Bachelor Project

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

3D Body Pose Estimation of Infants from RGB Images and Videos

Vojtěch Ježek

Supervisor: doc. Mgr. Matěj Hoffmann, Ph.D.
Supervisor–specialist: Ing. Filipe Gama
Field of study: Cybernetics and Robotics
May 2024

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Ježek Vojtěch**	Personal ID number:	**498988**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Cybernetics**		
Study program:	**Cybernetics and Robotics**		

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

3D Body Pose Estimation of Infants from RGB Images and Videos

Bachelor’s thesis title in Czech:

3D odhad polohy těla kojenců z RGB obrázků a videí

Guidelines:

Automatic extraction of infant poses and motion trajectories from videos is important in developmental and clinical psychology. The movement patterns extracted can be used for automatic diagnosis of psychomotor disabilities, such as cerebral palsy, without the need for bringing infants physically to the therapist. With rapid advances in computer vision, estimation of 2D and even 3D pose and shape from only RGB images is possible for adults (e.g., ViTPose [1] for 2D pose; TRACE [2] or 4DHumans [3] for 3D pose) and, sometimes, infants [4]. Current limitations of these methods include: (i) absence of model of infant shape (e.g., [2], [3]), (ii) repulsion of self-contact configurations (e.g., [5] for a possible solution).
The goal of this thesis is to evaluate the accuracy of state-of-the-art methods for 3D pose from RGB videos on infant datasets and then propose modifications of the existing algorithms and assess their performance.
1. Study the relevant literature and get state-of-the-art 3D human pose estimation algorithms to run. There are two main approaches.
a. 2D keypoints from image (e.g., [1]) and then 3D shape estimation (SMPL [6] or 4D Humans [3] for adults; SMIL [4]).
b. One-stage 3D pose estimation ([2], [7], [8]).
2. Compare the performance of these methods on infant data (using synthetic dataset [9] and recordings of real infants provided by the supervisor) using various metrics (e.g., Mean Per Joint Position error).
3. Based on the results of 2., identify the main sources limiting the performance and propose and implement modifications of the pose estimation pipelines(s). These may include:
a. Fine-tuning the best performing 2D pose estimation network [1] on infant data.
b. Investigate the possibility of incorporating the infant shape model (SMIL) into recent two-stage 3D pose estimation networks [3]. This may require retraining the model.
c. Investigate the possibility of retraining the one-stage 3D pose estimation networks [2], [7] on infant data.
d. Explore the possibilities of not penalizing self-collision in the cost functions for 3D pose estimation (following [5]).
4. Provide, evaluate, and document the best performing solution(s) to the problem of 3D pose estimation of infants in videos.

Bibliography / sources:

[1] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “ViTPose++: Vision Transformer for Generic Body Pose Estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[2] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, “TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8856–8866.
[3] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, “Humans in 4D: Reconstructing and Tracking Humans with Transformers,” in International Conference on Computer Vision (ICCV), 2023.
[4] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, “Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2540–2551, Oct. 2020, doi: 10.1109/TPAMI.2019.2917908.
[5] L. Muller, A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, “On self-contact and human pose,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9990–9999.
[6] G. Pavlakos et al., “Expressive Body Capture: 3D Hands, Face, and Body from a Single Image,” in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.
[7] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, “Monocular, one-stage, regression of multiple 3d people,” presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11179–11188.

[8] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, “Putting people in their place: Monocular regression of 3d people in depth,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13243–13252.
[9] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. Sebastian Schroeder, “Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set,” presented at the Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0. Accessed: Nov. 19, 2023. [Online]. Available: https://openaccess.thecvf.com/content_eccv_2018_workshops/w31/html/Hesse_Computer_Vision_for_Medical_Infant_Motion_Analysis_State_of_the_ECCVW_2018_paper.html

Name and workplace of bachelor’s thesis supervisor:

doc. Mgr. Matěj Hoffmann, Ph.D. Vision for Robotics and Autonomous Systems FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Ing. Filipe Gama Vision for Robotics and Autonomous Systems FEE

Date of bachelor’s thesis assignment: **15.01.2024** Deadline for bachelor thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

doc. Mgr. Matěj Hoffmann, Ph.D. Supervisor’s signature	prof. Dr. Ing. Jan Kybic Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.
Date of assignment receipt

Student’s signature

Acknowledgements

Foremost, I would like to thank my supervisor–specialist, Filipe Gama, for all his guidance and persistent assistance with my experiments.

I could not have undertaken this journey without my supervisor, Matěj Hoffmann, who never hesitated to share valuable insights and always showed a positive approach to any obstacles.

Conducting the experiments would not be possible without all the expertise Lea Müller shared with me regarding pose estimation methods.

I am also grateful to Valentin Marcel for all the advice he gave me and for all his help with data preparation.

I must thank Lukáš Rustler, who was always ready to aid me and who knew how to improve my mood.

My deepest gratitude belongs to my family, especially my parents, sister, and brothers, who accompanied me and strongly supported me throughout writing this thesis.

Last but not least, I would like to thank all my friends and classmates for all their emotional support.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských prací.

V Praze dne 15. května 2024

....
Vojtěch Ježek

Abstract

Automated estimation of infant pose and subsequent analysis of infant motion carries great potential for early diagnostics of developmental disorders like cerebral palsy. This thesis compares six state-of-the-art methods to 3D pose estimation from videos (ROMP, BEV, TRACE, 4D Humans, and SMPLify-X with OpenPose and ViTPose) on sequences of images of infants, with adaptations of standard metrics—mean per joint position error, bone length standard deviation, and the number of missed detections. Surprisingly, the SMPLify-X model, which, in our case, fits an infant body (SMIL) to the images, is outperformed by 4D Humans, which uses an adult model (SMPL). The comparison confirms that methods tracking people across multiple video frames output bodies with better bone length stability. We retrain the best-performing method, 4D Humans, with a model that fits infant bodies (SMIL) on infant data. We show that the use of infant bodies improves the estimation of depth. We provide an outline of possible future improvements to the training process.

Keywords: Pose estimation, 3D body models, infants, RGB videos, machine learning.

Supervisor: doc. Mgr. Matěj Hoffmann, Ph.D.

Abstrakt

Automatizovaný odhad polohy kojence a následná analýza jeho pohybu má velký potenciál pro včasnou diagnostiku vývojových poruch, jako je dětská mozková obrna. Tato práce porovnává šest moderních metod odhadu 3D polohy z videozáznamů (ROMP, BEV, TRACE, 4D Humans a SMPLify-X s OpenPose a ViTPose) na sekvencích snímků kojenců s úpravami standardních metrik – střední chyba polohy na kloub, směrodatná odchylka délky kosti a počet chybných detekcí. Překvapivě model SMPLify, který v našem případě používá model s tělem kojence (SMIL), je překonán modelem 4D Humans, který používá model dospělého člověka (SMPL). Srovnání potvrzuje, že metody sledující osoby na více videosnímcích produkují odhady tvaru a polohy těla s lepší stabilitou délky kostí. Nejúspěšnější metodu 4D Humans jsme přetrénovali na datech kojenců, aby používala model, který odpovídá tělům kojenců (SMIL). Ukazujeme, že použití dětských těl zlepšuje odhad hloubky. Uvádíme nástin možných budoucích vylepšení procesu trénování metody 4D Humans.

Klíčová slova: Odhad polohy těla, 3D modely těla, kojenci, RGB videa, strojové učení.

Překlad názvu: 3D odhad polohy těla kojenců z RGB obrázků a videí

Contents

Chapter 1

Introduction

Infant motion analysis is essential for understanding infant development over the course of time. Various research fields take an interest in infants' movements. For instance, in developmental psychology, it was found that during the first weeks of life, infants develop the ability to identify distinct entities in their environment. One example is that self-produced leg movements contribute to the development of infants' ability to be aware of their bodies [1]. In healthcare, infant motion analysis can serve to identify deviations from the infants' typical development. Trained professionals can assess the risk of early developmental disorders, such as cerebral palsy, using validated examinations like General Movement Assessment (GMA) [2] that evaluates spontaneous infantile movements or through Hammersmith Infant Neurological Examination (HINE) [3] that is based on tracking infants' movement, posture, and reflexes. However, the need for a trained expert for the evaluation creates a notable hurdle toward accessible early detection of developmental disorders.

Automated assessment of movement patterns is crucial for making these examinations accessible at a large scale with a minor cost. The data intended for the assessment can be obtained by various approaches. The gold standard in terms of accuracy is the use of motion capture systems. It has, however, severe drawbacks for employment on a larger scale. Recording with motion capture is limited to in-laboratory use, which excludes a lot of possible scenarios for infant observation. Another inconvenience is the need to place physical markers on the bodies of infants, which may affect the infants' behavior. Motion capture systems also require substantial initial financial investment. RGB-D cameras or multi-camera setups are another option for capturing the needed data. Contrary to the motion capture systems, they are cheaper and not reliant on in-laboratory use.

This thesis dips into the analysis of infant data acquired from single RGB videos. These videos can be recorded with everyday devices like smartphones. Therefore, this approach is available at a much lower cost than the aforementioned ones and can be used in any environment. The needed information could be either extracted by manually labeling it in each frame (which is time-consuming and expensive) or with machine learning-based methods. The literature indicates that spontaneous movements can be quantitatively analyzed from videos using deep learning-based methods [4].

One way that enables automatic motion analysis from videos is through human pose estimation. Its methods increased under rapid development their performance over the last years, and while being originally developed for adults, our interest lies in the analysis of infants. However, the unique morphology of infants poses challenges due to the different body proportions from those of an adult. The goal of this thesis is, therefore, to compare state-of-the-art 3D pose estimation methods from RGB videos on recordings of infants, indicate their shortcomings and propose and implement modifications to the most promising method. This includes retraining the model on infant data.

1.1 Structure of the Thesis

First, in Chapter 2, we list an overview of pose estimation methods from single videos. We start with two 2D pose estimation methods. Then we describe two currently used skinned models for fitting 3D bodies (of adults and infants) in images. The 3D pose estimation methods (models) to predict the needed parameters for the skinned models are discussed afterward.

In Chapter 3, we first describe datasets for the latter comparison of methods introduced previously in Chapter 2. This chapter also includes sections that include specifics of running the introduced models and retraining the best-performing one. We introduce this thesis's adaptation of multiple comparison metrics for the models.

Chapter 4 includes results of two experiments. The first compares the 3D pose estimation methods on image sequences of infants and determines which method to explore further, *i.e.*, retrain it on infant data. The second evaluates newly trained models.

We conclude the experiments in Chapter 5. We discuss the results and outline possible future paths to follow.

Chapter 2

Related Work

2.1 2D Pose Estimation

Keypoints in this work designate significant joints or spots that represent body parts (further discussed in Section 3.4), and their combination carries information about the pose. Estimating the position of keypoints in 2D is an essential step for some 3D pose estimation methods.

2.1.1 OpenPose

OpenPose [5] uses a bottom-up approach to estimate 2D keypoints. This approach determines each keypoint (or a group of keypoints) independently on the whole image before joining them to form each individual. In this way, the estimator does not rely on person detectors. OpenPose is based on convolutional neural network architecture (CNN). It was trained on COCO [6] and MPII [7] datasets.

2.1.2 ViTPose

The current state-of-the-art method, ViTPose [8], uses a top-down approach (although it is possible for ViTPose to adopt the bottom-up approach): it first uses a person detector and then estimates the poses of the individuals independently on the surroundings. Instead of CNNs, ViTPose employs vision transformers. It was trained on COCO [6], AIC [9] and MPII [7] datasets. A novel model, ViTPose++ [10], was recently introduced to extend ViTPose's capability to deal with heterogenous body keypoints.

2.2 3D Skinned Linear Models

2.2.1 Skinned Multi-Person Linear Model (SMPL)

Skinned Multi-Person Linear Model (SMPL) [11] is a skinned vertex-based model used to create 3D adult body models in various postures. The SMPL template consists of 6890 vertices and 23 body joints. SMPL represents shape deformations as a combination of a pose-dependent shape (pose blend shape) and an identity-specific shape (shape blend shape). The SMPL shape space is defined by a mean template shape and principal shape directions that are computed by *principal component analysis* (PCA) on pose-normalized meshes. The shape is then described by a vector of linear coefficients β that modify the template shape in the principal directions. Each body joint has 3 degrees of freedom (DOF), resulting in 69 parameters + 3 parameters for global rotation, *i.e.*, 72 pose parameters θ.

SMPL model parameters were trained to minimize the vertex reconstruction error. The pose blend shapes were learned from *multi-pose* dataset with 1786 registered meshes of 40 individuals in predefined positions, and shape blend shapes were learned on *multi-shape* dataset with 1700 male and 2100 female meshes from the CEASAR dataset [12].

SMPL-X [13] further extends the capabilities of SMPL with articulated hands and facial expressions.

2.2.2 Skinned Multi-Infant Linear Model (SMIL)

The body proportions of an adult are distinctively different from the ones of an infant. Therefore, using the SMPL model to estimate an infant's shape is insufficient (see Fig. 2.1). Hesse et al. proposed *Skinned Multi-Infant Linear Model* (SMIL) [14] to modify SMPL to fit the needs of infant pose estimation. SMIL was created by adapting SMPL as SMPL$_\mathrm{B}$ and then registering it to preprocessed data consisting of many low-quality RGB-D images. From these registrations, a new infant-specific shape space and a new pose prior were learned to form the final SMIL.

2.3 3D Pose Estimation Methods

There are two main approaches to developing models to estimate the 3D pose of humans from a single RGB video (see Fig. 2.2). The first generally consists of two stages—2D keypoints estimation and 3D lifting. 2D keypoints estimation is usually performed by an independently developed model, such as ViTPose [8], OpenPose [5], etc. Lifting from 2D to 3D is then performed by an optimization process specified by the model in question, in our case SMPLify-X [13] and HMR2.0 [16], that predicts input parameters to either SMPL [11] or SMIL [14] model.

