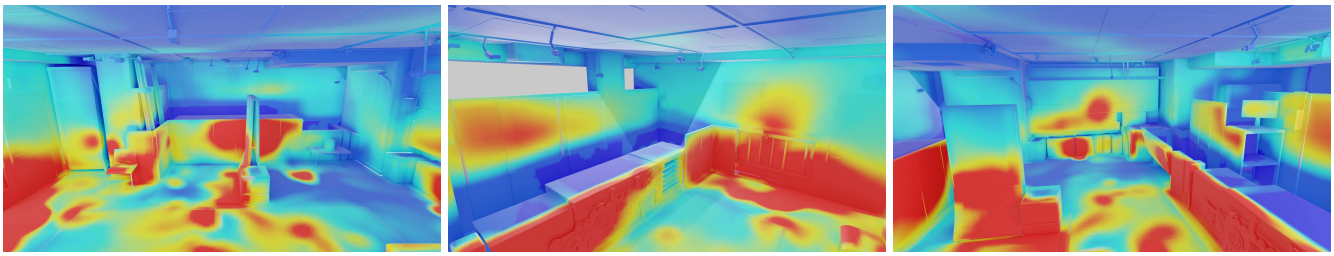

Fig. 5. **Visual attention map.** We show the accumulated attention heatmap from a 10-min session of a walking baby. We visualize three viewpoints of the digital twin of the recording environment (see supplementary video for full visualization). Warmer colors (red/yellow) indicate surfaces that received more visual attention, and cooler colors (blue) indicate regions that were rarely attended. Across all views, attention is concentrated at the baby's eye height, e.g., on the sides of low furniture, play structures, and nearby objects, whereas upper walls and the ceiling remain largely unattended.

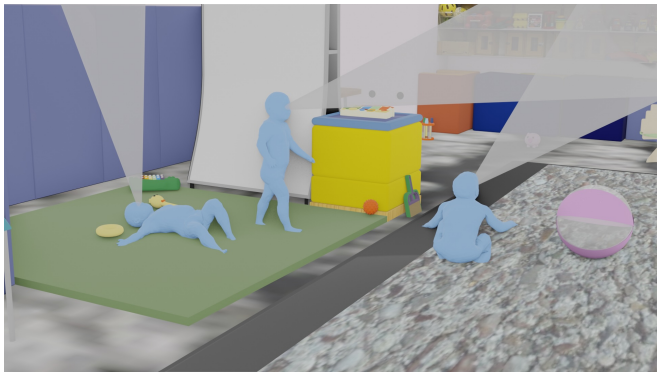

Fig. 6. **Visual attention cone.** We place the recovered baby mesh within a digital twin of the recording environment. Using the estimated head pose, we define a cone representing the infant's field of view. We show three instances of the baby in different postures, i.e., lying, squatting, and upright.

In each case, the cone projects from the head into the scene, intersecting the room geometry, which allows us to identify the surfaces that fall within the baby's field of view.

Attention maps. By accumulating the cone-intersected surface points over a full 10-minute session and rendering their density as a heatmap on the room geometry, we can obtain a spatial map of visual attention across the entire recording. We show such a result for one of the sessions in Figure 5. The heatmaps reveal a consistent pattern: Walking infants direct their attention to surfaces near their own eye height, such as the sides of low furniture, shelves, and nearby toys, whereas upper walls and ceilings receive little visual attention.

V. Summary and Future Work

In this paper, we present a pipeline for reconstructing 3D meshes of babies from calibrated and synchronized multi-view video. Previous work has either focused only on adults or relied on specialized pipelines tailored to specific attributes (e.g., wrist pose, head pose, posture), limiting the scope of analysis. In contrast, we propose a complete 3D reconstruction of the baby's body, enabling a wide range of downstream analyses relevant to developmental psychology and beyond. We evaluate our 3D reconstructions on multiple downstream tasks, including posture identification and gait (step) identification. Our results demonstrate that using our reconstructed 3D representation consistently outperforms alternative 3D representations or specialized algorithms, highlighting the benefits of a unified, model-based approach.

This work provides critical new avenues for future research and applications. For researchers in developmental psychology and clinical science, our approach can enable fully automated data processing of infant video, greatly accelerating the analysis of infant behavior without requiring laborious human annotation. Moreover, our approach provides data on 3D body and joint position—including infants' locomotor paths and gait parameters that are beyond the ability of human video annotation. Such a system facilitates understanding of infant motor behavior—replete with all the variability and idiosyncrasies of spontaneous, natural behavior in typically developing infants. It also promotes tools for

early screening and evaluation of therapeutic interventions for infants with atypical development.

Future work could leverage our reconstructions as high-quality training data for feedforward models that directly regress baby pose and shape from images or video, in the style of adult-focused methods [9], [11]. Additionally, our results point to the limitations of current parametric models, i.e., SMPL-A, which lead to improvements over adult-only models, but they do not perfectly match the proportions of babies (see Figure 4). Finally, we envision that our dataset could benefit other fields, such as robotics, by informing algorithms for teaching walking skills to robots. We plan to publicly release our code and processed data (3D reconstructions) to facilitate further work in these directions.