The second approach skips the first stage of estimating keypoints in 2D and directly estimates the SMPL or SMIL parameters from various feature maps. For this category VIBE [17], ROMP [18], BEV [19], and TRACE [20] are considered in this work. An overview of all the methods used is in Table 2.1.

	one/two-stage approach	3D model	tracking
ROMP	one	SMPL	no
BEV	one	SMPL+A	no
TRACE	one	SMPL	yes
4D Humans	two	SMPL	yes
SMPLify-X	two	SMIL (SMPL)	no

Table 2.1: Overview of the used models. SMPLify-X was originally introduced with SMPL; in this work, we utilized its version with SMIL.

Some of the methods from both approaches can track multiple people across multiple frames. This tracking feature comes into use within our purpose mainly on two occasions.

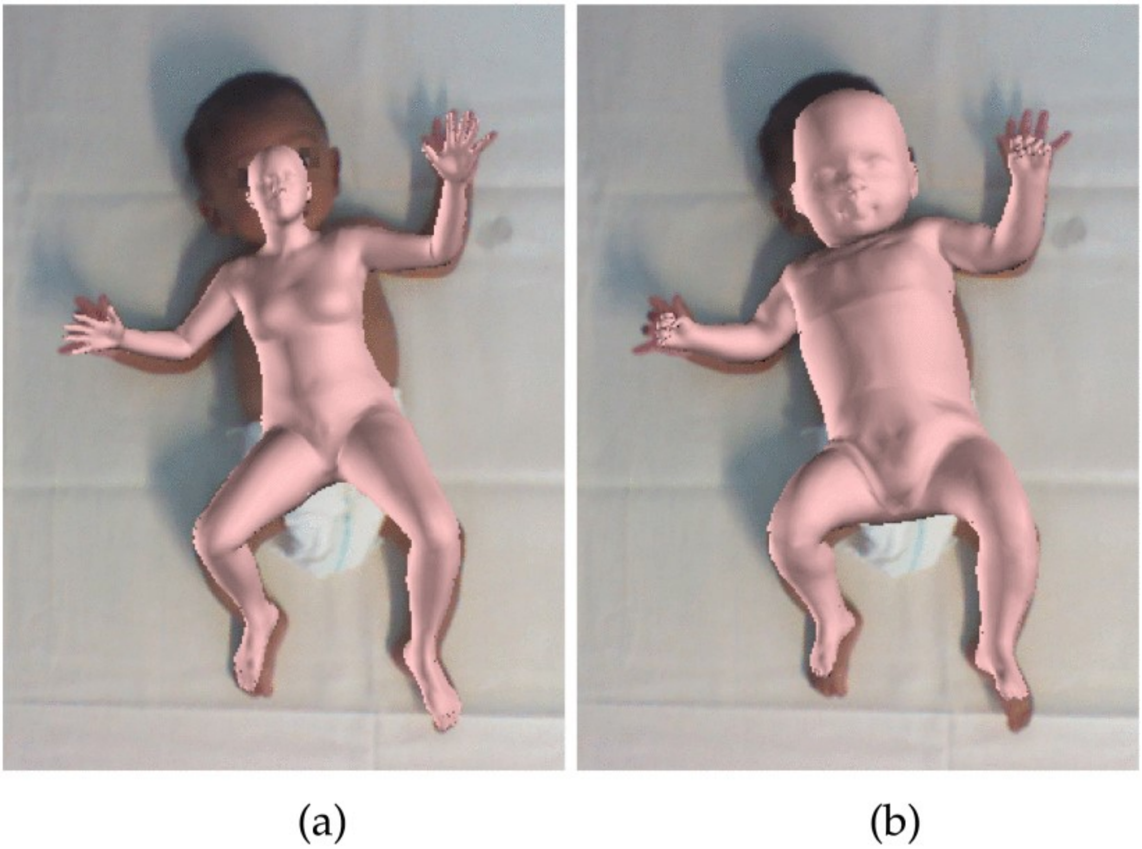


Figure 2.1: (a) Scaling the SMPL adult body model and fitting it to an infant does not work as body proportions significantly differ. (b) The proposed SMIL model properly captures the infants' shape and pose. Both the figure and caption taken from [15].

First, when an adult person would be present in the frame to manipulate or interact with the infant, the tracking helps maintain the identities of the two. Second, to ensure better temporal consistency of the estimations, for example, by reducing the need to reestimate the shape and camera translation parameters or increasing the coherence of the position estimation between two continuous frames.

2.3.1 SMPLify-X

SMPLify-X relies on an optimization process to predict SMPL-X parameters that minimizes an objective function that consists of these terms: priors for body pose, facial pose, facial expressions, and hand pose to penalize deviation from the neutral state, prior to penalizing extreme bending of elbows and knees, interpenetration penalty, variational-autoencoder-based body pose prior, and the data term. The data term relies on detecting 2D keypoints from OpenPose by default, although using different models for 2D keypoint estimation is possible. When using SMIL for this method instead of SMPL-X, the extra hands and face keypoints are not provided as SMIL does not support them. In this work, we used SMPLify-X with SMIL model in two configurations—with OpenPose and ViTPose. SMPLify-X was trained on the following datasets: Human3.6M [21], CMU [22], and PosePrior [23].

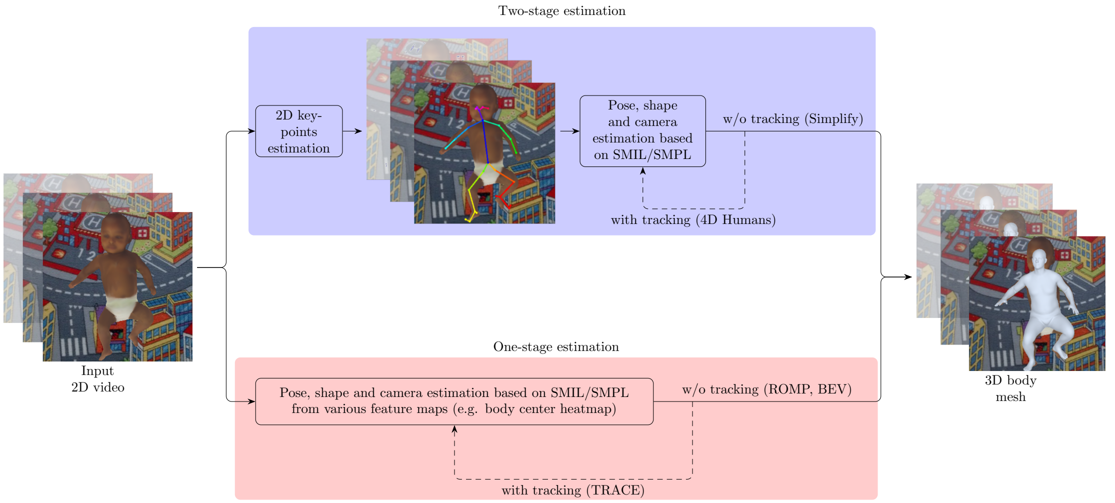

Two-stage estimation
2D key-points estimation
Pose, shape and camera estimation based on SMIL/SMPL
w/o tracking (Simplify)
with tracking (4D Humans)
Input 2D video
3D body mesh
One-stage estimation
Pose, shape and camera estimation based on SMIL/SMPL from various feature maps (e.g. body center heatmap)
w/o tracking (ROMP, BEV)
with tracking (TRACE)

Figure 2.2: Schematic of the two approaches to estimating the 3D body mesh from RGB videos.

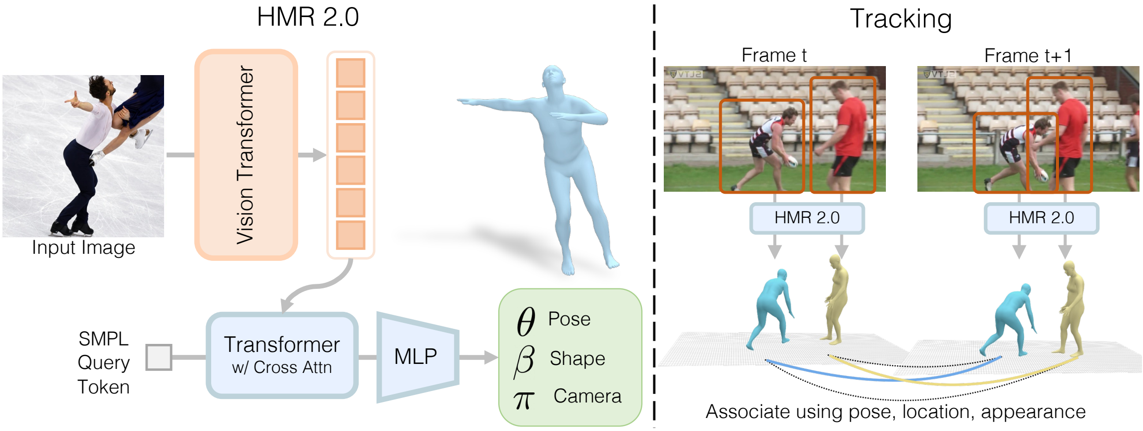


Figure 2.3: Overview of the 4D Humans framework, figure taken from [16].

2.3.2 4D Humans & HMR 2.0 Network

4D Humans [16] is the state-of-the-art approach for reconstructing and tracking humans in 4D, with fourth dimension being the time. It employs ViTPose as the backbone for 2D keypoint estimation. The network HMR 2.0 for human mesh recovery is then introduced to estimate SMPL shape and pose parameters along with global orientation. The body mesh reconstructions then serve as an input to the tracking system. For an overview of the 4D Humans approach, see Fig. 2.3.

4D Humans was trained on the following datasets: Human3.6M [21], MPI-INF-3DHP [24], COCO [6], MPII [7], and to generate extra pseudo-ground-truth fits: InstaVariety [25], AVA [26], and AIC [9].

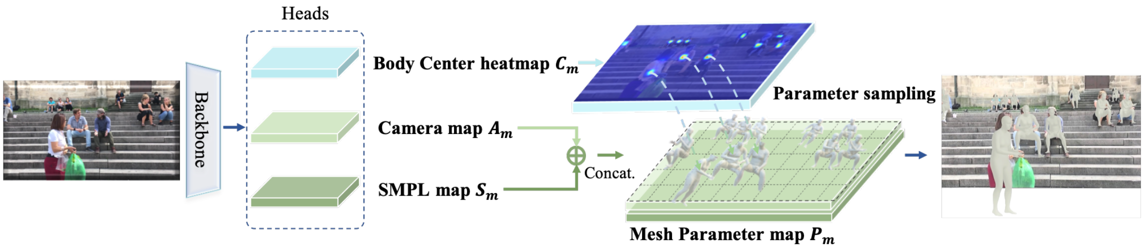


Figure 2.4: Overview of the ROMP model, figure taken from [18].

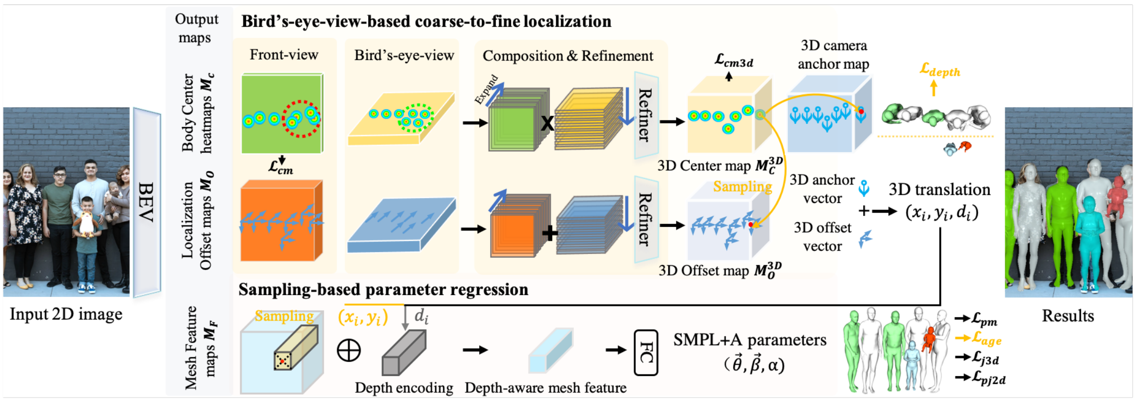


Figure 2.5: Overview of the BEV model, figure taken from [19].

2.3.3 ROMP

ROMP [18] uses three head networks. These output Body Center heatmap, Camera map, and SMPL map. By combining the camera map and the SMPL map and sampling them at the 2D body center locations, the SMPL parameters are extracted. For an overview of the ROMP model, see Fig. 2.4. ROMP was trained on the following datasets: Human3.6M [21], MPI-INF-3DHP [24], UP [27], COCO [6], MPII [7], LSP [28], and AIC [9].

2.3.4 BEV

BEV [19] builds on top of ROMP's heatmaps. It adds a heatmap representing an unseen bird's-eye-view. This aims to achieve a better estimation of the depth of corresponding subjects. Further, BEV detects if a target is an adult or an infant. For that purpose, BEV introduces SMPL+A. As SMPL and SMIL models parameterize body meshes of adults and infants, respectively, SMPL+A was introduced [29] to linearly blend SMPL and SMIL template shape, with age offset $\alpha \in [0, 1]$. For an overview of the BEV model, see Fig. 2.5. BEV was trained on the following datasets: Human3.6M [21], MuCo-3DHP [30], COCO [6], MPII [7], LSP [28], and CrowdPose [31].

2.3.5 TRACE

TRACE [20] uses an image and motion backbone to extract temporal feature and optical flow maps. These serve as an input to a four-head network that performs detection, tracking,

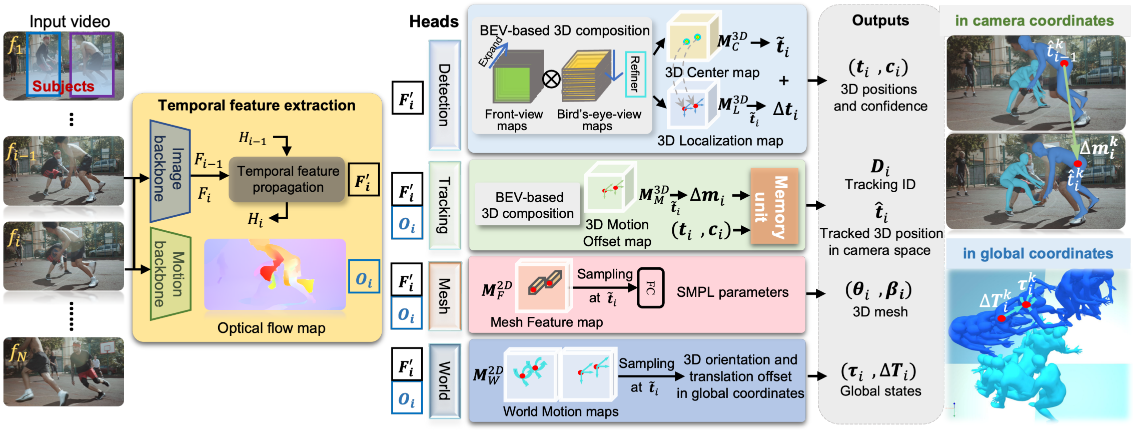


Figure 2.6: Overview of the TRACE model, figure taken from [20].

mesh parameter estimation, and global coordinates estimation. For an overview of the TRACE model, see Fig. 2.6. TRACE was trained on the following datasets: Human3.6M [21], MPI-INF-3DHP [24], 3DWP [32], PennAction [33], CrowdPose [31], and DynaCam [20].

2.4 Thesis Contribution

In most cases, the current models use the adult model, SMPL, and are trained on datasets focused on adults. In this work, we compare a selection of 3D pose estimation methods on infant videos, *i.e.*, ROMP, BEV, TRACE, 4D Humans, and SMPLify-X with OpenPose and ViTPose. Based on the comparison results, we identify the methods' current limitations and retrain the best-performing method—4D Humans—on infant data. We evaluate our retrained models and propose paths for future improvements.

Chapter 3

Materials and Methods

3.1 Datasets

There are very few publicly available datasets of infant recordings with 2D and 3D annotation, *i.e.*, with 2D and 3D keypoints for evaluation and SMIL parameters for training. Therefore, we use datasets with images of synthetic infants MINI-RGBD [34] and SyRIP [35]. SyRIP also provides some images of real infants. Then, we use one In-Lab recording of an infant.

3.1.1 MINI-RGBD

The MINI-RGBD [34] dataset consists of 12 artificially generated sequences. Each sequence represents a different synthetic infant with 1000 images. The pose and shape of each infant were extracted by registering SMIL to RGB-D recordings of real infants. The authors proposed that the sequences can be sorted into three groups according to difficulty: a) easy: simple movements of limbs, lying in a supine position (sequences 1-4); b) medium: more complex movements of limbs—self touches and legs crossings (sequences 5-9); and c) hard: limbs pointing towards the camera, grabbing legs with hands, infant turning to sides (sequences 10-12). The infants' movement should correspond to the development of motor skills in the first seven months of life.

The dataset's annotations that we use as the ground truth for this project are RGB images, pose and shape parameters, and 2D and 3D keypoints. Keypoints for ears and eyes, which are essential for head orientation, are missing in this dataset (see Fig. 3.3d).

3.1.2 SyRIP

SyRIP [35] dataset provides two sets of images—one for training and the second for validation. The training dataset consists of 1000 images of synthetic infants on a natural background and 200 images of real infants collected from Google Images and YouTube. The validation dataset contains 500 images of real infants collected from Google Images and YouTube.

The dataset presents the annotated information in JSON files, following the COCO format[1]. The dataset's information that we can use as the ground truth for this project are RGB images and the 2D keypoints.

[1] `https://cocodataset.org/#format-data`

3.1.3 In-Lab Recorded Infant Data

Our team recorded a video of a 36-week-old infant from different angles using two RealSense D435 cameras and one Miqus-RGB camera (the setup comes from [36]). The frames needed to be manually synchronized, as the Miqus camera uses a regular 25 Hz frame rate, and the RealSense cameras give temporal stamps to each frame. The cameras were calibrated to determine extrinsic parameters by taking nine images of a known pattern from all three cameras in different configurations. These extrinsic parameters hold information about camera translation and rotation in the world. Intrinsic parameters are then the focal lengths of the lenses and sizes of the pixels' sensors. Taking inspiration from the training datasets of 4D Humans [16], pseudo-ground-truth fits were generated for our data: The first step in 3D keypoint estimation was to estimate 2D keypoints with ViTPose from each camera. To find the 3D position of the keypoints, triangulation methods were used with known extrinsic and intrinsic parameters from calibration [37].

The video of the infant in a supine position consists of 3418 RGB images. All of those are annotated with VitPose. However, only 1285 are annotated in 3D due to the lower frame rate of the other cameras.

Experiments with the infant recordings and the disclosure of the images were approved by the Committee for Research Ethics at the Czech Technical University in Prague under the reference number 0000-02/23/51902/EKČVUT on March 14, 2023.

3.2 Setting Up the Models for Comparison

We created Docker images with the ROMP, BEV, and TRACE installations. ROMP and BEV share a simplified implementation that enables running both models after its setup. This particular implementation is distributed under version 2.0 on the models' GitHub[2], including the model data and the source code. TRACE shares the same GitHub page with ROMP and BEV but is distributed under version 3.0. These three models need the SMPL 1.1.0 for Python 2.7, and BEV further needs the SMIL model. Those can be downloaded from the official SMPL and SMIL websites[3][4]. The installation guide with the needed dockerfiles is available on our project's GitLab[5]. Built docker images are accessible at the Humanoid and Cognitive Robotics Group's Docker Hub[6].

4D Humans was installed in a conda environment after the commit 124e8b2 on the project's GitHub[7]. For the 4D Humans, we have modified the inference script. That, along with the specifics of the installation, is also accessible on our project's GitLab[5].

Our team conducted experiments with SMPLify-X in the past [38], and all the data

[2] https://github.com/Arthur151/ROMP

[3] https://smpl.is.tue.mpg.de/

[4] https://www.iosb.fraunhofer.de/en/competences/image-exploitation/object-recognition/sensor-networks/motion-analysis.html

[5] https://gitlab.fel.cvut.cz/body-schema/baby-keypoint-extraction/code-3d-keypoints-and-models

[6] https://hub.docker.com/u/humanoidsctu

[7] https://github.com/shubham-goel/4D-Humans

needed for our comparison was already extracted before starting this thesis, along with creating its Docker. Therefore, this process was not in the scope of this thesis. The SMPLify-X Docker is currently not publicly available.

3.3 Retraining HMR 2.0 to estimate SMIL parameters

We retrained HMR 2.0 from scratch. SMIL is implemented in the same way as SMPL. Therefore, we just replaced the SMPL model `.pkl` file with the one of SMIL. However, the original implementation of SMPL (or 4D Humans) uses only the first 10 shape parameters from the PCA space. Because the data used for training includes 20 shape parameters, we extended the number of shape parameters to 20. That led to a few changes in the overall implementation code, listed on our project's GitLab[8].

3.3.1 Loss Functions

By default, 4D Humans employs four different loss functions in training the predictor, which outputs SMIL parameters $\Theta = [\theta, \beta, \pi]$ [16]. First is the parameter loss, which is computed as MSE loss, using ground-truth SMIL parameters—shape β^* and pose θ^*:

$$\mathcal{L}_{\text{smil}} = \|\theta - \theta^*\|_2^2 + \|\beta - \beta^*\|_2^2. \tag{3.1}$$

Second, 4D Humans uses a 3D keypoints loss, which is obtained by L1 (Manhattan) norm for predicted keypoints X and the annotation X^* in the 3D space:

$$\mathcal{L}_{\text{3D}} = \|X - X^*\|_1. \tag{3.2}$$

During our experiment, the $\mathcal{L}_{\text{3D}}$ is left out from training as the training data either did not have 3D annotations or it did have them, but we did not manage within the project's timeframe to properly convert the 3D keypoints to the reference frame corresponding to the one needed by HMR 2.0. Consequences of using $\mathcal{L}_{\text{3D}}$ with incorrectly aligned reference frames can be seen in Fig. 3.1.

The third loss is computed using the L1 norm between the 2D annotated keypoints x^* and the projections of predicted 3D keypoints $\pi(X)$:

$$\mathcal{L}_{\text{2D}} = \|\pi(X) - x^*\|_1. \tag{3.3}$$

The last loss is the generator loss. Using the adversarial prior, the discriminator D_k is trained to ensure realistic 3D poses for each part of the infant model. The SMIL parameters are split into shape parameters β, body pose parameters θ_b, and per-part relative rotations θ_i.

$$\mathcal{L}_{\text{adv}} = \sum_k (D_k(\theta_b, \beta) - 1)^2. \tag{3.4}$$

[8] `https://gitlab.fel.cvut.cz/body-schema/baby-keypoint-extraction/code-3d-keypoints-and-models`

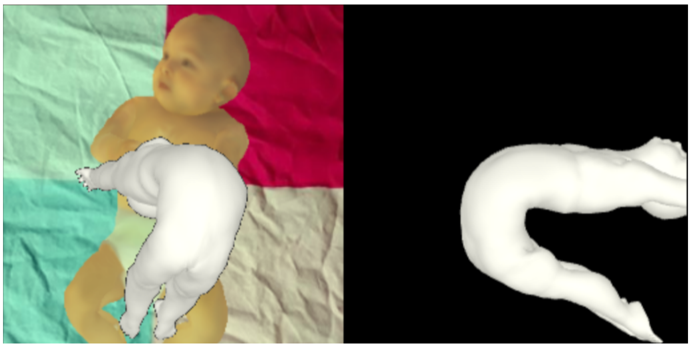


Figure 3.1: When training with the $\mathcal{L}_{3D}$, estimated infants' pose was extremely bent. The most likely cause is incorrect alignment of the 3D ground-truth annotation reference frame and the 4D Humans's internal reference frame.

3.3.2 SMIL Mean Parameters for Initialization

SMPL used in 4D Humans needs to have the parameters for SMPL decoder head initialized. The mean SMPL parameters are loaded from an `.npz` file for that purpose. To accommodate SMIL, it is needed to change those values to ones compatible with the SMIL shape space. We averaged the shape parameters from the 12 MINI-RGB sequences to form a new mean shape parameter vector. These new shape parameters should represent the average synthetic infant and are used as the initial parameters for the SMPL decoder head.

The mean pose parameters were not changed as their meaning is the same for both SMPL and SMIL (although the mean pose of an adult varies from the one of an infant, it should not influence the results by a large margin after the inference).

3.3.3 Replacing CMU Mocap Dataset for Training

The discriminator in the training of HMR 2.0 uses the CMU Mocap dataset [22] to estimate the adversarial loss Eq. (3.4). CMU Mocap provides approximately 300,000 SMPL body shape parameters β and corresponding pose parameters θ. However, the poses of an infant differ from those of an adult. Furthermore, the shape parameters are also different for SMIL. Therefore, we substituted the dataset with the shape and pose parameters from the MINI-RGBD dataset consisting of 12 different shapes, with 1000 poses each.

3.3.4 Training Parameters

We trained with learning rate 10^{-5} and weight decay of 10^{-4}. The weights assigned to each loss are described in Table 3.1. Due to the issue with the 3D error loss (Fig. 3.1), we decided to test training the model in different situations by disabling or enabling losses to confirm that 3D error loss was the cause of the issue and that the other losses led to improving the model.

	without $\mathcal{L}_{\text{adv}}$	with $\mathcal{L}_{\text{adv}}$
KEYPOINTS_2D	0.01	0.01
GLOBAL_ORIENT	0.001	0.001
BODY_POSE	0.001	0.001
BETAS	0.0005	0.0005
ADVERSARIAL	0.0	0.0005

Table 3.1: Configuration of weights for losses. GLOBAL_ORIENT denotes the first three parameters from body pose parameters.

3.4 Preparation of Data for Training and Evaluation

Each dataset provides information about keypoints in different ways. That is the same for the used models. All keypoint formats are listed in Table 3.2 and visualized in Fig. 3.3. Generally, the OpenPose 25 keypoints were used as the common format throughout this work due to the original use of this format in our group from SMPLify-X + OpenPose pose estimation [38]. Other formats often do not offer the same keypoints, reducing the final number of usable keypoints for each method. Occasionally, the names of the keypoints do not indicate the same position in different formats—what corresponds to shoulders in COCO17/OP18/OP25 is called UpperArm in MINI-RGBD. Shoulders in MINI-RGBD lie more inside the torso, closer towards the neck.

The keypoints, which are supposed to be the same, are sometimes placed in slightly different positions by the different methods (both for 2D and 3D), as in example Fig. 3.2. That might be caused by them having been trained on data annotated in different formats with different keypoint positioning guidelines. We can see in Fig. 3.3 that, *e.g.*, in MINI-RGB and OP25, the hips and the mid-hip keypoints are located in different places (MINI-RGBD format does not specifically have MidHip keypoints, although the global keypoint seems to be located in its position). SMPLify-X, for that reason, ignores the hip keypoints to calculate the loss when estimating the 3D pose.

3.4.1 WebDataset Format

To retrain HMR 2.0, we needed to set up the dataset in a structure compatible with the one used in the authors' original training scripts (accessible at the 4D Humans Github repo [16]), administered in WebDataset implementation of PyTorch Dataset. WebDataset stores the data in POSIX tar archives and uses sequential data access.

The tar archives contain training samples following the convention that the files belonging together have the same basename. The images are saved either with the `.jpg` or `.png` extension. The annotations are in `.data.pyd` files and hold the following mandatory information in a Python dictionary: 3D keypoints, 2D keypoints, betas (identity-dependent shape parameters), pose parameters, scale (dimensions of the infants' bounding box), and center (center of the bounding box).

kp ID \ format	COCO 17	OP 18	OP 25	MINI-RGBD
0	nose	nose	Nose	global
1	left_eye	neck	Neck	leftThigh
2	right_eye	right_shoulder	RShoulder	rightThigh
3	left_ear	right_elbow	RElbow	spine
4	right_ear	right_wrist	RWrist	leftCalf
5	left_shoulder	left_shoulder	LShoulder	rightCalf
6	right_shoulder	left_elbow	LElbow	spine1
7	left_elbow	left_wrist	LWrist	leftFoot
8	right_elbow	right_hip	MidHip	rightFoot
9	left_wrist	right_knee	RHip	spine2
10	right_wrist	right_ankle	RKnee	leftToes
11	left_hip	left_hip	RAnkle	rightToes
12	right_hip	left_knee	LHip	neck
13	left_knee	left_ankle	LKnee	leftShoulder
14	right_knee	right_eye	LAnkle	rightShoulder
15	left_ankle	left_eye	REye	head
16	right_ankle	right_ear	LEye	leftUpperArm
17		left_ear	REar	rightUpperArm
18			LEar	leftForeArm
19			LBigToe	rightForeArm
20			LSmallToe	leftHand
21			LHeel	rightHand
22			RBigToe	leftFingers
23			RSmallToe	rightFingers
24			RHeel	noseVertex

Table 3.2: Several formats are used by the aforementioned models listed in the table. To use data from different formats, it must be rearranged to align the keypoint correctly. That is, *e.g.*, right shoulder is designated by '6' in COCO 17, '3' in OpenPose 18 (OP 18) and OpenPose 25 (OpenPose 25), and '17' in MINI-RGBD format.