Acknowledgements: This work was supported by NICHD Grant R01HD033486 and DARPA Grant N66001-19-2-4035 to KEA and NSFC Grant 32400886 to DH. GP was supported by NSF-2504906, and 2544200; gifts from Adobe and Google; and computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. This work has partly taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-24-1-2550), ARO (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork) Lockheed Martin, and UT Austin's Good Systems grand challenge. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] Karen E Adolph and Justine E Hoch. Motor development: Embodied, embedded, enculturated, and enabling. *Annual review of psychology*, 70(1):141–164, 2019.

[2] Christina M Hospodar and Karen E Adolph. The development of gait and mobility: Form and function in infant locomotion. *Wiley Interdisciplinary Reviews: Cognitive Science*, 15(4):e1677, 2024.

[3] Regina T Harbourne and Nicholas Stergiou. Nonlinear analysis of the development of sitting postural control. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 42(4):368–377, 2003.

[4] Sandra L Saavedra, Paul van Donkelaar, and Marjorie H Woollacott. Learning about gravity: segmental assessment of upright control as infants develop independent sitting. *Journal of Neurophysiology*, 108(8):2215–2229, 2012.

[5] Do Kyeong Lee, Whitney G Cole, Laura Golenia, and Karen E Adolph. The cost of simplifying complex developmental phenomena: A new perspective on learning to walk. *Developmental science*, 21(4):e12615, 2018.

[6] Kari S Kretch, Natalie A Koziol, Emily C Marcinowski, Audrey E Kane, Ketaki Inamdar, Elena Donoso Brown, James A Bovaird, Regina T Harbourne, Lin-Ya Hsu, Michele A Lobo, et al. Infant posture and caregiver-provided cognitive opportunities in typically developing infants and infants with motor delay. *Developmental Psychobiology*, 64(1):e22233, 2022.

[7] Susan K Patrick, J Adam Noah, and Jaynie F Yang. Developmental constraints of quadrupedal coordination across crawling styles in human infants. *Journal of Neurophysiology*, 107(11):3050–3061, 2012.

[8] Karen E Adolph and Christina M Hospodar. Motor development. In *Developmental Science*, pages 234–269. Routledge, 2024.

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[10] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023.

[11] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024.

[12] Christina M Hospodar, Tieqiao Wang, Yasmine Elasmar, Sinisa Todorovic, and Karen E Adolph. Infant gait modifications: Integration of computer vision with human annotation provides accurate step classification and location as infants navigate varied terrain. In *ICDL*, 2024.

[13] Danyang Han, Nicolas Aziere, Tieqiao Wang, Ori Ossmy, Ajay Krishna, Hanzhi Wang, Ruiting Shen, Sinisa Todorovic, and Karen Adolph. Infants' developing environment: Integration of computer vision and human annotation to quantify where infants go, what they touch, and what they see. In *ICDL*, 2024.

[14] Nikolas Hesse, Sergi Pujades, Michael J Black, Michael Arens, Ulrich G Hofmann, and A Sebastian Schroeder. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *PAMI*, 42(10):2540–2551, 2019.

[15] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148–168, 1985.

[16] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[19] 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 *ICCV*, 2019.

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[21] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *CVPR*, 2022.

[22] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022.

[23] Lennart Jahn, Sarah Flügge, Dajie Zhang, Luise Poustka, Sven Bölte, Florentin Wörgötter, Peter B Marschik, and Tomas Kulvicius. Comparison of marker-less 2D image-based methods for infant pose estimation. *Scientific Reports*, 15(1):12148, 2025.

[24] 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 *CVPR*, 2019.

[25] 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 *CVPR*, 2021.

[26] 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 *CVPR*, 2022.

[27] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[28] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[29] Anshul Shah, Shlok Mishra, Ankan Bansal, Jun-Cheng Chen, Rama Chellappa, and Abhinav Shrivastava. Pose and joint-aware action recognition. In *WACV*, 2022.

[30] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3D pose and tracking for human action recognition. In *CVPR*, 2023.

[31] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *CVPR*, 2023.

[32] Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, and Sarah Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *WACV*, 2023.

[33] Elaheh Hatamimajoumerd, Pooria Daneshvar Kakhaki, Xiaofei Huang, Lingfei Luan, Somaieh Amraee, and Sarah Ostadabbas. Challenges in video-based infant action recognition: A critical examination of the state of the art. In *WACV*, 2024.

[34] Xiaofeng Han, Diego Guffanti, and Alberto Brunete. A comprehensive review of vision-based sensor systems for human gait analysis. *Sensors*, 25(2), 2025.

[35] Qingkai Zhen, Yongqing Liu, Qi Hu, and Qi Chen. Video based accurate step counting for treadmills. *Proceedings*, 2:301, 02 2018.

[36] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 4:5–21, 1987.
[37] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.
[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.