3.4.2 Configuration of Datasets For Training and Evaluation

We split the data into three categories: training, validation, and testing (as shown in Table 3.3). The training dataset is the largest and should consist of precisely annotated data. Therefore, we choose nine sequences from MINI-RGBD (1–5, 7–10) and the training subset of SyRIP.

For validation, we selected sequences 11 and 12 from MINI-RGBD, as they contain more complex poses. We also used SyRIP validation data, although its informative value is slightly diminished, as it lacks annotation for SMIL parameters,

The In-Lab data was set for testing. Considering the data is not manually coded, and the annotation is approximate given the way of getting the information (its accuracy depends on the precision of ViTPose), we deemed it not suitable for training at the moment as the learning and the final model would then be influenced by the data's inaccuracies. However, the data could still be used for testing as long as it is reflected on the impact of the data's

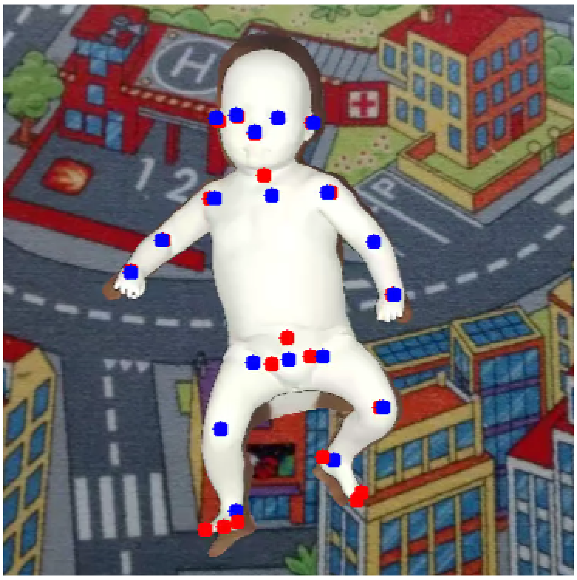


Figure 3.2: Processed image with ViTPose + SMPLify-X with SMIL model. It can be seen that ViTPose (blue dots) placed hips in a line. That is contrary to where the 3D keypoints from SMPLify-X (red dots) are projected—triangle.

reservations.

	training	validation	testing
MINI-RGBD	sequences 1–5, 7–10	sequences 11, 12	sequence 6
SyRIP	1000 synthetic + 200 real	500 real	N/A
In-Lab	N/A	N/A	~3400 images

Table 3.3: Configuration of datasets subsets for training, evaluating and testing HMR 2.0.

3.5 Comparison Metrics

Some of the commonly used metrics for 3D pose estimation in the literature include Mean per Joint Position Error (MPJPE) [39], Percentage of Correcly Estimated Parts (PCP) [40,41], Percentage of Correct Keypoints (PCK) [42], Mean per Bone Position Error (Bone Error), Bones Standard Deviation (Bone Std) [43], Illegal Angle [23], and Mean of the Root Position Error (MRPE) [44]. For our work, we employ various implementations of MPJPE, Bone Std, and the Number of Missed Detections. Bone Std and the Number of Missed Detections do not need corresponding ground-truth data.

3.5.1 Mean per Joint Position Error (MPJPE)

MPJPE averages the absolute error between estimated joints and the ground truth. Joints are here interchangeable with keypoints (it will be referred to keypoints when talking about joints/keypoints further on in this work). MPJPE for one frame can be obtained (in any dimension and space) by averaging the Euclidean distances of the keypoints:

$$E_{\mathrm{MPJPE}}(f) = \frac{1}{N}\sum_{j} \|X_{\mathrm{gt}}^{(f)}(j) - X_{\mathrm{e}}^{(f)}(j)\|, \tag{3.5}$$

where N is the number of keypoints, $X_{\mathrm{gt}}^{(f)}(j)$ denotes the ground truth position of keypoint j in frame f and $X_{\mathrm{e}}^{(f)}(j)$ denotes the estimated position of keypoint j in frame f. The final MPJPE is then computed as the average of the l individual frames errors [39] as

$$E_{\mathrm{MPJPE}} = \frac{1}{l}\sum_{f} E_{\mathrm{MPJPE}}(f). \tag{3.6}$$

MPJPE in 3D

All the methods process the videos frame-by-frame; consequently, the estimated camera settings and shape parameters vary across each video and sometimes even between two subsequent frames. Each method has possibly a different reference frame. That leads to changes in absolute 3D keypoints coordinates; hence, we need to establish a routine that helps us synchronize the coordinate systems of estimated 3D keypoints and the ground truth.

The keypoints in each frame were canonicalized according to a four-step process proposed by Khoury et al. [38]. Firstly, all the keypoints are translated, so the neck keypoint lies in $(0, 0, 0)$. The 3D coordinates are then scaled so that the distance between the Neck and MidHip keypoints is 1. Another step is to align the Neck-MidHip link with the y-axis to place the MidHip coordinate to $(0, -1, 0)$. Lastly, the 3D keypoints are rotated around the y-axis so that the z-axis is orthogonal to the segment between the left and the right hip.

MPJPE in 2D

For comparison in 2D, the estimated 3D keypoints must be projected to the common plane with the ground truth. We evaluated MPJPE in 2D in two ways as each gives us different kinds of information. First, MPJPE in 2D (Type I) was measured in the plane determined by the neck and hip keypoints after canonicalization. This information, together with MPJPE in 3D, can reveal the proportion of error in different axes, *e.g.*, whether the majority of the MPJPE comes from the depth estimation.

Second, MPJPE in 2D (Type II) was measured using the plane of the original image. The units of the error are, therefore, pixels. This metric shows the degree of consensus between what we can visually see in the image and what the model estimates. If we wanted to use it across multiple videos, the videos would need to be of a specific infant's age and session condition. The resolution and camera parameters could also differ, which would change the meaning of what an absolute error of a fixed number of pixels actually describes, *e.g.*, an estimation with a 40-pixel error in a first video can actually be more accurate than a 10-pixel error in a second video with much lower resolution. Therefore, we normalize the computed distances in pixels to the median Neck-MidHip distance, which we obtain from the ground truth for each video separately.

3.5.2 Bone Length Standard Deviation

Bone length standard deviation (Bone Std) measures the stability of estimated bone length across video frames [43]. In reality, the bone lengths of a subject are constant, so any change in bone length is a sign of inaccuracy. In practice, with regards to evaluating pose estimation methods, the keypoints estimations do not locate the extremities of bones, so some variability is expected, and only some pairs of keypoints that are close enough to bone extremities positions can be used to estimate a bone length. Furthermore, this can only done in 3D because 2D locations would be too often subject to distortions from camera angles, *e.g.*, an arm stretched perpendicular to the camera could end up with close to-aligned estimations of the elbow and wrist keypoints that will have close to no error in the 2D space and as such end up with a very short bone length, leading to high variability from factors unrelated to the pose estimation itself.

The standard deviation for separate bone lengths can be calculated as

$$\sigma(b) = \sqrt{\mathbb{E}\left[(l(b) - \mu(b))^2\right]}, \tag{3.7}$$

where $b \in \mathcal{B}$ is the bone, μ is the mean bone length across a video and l is the bone length in each frame. We can then average the standard deviation for different bones to get the mean Bone Std. Bone lengths in this work are calculated from the keypoints after the aforementioned skeleton alignment.

For this metric, the following body segments were considered: ankles-knees (tibia), knees-hips (femur), elbows-shoulders (humerus), and wrist-elbows (radius).

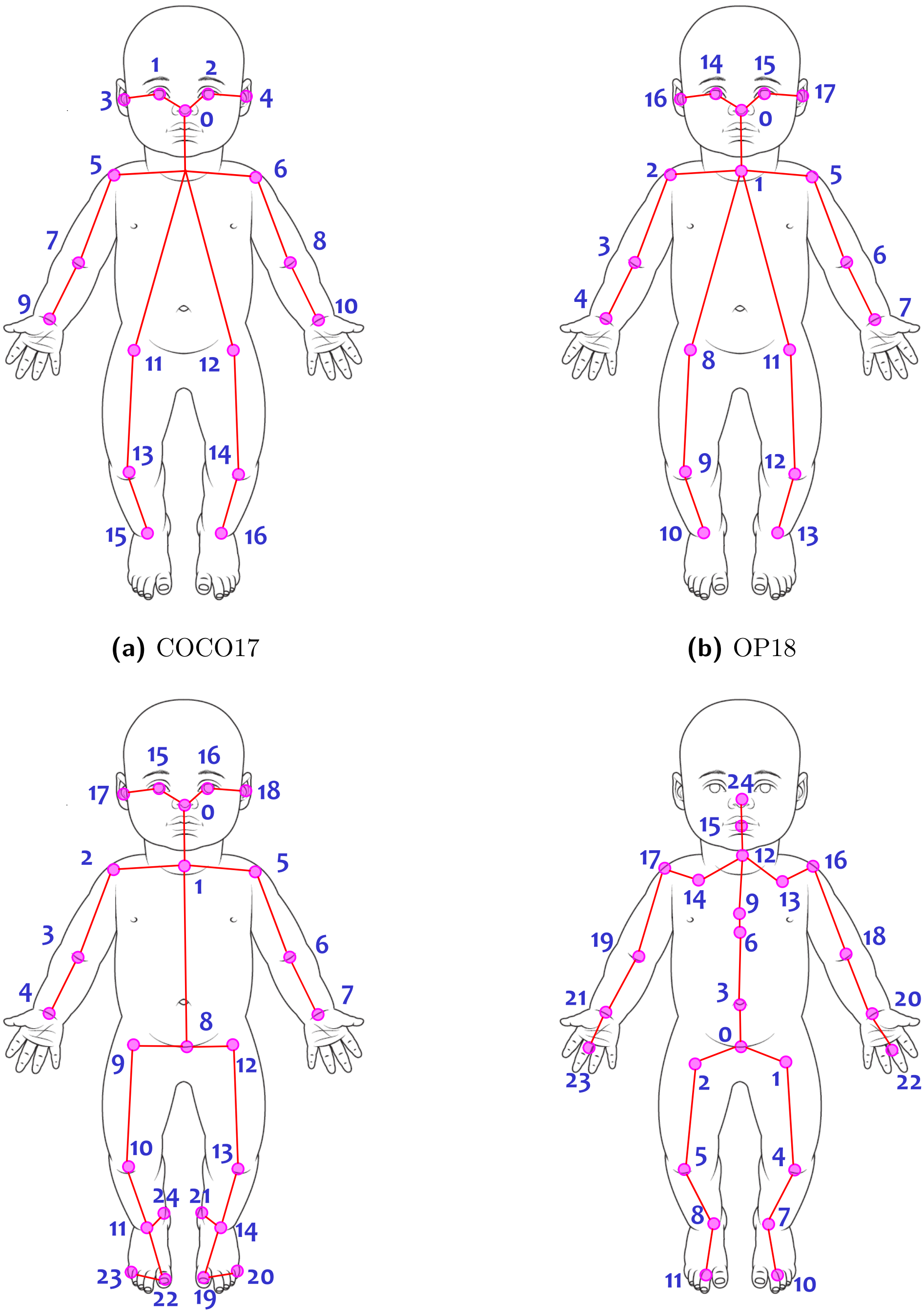
3.5.3 Missed Detections

Each 3D pose estimation usually uses some detector for the initial location of the person of interest (except the bottom-up approaches, such as SMPLify-X + OpenPose). High amounts of missed detections cause breaks in continuity for motion analysis. It also complicates any possible filtering of the data. Although the number of missed detections does not specifically describe the quality of the pose estimation itself, it is a relevant piece of information for the sake of possible improvements to the model, *e.g.*, replacing the detector with another.

It is also possible for the methods to miss keypoints. That sometimes occurs with the two-stage approaches when the 2D pose estimation fails to produce all the keypoints. Therefore, in our case, OpenPose goes through an interpolation step to fill in the missing keypoints. We do not monitor missed keypoints further.

3.6 Computing Hardware Specification

We trained using a workstation with the following specifications: CPU: Intel(R) Xeon(R) W-2295 (18C / 36T, 3.0 / 4.8GHz, 24.75MB), GPU: Two NVIDIA TU104GL Quadro RTX 5000 16 GB, RAM: 251 GB DDR4, OS: Ubuntu 20.04.5 LTS. Only one of the GPUs was used, and its memory limits the batch size to 6. The number of workers for the dataloader is set consequently to 6.



(a) COCO17

(b) OP18

(c) OP25

(d) MINI-RGBD

Figure 3.3: Keypoints sets—overview of different formats. Out of the used formats, only the MINI-RGBD does not include eyes and ears keypoints. A few of the keypoints in Fig. 3.3d need to be more specified: (3) is placed around the belly button, (6) is located at the xiphisternum, (9) is between the nipples, and (15) is at the imagined center of the head. Background infant schematic taken from `https://www.formsbank.com/template/30358/baby-body-chart-medical-assessment-victorian-forensic-paediatric-medical-service.html`

Chapter 4

Experiments and Results

This chapter consists of two sections. The first aims at a comparison of 3D pose estimation methods introduced in Section 2.3 on infant videos. Based on the results of the comparison, the second section then focuses on retraining the most promising method with the SMIL model on infant data. At the start of the project, only the MINI-RGBD dataset was available to us. Therefore, the comparison was made only on the synthetic infants.

4.1 Comparison of 3D Pose Estimation Models

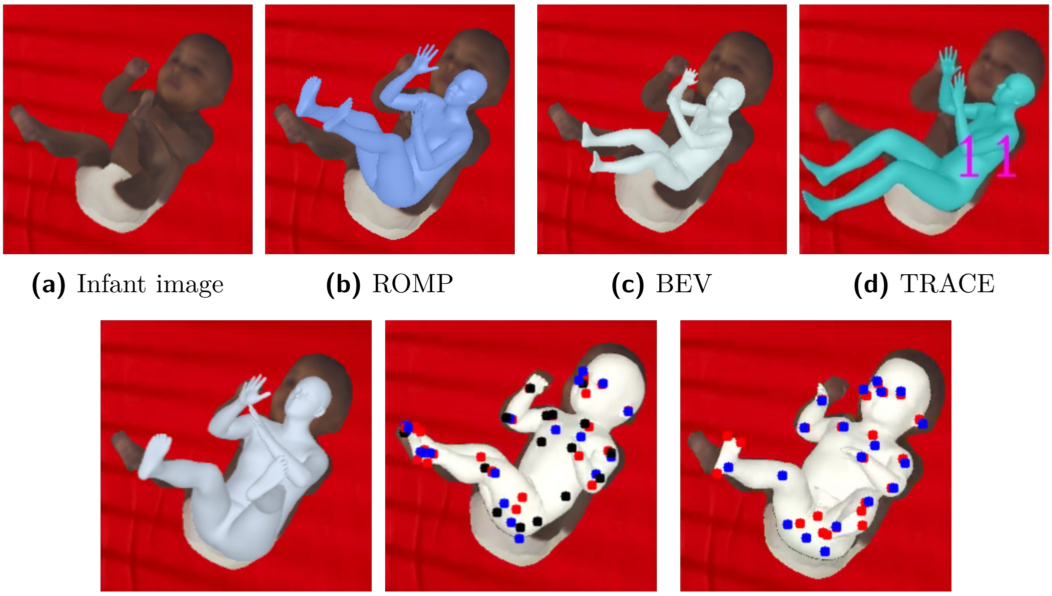
We used the methods for comparison as we set them up in Section 3.2, *i.e.*, in their 'vanilla' versions developed for adults, except for SMPLify-X. ROMP, TRACE, and 4D Humans use SMPL, BEV uses SMPL+A, and SMPLify employs SMIL in our comparison (see Fig. 4.1). Therefore, we visually separate SMPLify with vertical lines from the other methods in the following tables (BEV should be able to estimate infant shapes; however, in our comparison, it generally failed to recognize infants and treated them as adults). Each considered 3D pose estimation model was tested on five synthetic sequences of infants (1, 2, 6, 10, 12) from the MINI-RGBD dataset.

4.1.1 Results

The results comparing the aforementioned models are shown for each metric described in Section 3.5. When comparing models using MPJPE, it must be noted that because the MINI-RGBD dataset lacks eyes and ears notation, we cannot judge the extent of SMPL's influence on the estimation of infants' heads (as visualized in Fig. 2.1).

MPJPE in 3D

The best-performing method in MPJPE in 3D, shown in Table 4.1, is 4D Humans. It accomplished the lowest mean error of 22.2 % of the Neck-MidHip distance. SMPLify-X with ViTPose comes close to 4D Humans, being noticeably worse only on wrists and ankles but better on the nose, elbows, or knees. ROMP shows the best results on wrists. TRACE estimates the elbows the most correctly. We observe the tendency for the leg keypoints to have the highest errors.



(a) Infant image **(b)** ROMP **(c)** BEV **(d)** TRACE

(e) 4D Humans **(f)** SMPLify-X+OP **(g)** SMPLify-X+ViT

Figure 4.1: Qualitative comparison of the models on a synthetic infant image. ROMP (b), TRACE (d), and 4D Humans (e) fit an adult model. BEV (c) has the potential to recognize the adult or infant, but in this case, BEV failed to recognize the infant and fitted an adult body to the infant instead. SMPLify-X both use the SMIL model.

MPJPE in 2D (Type I)

The MPJPE in 2D (Type I) are shown in Table 4.2. 4D Humans absolutely outperformed other methods using this metric. Considering the MPJPE in 2D of 4D Humans is approximately two times smaller than for MPJPE in 3D, we can estimate that the error originates $\sim\sqrt{3}$ times more in the z-axis than in the xy-plane. 4D Humans is followed by SMPLify-X + ViTPose and ROMP. The SMPLify-X + OpenPose behaved the worst. It can be noted from experience that OpenPose more often produces completely displaced keypoints than ViTPose, such as in Fig. 4.2, which results in larger error after SMPLify-X processing. It is surprising to see TRACE, which is one of the newest 3D pose estimation methods with 4D Humans and should improve on BEV and ROMP, perform at a similar level as SMPLify-X + Openpose.

MPJPE in 2D (Type II)

The MPJPEs in 2D (Type II) are shown in Table 4.3. It is also visualized in Fig. 4.3. 4D Humans scored the lowest normalized error of 8.7 % of the Neck-MidHip distance. SMPLify-X + ViTPose came a close second, although it performed noticeably better only on the shoulders and the nose.

Bone Length Standard Deviation

We investigate values of Bone Length Std in Table 4.4. The methods using tracking—4D Humans and TRACE—perform noticeably better than the rest. Generally, the Bone

	ROMP	BEV	TRACE	4D Humans	SMPLify-X + OP	SMPLify-X + ViT
Nose	20.5	21.8	21.3	20.1	17.9	**17.7**
Neck	0	0	0	0	0	0
RShoulder	**6.0**	6.2	7.3	7.4	8.6	9.7
RElbow	21.0	19.2	**19.1**	23.4	19.2	21.5
RWrist	25.6	23.5	24.5	**20.4**	29.6	28.1
LShoulder	**4.4**	4.9	5.9	6.3	12.0	13.8
LElbow	20.8	18.9	**18.5**	27.0	22.4	25.8
LWrist	**23.9**	25.2	27.2	25.9	33.3	37.1
MidHip	16.8	18.2	17.4	**17.2**	23.0	23.0
RHip	0.9	1.5	1.2	**0.9**	1.4	1.3
RKnee	43.3	36.5	41.8	37.1	23.2	**22.1**
RAnkle	44.2	42.7	43.5	**30.0**	39.2	38.6
LHip	0.9	1.5	1.2	**0.9**	1.4	1.3
LKnee	41.0	35.2	42.4	35.0	24.5	**21.1**
LAnkle	45.3	43.1	43.2	**30.4**	43.5	41.0
LBigToe	64.0	59.1	62.0	48.1	52.8	**46.0**
RBigToe	62.8	58.1	61.0	46.7	47.0	**44.0**
$\bar{x}$	26.0	24.5	25.7	**22.2**	23.5	23.1

Table 4.1: MPJPE in 3D in the percentage of the Neck-MidHip distance averaged across all samples.

Length Std is lower on upper limbs as their length is shorter. SMPLify-X, regardless of the 2D pose estimation method, scores the worst. It is surprising that the Bone Length Stds are quite low in contrast to much higher individual keypoint errors for all the methods.

Missed Detections

The percentages of missed detection are shown in Table 4.5. 4D Humans has the largest number of missed detections, contrary to its previous good results in other metrics. This high number of misses means that 4D Humans could ignore a lot of difficult poses, which might be the reason for better performance in other metrics (whereas other methods may achieve high errors on those particular images). Both ROMP and BEV have a relatively small percentage of missed detections compared to 4D Humans and TRACE. SMPLify-X did not miss once. Although SMPLify-X with OpenPose had no missing detections after the 2D step, it had missing keypoints. Therefore, the 2D step for the data that went through SMPLify and OpenPose had been through an interpolation step.

To ensure 4D Humans did not have an unfair advantage in the comparison due to the large number of missed detections on sequence 10, we have further compared the methods with sequence 10 left out (see Fig. 4.6). These results are very similar to the ones with sequence. The only notable improvement can be seen for SMPLify-X + OpenPose on all metrics, and TRACE improved its Bone Length Std. Otherwise, the differences are negligible, and the results do not indicate 4D Humans gaining an unfair advantage regarding

	ROMP	BEV	TRACE	4D Humans	SMPLify-X + OP	SMPLify-X + ViT
Nose	17.3	18.0	17.2	**14.4**	16.8	16.8
Neck	0	0	0	0	0	0
RShoulder	4.7	4.3	6.3	5.2	4.4	**4.0**
RElbow	15.3	13.1	13.9	**10.9**	12.1	11.3
RWrist	19.1	15.9	17.8	**9.2**	18.6	15.4
LShoulder	2.8	**2.7**	4.5	3.3	7.6	7.5
LElbow	14.9	13.5	12.8	**10.3**	13.4	13.9
LWrist	14.9	14.8	15.6	**10.6**	18.9	18.2
MidHip	**12.5**	13.5	12.7	12.7	15.1	15.0
RHip	**0.9**	1.5	1.2	**0.9**	1.4	1.3
RKnee	12.6	13.3	14.9	**10.8**	17.2	15.0
RAnkle	28.9	32.6	33.0	**19.3**	30.5	27.1
LHip	0.9	1.5	1.2	**0.9**	1.4	1.3
LKnee	12.8	15.7	18.9	**9.1**	19.7	16.6
LAnkle	34.2	34.3	34.4	**19.7**	34.3	29.1
LBigToe	41.1	40.6	42.9	**28.7**	41.8	34.5
RBigToe	32.8	36.5	38.0	**26.2**	35.8	30.8
$\bar{x}$	15.6	16.0	16.8	**11.3**	17.0	15.6

Table 4.2: MPJPE in 2D (Type I) in the percentage of the Neck-MidHip distance averaged across all samples.

its larger number of missed detections.

Given that the 4D Humans approach performed the best in most considered metrics (even outperformed SMPLify-X with ViTPose and SMIL) and includes tracking as a feature, we considered it the most promising method to incorporate SMIL to its pipeline and retrain the HMR 2.0 model.

4.2 Retraining the HMR 2.0 with Infant Data

We retrained HMR 2.0 from scratch in four configurations:

- only on the MINI-RGBD synthetic infants (**MR**),
- on MINI-RGBD with SyRIP (**MR** + **S**),
- on MINI-RGBD with adversarial loss ($\mathbf{MR} + \boldsymbol{\mathcal{L}}_{\mathbf{adv}}$),
- on MINI-RGBD with SyRIP with adversarial loss ($\mathbf{MR} + \mathbf{S} + \boldsymbol{\mathcal{L}}_{\mathbf{adv}}$).

When training with SyRIP, the weights of datasets are 0.15 for SyRIP and 0.85 for MINI-RGBD. The weights were chosen so because of the unequal dataset sizes, with the MINI-RGBD being 7.5 times larger than the SyRIP. We then slightly emphasized the

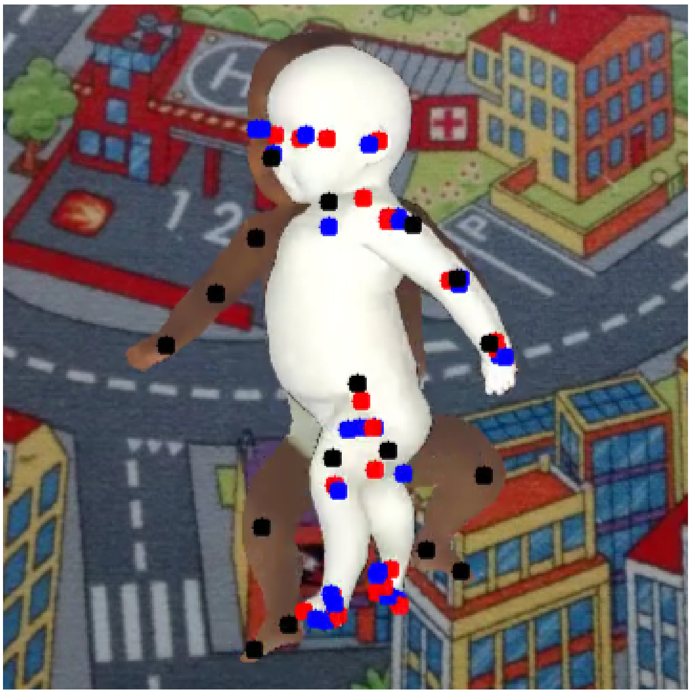


Figure 4.2: OpenPose sometimes places the 2D keypoints in entirely wrong positions, causing SMPLify-X to fail. Black points are the ground truth, blue ones denote the OpenPose placement, and red points are the 3D keypoints estimated by SMPLify-X reprojected to the plane of the image.

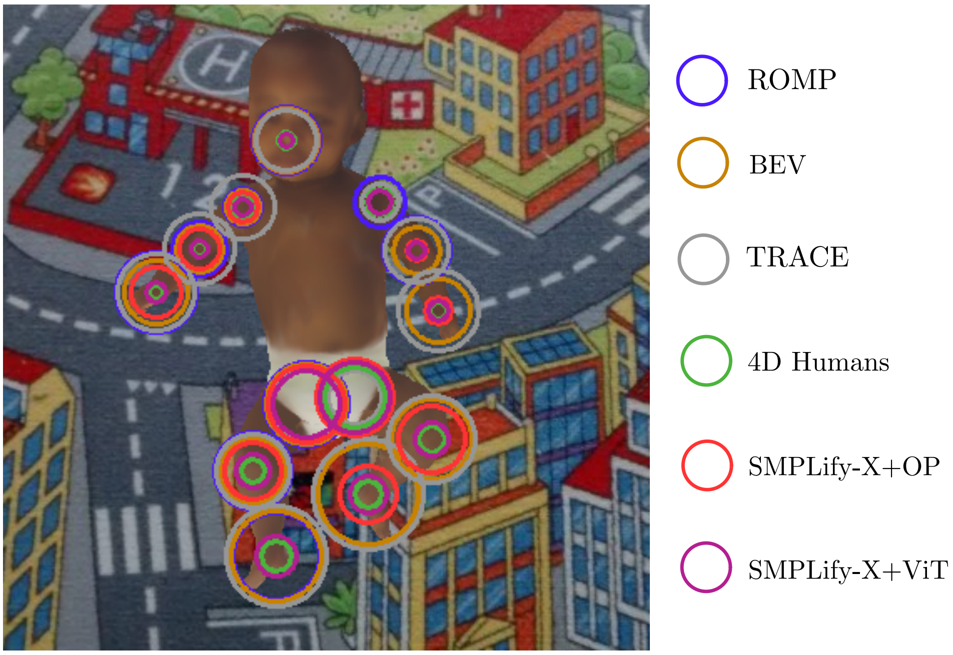


Figure 4.3: Visualization of MPJPE in 2D (Type II). The error measured in the MidHip-Neck distance is scaled to the image so that the number of pixels between the Neck and MidHip corresponds to 1 distance unit.

	ROMP	BEV	TRACE	4D Humans	SMPLify-X + OP	SMPLify-X + ViT
Nose	17.1	16.6	16.6	4.7	**3.5**	**3.5**
Neck	5.8	**5.0**	7.5	6.0	5.9	8.0
RShoulder	10.3	8.9	16.7	5.7	8.0	**4.8**
RElbow	14.5	12.1	18.7	**4.2**	11.0	4.6
RWrist	20.7	17.0	20.5	**4.3**	14.2	5.5
LShoulder	13.0	10.2	9.6	8.0	**6.3**	6.6
LElbow	17.3	12.5	16.8	**4.6**	5.8	4.8
LWrist	20.8	16.5	21.2	**5.1**	7.0	5.6
MidHip	8.5	7.4	10.0	**6.6**	11.8	11.0
RHip	21.5	20.6	18.4	17.9	19.9	**17.8**
RKnee	19.6	16.9	19.7	**6.6**	15.5	9.1
RAnkle	22.1	23.3	26.7	**8.4**	11.0	10.5
LHip	20.0	18.6	**14.5**	16.6	20.0	18.6
LKnee	22.1	20.9	22.8	**7.6**	17.2	9.5
LAnkle	26.3	25.1	29.1	**7.5**	14.5	9.6
LBigToe	34.7	36.8	45.5	**18.3**	20.6	21.6
RBigToe	25.4	34.4	39.0	16.4	**14.1**	25.0
$\bar{x}$	18.8	17.8	20.8	**8.7**	12.1	10.4

Table 4.3: MPJPE in 2D (Type II) in the percentage of the Neck-MidHip distance averaged across all samples.

	ROMP	BEV	TRACE	4D Humans	SMPLify-X + OP	SMPLify-X + ViT
R–Rad	1.8	1.9	**1.1**	**1.1**	2.2	2.0
L–Rad	1.7	1.7	**1.1**	**1.1**	2.0	1.8
R–Hum	1.7	1.9	**1.1**	1.2	2.3	2.1
L–Hum	1.8	2.1	**1.1**	1.2	2.4	2.2
R–Fem	2.6	3.5	**1.7**	**1.7**	3.5	3.3
L–Fem	2.6	3.3	**1.7**	**1.7**	3.4	3.3
R–Tib	2.7	3.1	**1.7**	1.8	4.3	3.7
L–Tib	2.7	3.1	**1.7**	1.8	4.4	3.8
$\bar{x}$	2.2	2.6	**1.4**	1.5	3.1	2.8

Table 4.4: Bone length standard deviations measured in the percentage of the Neck-MidHip distance, averaged across all samples. R/L denotes the right or the left side of the body, the Rad is the radius, Hum is the humerus, Fem is the femur, and Tib is the tibia.

SyRIP weight to raise its influence even more. We trained for 400,000 iterations and created model checkpoints after every 10,000 training iterations. We then chose the checkpoint with the overall lowest validation loss. If the training loss for that specific checkpoint had been considerably higher compared to the course of the training, we moved on to the checkpoint with the second lowest validation loss. The checkpoints that were used for the performance comparison were created after a) 310,000, b) 360,000, c) 370,000, and d) 360,000 iterations for each configuration, respectively.

	ROMP	BEV	TRACE	4DHumans	SMPLify-X + OP	SMPLify-X + ViT
Sq 1	**0**	**0**	**0**	**0**	**0**	**0**
Sq 2	**0**	**0**	**0**	10	**0**	**0**
Sq 6	**0**	**0**	**0**	**0**	**0**	**0**
Sq 10	3	1	36	662	**0**	**0**
Sq 12	**0**	4	262	1	**0**	**0**
$\sum$	3	5	298	673	**0**	**0**
%	0.06	0.10	5.96	13.46	**0.00**	**0.00**

Table 4.5: Missed detections. Each sequence consists of 1000 consecutive images.

	ROMP	BEV	TRACE	4D Humans	SMPLify-X + OP	SMPLify-X + ViT
MPJPE 3D	26.3	25.0	26.3	22.2	**22.1**	23.1
MPJPE 2D (I)	15.9	16.6	16.8	**11.3**	15.9	14.6
MPJPE 2D (II)	19.5	17.9	20.3	**8.7**	12.0	10.1
Bone Std	2.1	2.2	**0.9**	1.5	2.3	2.6

Table 4.6: Overview of results with sequence 10 left out of comparison. All the metrics are measured in percentage of the Neck-MidHip distance.

4.2.1 Results

MPJPE in 3D

We tested the behavior of models on synthetic and real datasets separately. On the synthetic dataset (sequence 6) using MPJPE in 3D (Table 4.7), our trained models performed better than the original model, in particular improving on the keypoints estimation of the legs. However, on our In-Lab dataset with a real infant, the original model is better by 3 % of the Neck-MidHip error on average than our best model, and our trained models' errors are notably higher, as shown in Table 4.8. The trained models with adversarial loss produce slightly lower errors. Models trained on both datasets also have better results in MPJPE in 3D. The visualized 3D meshes for all the models are shown in Fig. 4.5 on five samples.

MPJPE in 2D (Type I)

MPJPE in 2D (Type I) is presented for synthetic data in Fig. 4.9 and for real data in Fig. 4.10. It shows similar results as the MPJPE in 3D, although the original model does not fall behind the trained models on the synthetic data. We can also see that the trained models do not have substantially lower errors in 2D on the real data. We can estimate that the error in the z-axis is ~2/3 of the error in the xy-plane, which is contrary to what we have observed in Section 4.1.1. One drawback of this observation is that the initial comparison also included MINI-RGBD sequences 10 and 12, which include a lot of complex leg postures in the air that sequence 6 lacks.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Nose	26.6	23.9	20.5	**19.8**	21.5
Neck	0	0	0	0	0
RShoulder	7.7	7.6	**6.1**	8.4	6.7
RElbow	16.5	17.5	**15.4**	17.6	20.5
RWrist	29.3	29.6	26.3	28.1	**18.0**
LShoulder	7.0	4.9	7.2	**4.6**	6.3
LElbow	15.8	12.1	16.2	**10.9**	24.4
LWrist	24.2	**20.0**	26.1	20.1	20.2
MidHip	24.5	24.4	24.8	24.6	**19.0**
RHip	2.5	2.4	2.4	2.3	**1.0**
RKnee	10.2	10.5	**9.4**	13.3	34.6
RAnkle	26.1	**22.6**	24.8	25.8	26.3
LHip	2.5	2.4	2.4	2.3	**1.0**
LKnee	**12.3**	14.0	13.3	14.0	34.6
LAnkle	19.2	20.8	19.1	**18.7**	28.7
LBigToe	28.3	31.3	**28.1**	29.5	53.1
RBigToe	34.4	34.1	**33.7**	34.2	47.4
$\bar{x}$	16.9	16.4	16.2	**16.1**	21.4

Table 4.7: MPJPE in 3D in the percentage of the Neck-MidHip distance from MINI-RGBD sequence 6.

MPJPE in 2D (Type II)

The values for MPJPE in 2D (Type II) are shown in Table 4.11 for synthetic infants and in Table 4.12 for the real infants. The original model achieves at least twice as low an error as our trained models on both the synthetic and the real data. The models trained with adversarial loss produce lower errors than when trained without it. The models also trained on SyRIP, which contains a few examples of real infants, perform better on the real infant than their counterparts trained strictly on the MINI-RGBD synthetic data. The visualized keypoints reprojected from 3D to the plane of the image for all the models are shown in Fig. 4.6 on five samples. The visualized error for each keypoint is in Fig. 4.4.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Neck	0	0	0	0	0
RShoulder	14.1	12.7	14.5	**11.1**	14.6
RElbow	19.7	16.4	19.8	**16.3**	20.5
RWrist	26.9	25.0	21.2	22.8	**20.6**
LShoulder	**7.3**	13.0	12.0	11.8	13.9
LElbow	21.7	15.1	13.6	**13.2**	17.1
LWrist	45.9	30.4	33.6	27.3	**20.0**
RHip	13.9	13.9	13.8	13.8	**11.9**
RKnee	25.9	24.1	26.9	26.1	**20.2**
RAnkle	29.5	27.4	35.2	31.7	**17.9**
LHip	13.9	13.9	13.8	13.8	**11.9**
LKnee	24.9	26.5	25.6	24.5	**17.1**
LAnkle	44.9	35.1	49.2	38.1	**20.0**
REye	17.6	**16.6**	17.7	16.7	18.7
LEye	19.6	16.8	17.0	17.9	**15.1**
$\bar{x}$	21.7	19.1	20.9	19.0	**16.0**

Table 4.8: MPJPE in 3D in % of the Neck-MidHip distance from In-Lab data.

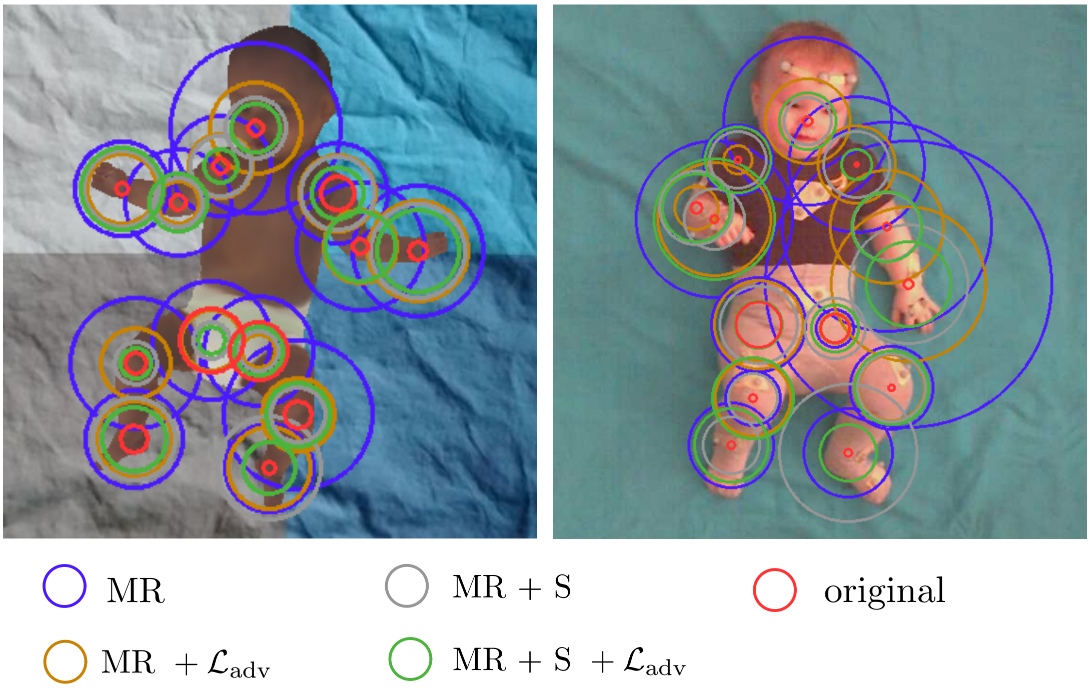

MR
MR + S
original
MR + $\mathcal{L}_{\text{adv}}$
MR + S + $\mathcal{L}_{\text{adv}}$

Figure 4.4: Visualization of MPJPE in 2D (Type II). The error measured in the MidHip-Neck distance is scaled to the image so that the number of pixels between the Neck and MidHip corresponds to 1 distance unit. When looking back to Fig. 4.3, we can see that our best model (MR + S + $\mathcal{L}_{\text{adv}}$) performed in a similar manner as ROMP, BEV, and TRACE. 4D Humans and the methods using SMIL (SMPLify-X) still produce more visually correct bodies.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Nose	25.6	23	19.9	19.1	**16.4**
Neck	0	0	0	0	0
RShoulder	5.4	5.9	**4.1**	6.9	5.7
RElbow	10.4	11.9	8.2	11.4	**9.7**
RWrist	20.7	21.3	16.7	20.5	**9.3**
LShoulder	6.1	**3.7**	5.5	**3.7**	3.9
LElbow	11.5	**6.8**	8.7	7.5	10.0
LWrist	15.5	12.8	11.0	14.4	**8.1**
MidHip	17.2	17.1	17.2	17.1	**13.9**
RHip	2.5	2.4	2.4	2.3	**1.0**
RKnee	**6.6**	8.0	6.8	7.8	10
RAnkle	20.3	**14.8**	17.6	18.3	19.9
LHip	2.5	2.4	2.4	2.3	**1.0**
LKnee	**8.7**	10.9	9.6	10.7	9.3
LAnkle	16.3	17.1	17.0	**12.5**	19.4
LBigToe	21.1	23.7	23.4	**17.0**	29.1
RBigToe	26.0	**18.8**	22.1	25.8	27.4
$\bar{x}$	12.7	11.8	**11.3**	11.6	11.4

Table 4.9: MPJPE in 2D (Type I) in the percentage of the Neck-MidHip distance from MINI-RGBD sequence 6.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Neck	0	0	0	0	0
RShoulder	11	12.1	13.5	**10.6**	13.3
RElbow	**11.4**	12	16.5	11.5	12.7
RWrist	19.5	21	17.3	19.5	**8.9**
LShoulder	**6.1**	10.6	10.7	10.3	13.4
LElbow	19.7	**8.8**	11.1	9.6	13.2
LWrist	42.2	21.7	28.8	21	**9.8**
RHip	13.9	13.9	13.8	13.8	**11.9**
RKnee	21.4	20.1	24.9	22	**6.8**
RAnkle	23.3	20.6	22.1	25.6	**6.9**
LHip	13.9	13.9	13.8	13.8	**11.9**
LKnee	18.9	15.1	19	15.3	**7.9**
LAnkle	40.9	30.4	45.5	30.6	**14.1**
REye	14.9	**13.6**	15.1	14.3	16.5
LEye	16.7	13.6	13.9	14.8	**12.6**
$\bar{x}$	18.2	15.2	17.7	15.5	**10.7**

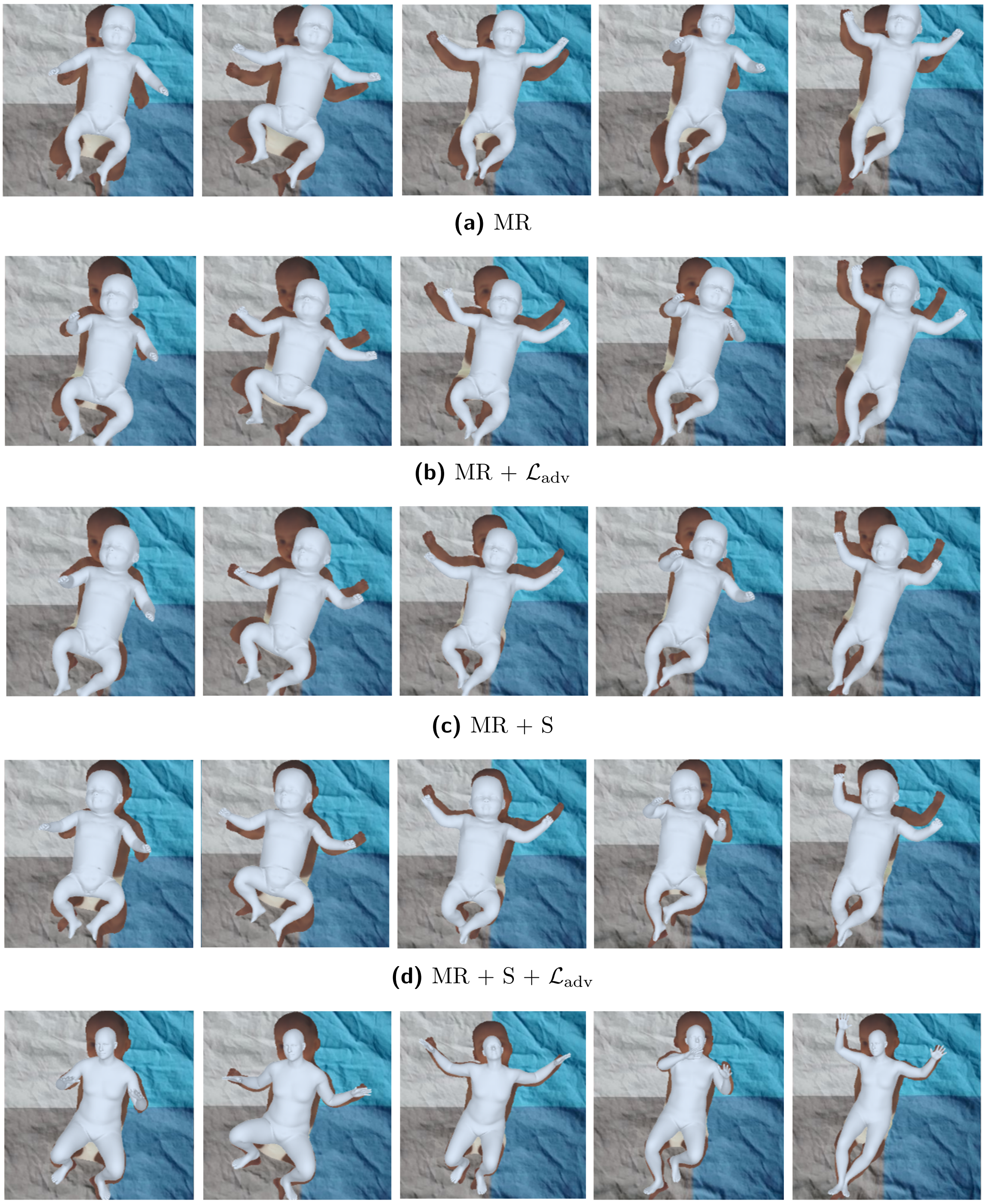
Table 4.10: MPJPE in 2D (Type I) in the percentage of the Neck-MidHip distance from In-Lab data.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Nose	50.2	27.2	18.5	14.7	**5.3**
Neck	21.4	22.2	23.9	17.3	**6.4**
RShoulder	29.4	18.8	18.6	10.5	**4.5**
RElbow	31.1	12.9	14.2	17.4	**4.6**
RWrist	27.6	21.0	22.2	24.6	**4.1**
LShoulder	29.7	22.2	21.5	15.7	**9.8**
LElbow	37.6	21.9	21.2	21.2	**5.3**
LWrist	38.7	29.7	28.2	24.5	**5.9**
MidHip	11.6	29.9	34.8	26.7	**6.9**
RHip	35.4	17.8	10.7	**8.8**	19.0
RKnee	38.6	20.5	12.6	9.7	**6.5**
RAnkle	28.9	22.9	24.8	19.1	**8.6**
LHip	32.5	**10.0**	11.8	14.6	17.1
LKnee	43.7	22.0	16.6	19.3	**8.0**
LAnkle	25.8	24.3	30.2	15.2	**4.2**
LBigToe	21.7	30.6	36.9	**15.3**	18.0
RBigToe	27.6	23.2	27.5	25.3	**15.6**
$\bar{x}$	21.3	22.1	22.0	17.6	**8.8**

Table 4.11: MPJPE in 2D (Type II) in the percentage of the Neck-MidHip distance from MINI-RGBD sequence 6.

	MR	MR + $\mathcal{L}_{\text{adv}}$	MR + S	MR + S + $\mathcal{L}_{\text{adv}}$	original
Nose	51.2	26.1	14.8	17.5	**2.9**
Neck	33.2	19.5	**4.4**	17.4	14.1
RShoulder	21.7	8.9	17.6	20.3	**2.1**
RElbow	22.0	13.4	22.7	20.8	**3.6**
RWrist	47.7	34.9	18.8	37.2	**2.8**
LShoulder	41.6	23.9	20.1	9.5	**1.7**
LElbow	64.0	34.6	21.2	21.3	**2.4**
LWrist	87.3	47.2	36.3	26.0	**3.1**
MidHip	24.8	21.8	**8.0**	19.3	22
RHip	29.2	27.5	24.0	14.8	**14.1**
RKnee	17.0	23.2	15.2	25.1	**2.4**
RAnkle	25.8	17.4	17.7	22.5	**2.4**
LHip	11.6	9.5	18.4	14.9	**9.1**
LKnee	25.4	20.6	21.1	22.6	**1.9**
LAnkle	27.3	17.7	42.3	17.6	**2.5**
REye	44.6	21.5	11.9	19.6	**6.0**
LEye	53.3	28.6	15.1	19.4	**4.0**
REar	25.2	**12.2**	20.3	24.8	15.8
LEar	44.6	28.6	**11.4**	22.6	14.6
$\bar{x}$	36.7	23.0	19.0	20.7	**6.7**

Table 4.12: MPJPE in 2D (Type II) in the percentage of the Neck-MidHip distance from In-Lab data.



(a) MR

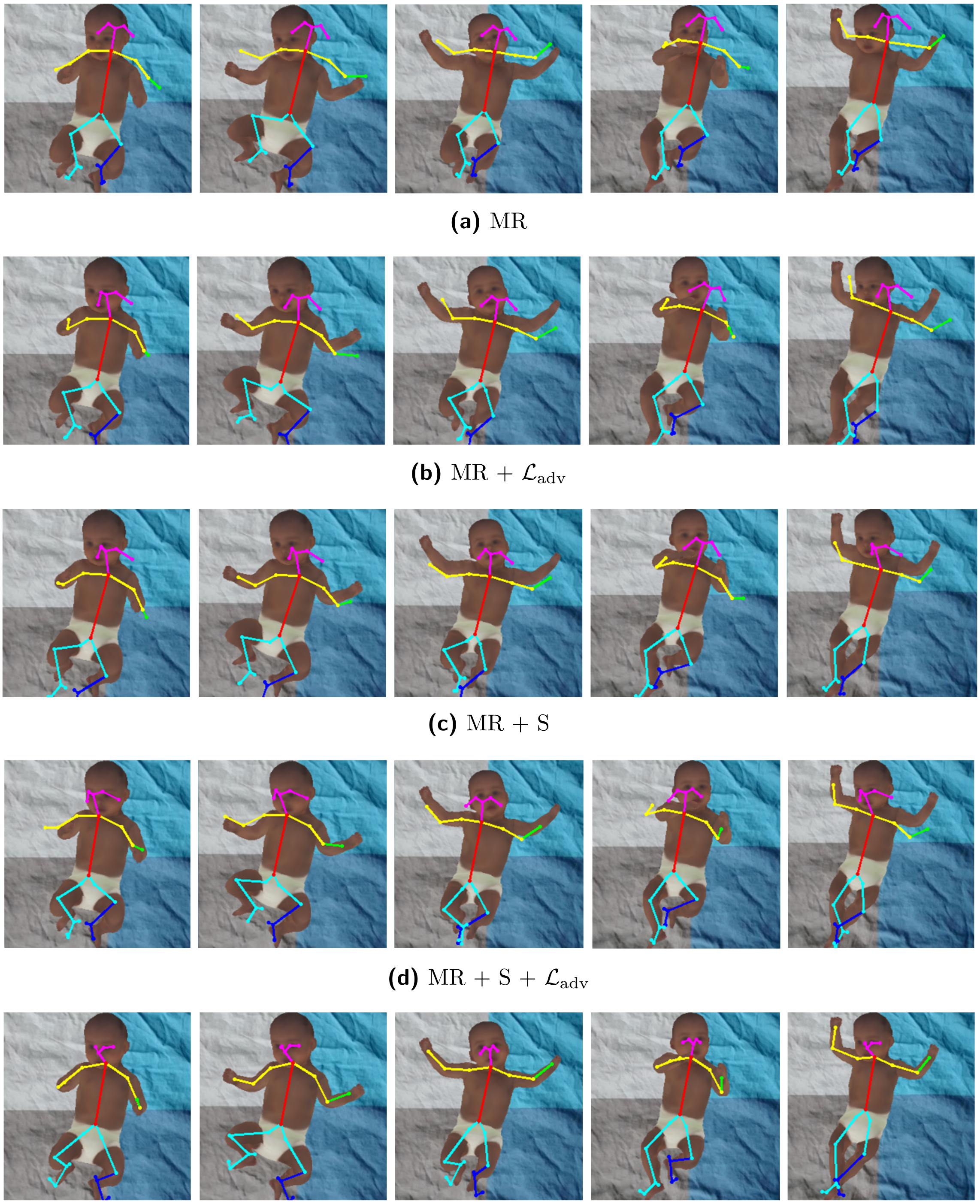
(b) MR + $\mathcal{L}_{\text{adv}}$

(c) MR + S

(d) MR + S + $\mathcal{L}_{\text{adv}}$

(e) Original 4D Humans

Figure 4.5: Comparison of estimated body meshes on differently trained models.



(a) MR

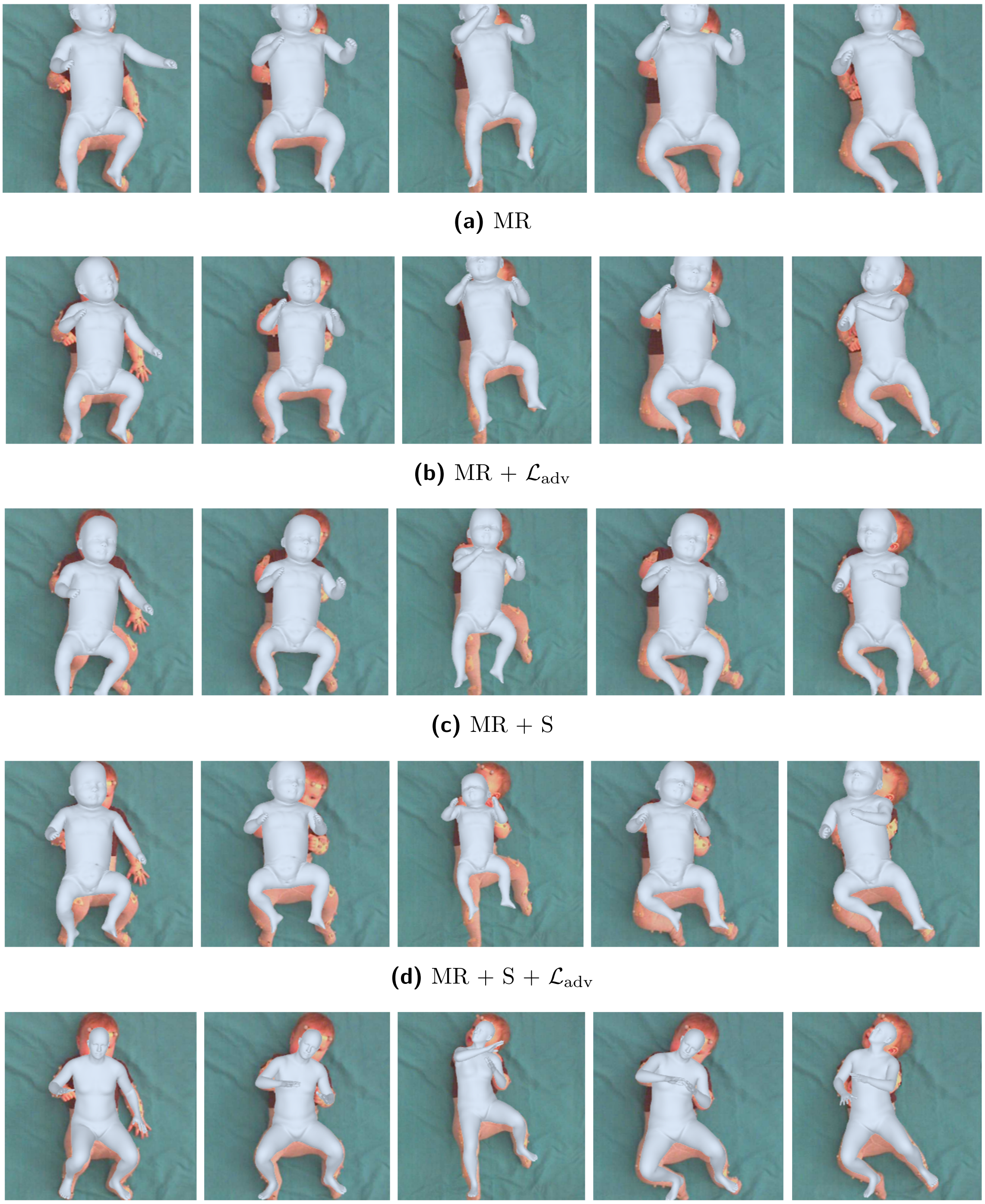
(b) MR + $\mathcal{L}_{\text{adv}}$

(c) MR + S

(d) MR + S + $\mathcal{L}_{\text{adv}}$

(e) Original 4D Humans

Figure 4.6: Comparison of projected 3D keypoints to the image plane on differently trained models.



(a) MR

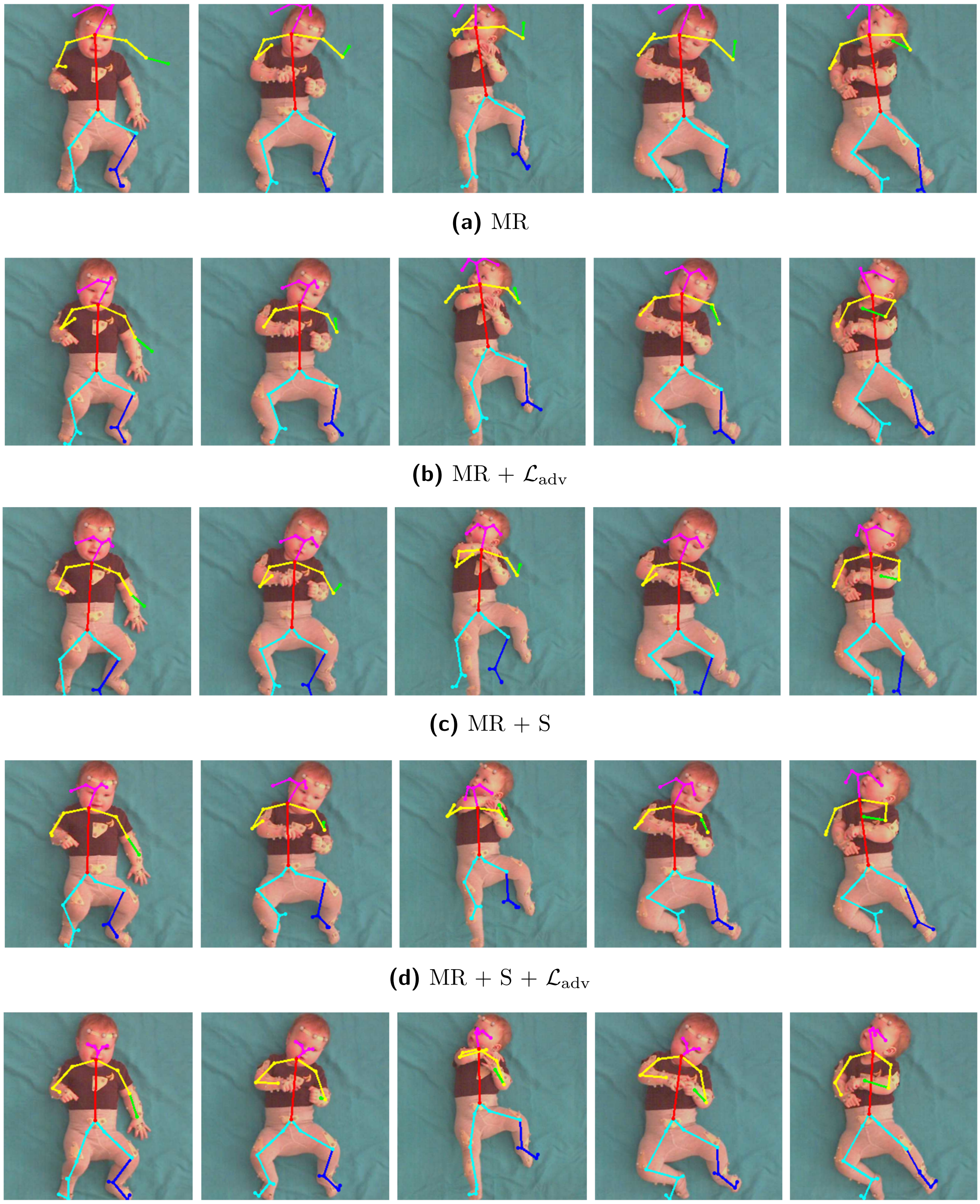
(b) MR + $\mathcal{L}_{\text{adv}}$

(c) MR + S

(d) MR + S + $\mathcal{L}_{\text{adv}}$

(e) Original 4D Humans

Figure 4.7: Comparison of estimated body meshes on differently trained models.



(a) MR

(b) MR + $\mathcal{L}_{\text{adv}}$

(c) MR + S

(d) MR + S + $\mathcal{L}_{\text{adv}}$

(e) Original 4D Humans

Figure 4.8: Comparison of projected 3D keypoints to the image plane on differently trained models.

Chapter 5

Conclusion, Discussion, and Future Work

This work presents a review of several current state-of-the-art methods for 3D pose estimation from RGB videos. We used the methods for comparison in their ‘vanilla’ versions developed for adults, except for SMPLify-X, which was modified in our team’s previous work. ROMP, TRACE, and 4D Humans use SMPL, BEV uses SMPL+A, and SMPLify employs SMIL. The methods were compared on five synthetic sequences of infants from the MINI-RGBD dataset, using the following metrics: Mean Per Joint Position Error (MPJPE) in three forms, bone length standard deviation, and missed detections. The most promising method was then retrained in several configurations to explore the possible areas of its improvement.

In the initial comparison, the best-performing method overall was 4D Humans. However, 4D Humans had the substantially largest number of missed detections. That could have resulted in 4D Humans ignoring some of the difficult poses, on which it would have otherwise achieved higher errors, although further testing did not indicate 4D Humans gaining an advantage over other methods because of that (see Table 4.6). The problem of the high number of misses could be solved by using an alternative detector. We could also observe that methods having tracking (listed in Table 2.1) produced body skeletons with better stability of bone lengths across full videos. Concluding that 4D Humans performed the best in most considered metrics and included tracking, we found it the most promising method to incorporate the SMIL (infant) model into its pipeline and retrain the HMR 2.0 network on infant data.

We used three datasets for the training, validation, and evaluation: the MINI-RGBD dataset with synthetic infants, SyRIP with a mix of real and synthetic infants, and an In-Lab created data with generated pseudo-ground truth. Before retraining the HMR 2.0 network to output SMIL parameters instead of SMPL, we could observe several things: a large portion of the error emerged in the z-axis after the skeleton alignment, *i.e.*, in the depth counting only the estimated skeleton itself. The trained models improved the estimation of depth, likely because the SMIL model better takes account of the infants’ features that are different from those of an adult. We could also see that training with the adversarial loss generally improved the final models. Therefore, we deem the proposed substitution of CMU Mocap data with MINI-RGBD SMIL parameters for training the discriminator justifiable.

Our models are predominantly trained on synthetic data. Our results showed that the trained models improved on the original model after skeleton alignment on the synthetic infants but fell behind on the real infant. The configuration of datasets proposed in

Section 3.4.2 assigned two out of three of the most complex sequences from the MINI-RGBD dataset for validation, leaving those from the training. It might be the reason why our models do not perform well on real infants—they lack the 'hard' synthetic infants (with higher complexity of leg and arm movements) in training, and the In-Lab data that we used for testing consisted of an infant that would be considered 'hard' regarding its movements and poses. Moreover, the models could be overfitted on some properties characterizing images with synthetic infants, such as skin texture and contrast between the infant and background.

We suggest multiple areas of improvement in retraining and fine-tuning the HMR 2.0 network: The first is to convert all the ground truth pose parameters that are in the Euler angle representation to a 6D rotation matrix representation. Euler angle representation suffers from *gimbal lock*, in which the system loses one degree of freedom, and consequently, *e.g.*, there are infinitive rotation parameterizations for the same head pose. The literature indicates that 6D matrix representation positively impacts neural networks' learning of the accurate head pose [45].

Secondly, two options exist to solve the lack of 3D loss during training. One is to properly align the internal reference frame of HMR 2.0 and the ground truth's own reference frame. If unsuccessful, the second is to compute the 3D loss after minimizing Procrustes distance between the ground truth and 4D Humans's estimation.

To create a more robust and overall better estimating model, it needs to be trained on a much larger dataset, preferably consisting of real infants—the bias towards the synthetic data could be solved this way. To obtain more data, pseudo-ground-truth fits (like those from our In-Lab recorded data that we used for testing) can be used, taking inspiration from the training of the original HMR 2.0 [16]. Another possible approach to achieve a more robust model would be to fine-tune the original HMR 2.0 network instead of completely retraining it from scratch.

One possible area of future pursuit is the improvement of the estimation of poses that are close to self-contact and further detection of self-touches. Self-touches are an important behavioral pattern in early infancy, and their successful automatic monitoring carries great research potential. Current 3D pose estimation methods typically fail to detect self-contacts—due to commonly used repulsion mechanisms to avoid collisions, the estimations usually end up farther from real positions when there is self-touch or in configurations close to self-touch. Müller et al. [46] propose a new method, TUCH, to improve pose estimation with self-contact. Another possible direction of research, taking inspiration from BEV and its use of SMPL+A, is to explore further age-based model fitting to achieve better pose estimations on infant-parent interaction videos.

Bibliography

[1] P. Rochat, "Self-perception and action in infancy," *Experimental brain research*, vol. 123, pp. 102–109, 1998.

[2] C. Einspieler and H. F. Prechtl, "Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system," *Mental retardation and developmental disabilities research reviews*, vol. 11, no. 1, pp. 61–67, 2005.

[3] D. M. Romeo, D. Ricci, C. Brogna, and E. Mercuri, "Use of the hammersmith infant neurological examination in infants with cerebral palsy: a critical review of the literature," *Developmental Medicine & Child Neurology*, vol. 58, no. 3, pp. 240–245, 2016.

[4] H. I. Shin, H.-I. Shin, M. S. Bang, D.-K. Kim, S. H. Shin, E.-K. Kim, Y.-J. Kim, E. S. Lee, S. G. Park, H. M. Ji *et al.*, "Deep learning-based quantitative analyses of spontaneous movements and their association with early neurological development in preterm infants," *Scientific Reports*, vol. 12, no. 1, p. 3138, 2022.

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[7] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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*, 2014, pp. 3686–3693.

[8] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.

[9] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu *et al.*, "AI Challenger: A large-scale dataset for going deeper in image understanding," *arXiv preprint arXiv:1711.06475*, 2017.

[10] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision transformer for generic body pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[11] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[12] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides, "Civilian american and european surface anthropometry resource (CAESAR), final report, volume I: Summary," *Sytronics Inc Dayton Oh*, p. 3, 2002.

[13] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 975–10 985.

[14] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I.* Springer, 2018, pp. 792–800.

[15] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, "Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2540–2551, 2019.

[16] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 783–14 794.

[17] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5253–5263.

[18] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3D people," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 179–11 188.

[19] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3D people in depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 243–13 252.

[20] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, "TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8856–8866.

[21] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[22] CMU MoCap: Carnegie Mellon University Motion Capture Database. [Online]. Available: http://mocap.cs.cmu.edu/

[23] I. Akhter and M. J. Black, “Pose-conditioned joint angle limits for 3D human pose reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1446–1455.

[24] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, “Monocular 3D human pose estimation in the wild using improved CNN supervision,” in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 506–516.

[25] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, “Learning 3D human dynamics from video,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5614–5623.

[26] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. 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*, 2018, pp. 6047–6056.

[27] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, “Unite the people: Closing the loop between 3d and 2d human representations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6050–6059.

[28] S. Johnson and M. Everingham, “Learning effective human pose estimation from inaccurate annotation,” in *CVPR 2011*. IEEE, 2011, pp. 1465–1472.

[29] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, “AGORA: Avatars in geography optimized for regression analysis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 468–13 478.

[30] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, “Single-shot multi-person 3D pose estimation from monocular RGB,” in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 120–130.

[31] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, “CrowdPose: Efficient crowded scenes pose estimation and a new benchmark,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 863–10 872.

[32] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. 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)*, 2018, pp. 601–617.

[33] W. Zhang, M. Zhu, and K. G. Derpanis, “From actemes to action: A strongly-supervised representation for detailed action understanding,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.

[34] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. Sebastian Schroeder, “Computer vision for medical infant motion analysis: State of the art and RGB-D data set,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[35] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, “Invariant representation learning for infant pose estimation with small data,” in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[36] N. Vaculínová, “Evaluation framework for infant 3D pose extraction from RGB images using RGB-D cameras and motion capture system,” May 2023.

[37] S. Nousias, M. Lourakis, and C. Bergeles, “Large-scale, metric structure from motion for unordered light fields,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3287–3296.

[38] J. Khoury, S. T. Popescu, F. Gama, V. Marcel, and M. Hoffmann, “Self-touch and other spontaneous behavior patterns in early infancy,” in *2022 IEEE International Conference on Development and Learning (ICDL)*. IEEE, 2022, pp. 148–155.

[39] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, “Deep 3D human pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.

[40] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Progressive search space reduction for human pose estimation,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[41] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, “3D human pose estimation: A review of the literature and analysis of covariates,” *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[42] Y. Yang and D. Ramanan, “Articulated human detection with flexible mixtures of parts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.

[43] X. Sun, J. Shang, S. Liang, and Y. Wei, “Compositional human pose regression,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2602–2611.

[44] G. Moon, J. Y. Chang, and K. M. Lee, “Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 133–10 142.

[45] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, “6D rotation representation for unconstrained head pose estimation,” in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2496–2500.

[46] L. Müller, A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, “On self-contact and human pose,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9990–9999.