Automatic infant 2D pose estimation from videos: comparing seven deep neural network methods (Supplementary Materials)

Filipe Gama[1], Matěj Mísař[1], Lukáš Navara[1], Sergiu T. Popescu[1], and Matej Hoffmann*[1]

[1]Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics, Prague, Czech Republic
*matej.hoffmann@fel.cvut.cz

Materials and Methods

2D Pose estimation techniques

A summary of the characteristics of the different methods is available in Table S1.

Properties	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTpose
TD/BU	TD	BU	TD	TD	BU	TD	BU	TD
Architecture	R-CNN	R-CNN	R-CNN	CNN	HRNet	HRNet	CNN	Transformer
Scores (Y/N)	Y	N	Y	N	Y	N	N	N
Confidences (Y/N)	Y	Y	Y	N	N	Y	Y	Y
Training dataset	COCO	MPII	COCO	COCO	COCO	COCO	COCO+MPII	COCO+MPII+AIC

Table 1. Summary comparison overview of the methods used. TD: Top-Down, BU: Bottom-Up.

To make the comparisons between the methods fair, we used the versions of the models trained on the COCO dataset, but some of the methods have models trained on other datasets, which sometimes provide a different amount of keypoints and different training dataset sizes and examples, such as BODY 25 (25 keypoints), MediaPipe/BlazePose (33 keypoints), COCO-WholeBody (133 keypoints), or Halpe (136 keypoints). These other variants may or may not perform better than the ones we used and might be more suited to study some tasks (e.g., finger estimation would be needed to study grasping, and more facial keypoints could help to study head orientation).

The details of the parameters and weights version used are described below, for each method.

AlphaPose

AlphaPose, publicly available at (https://github.com/MVIG-SJTU/AlphaPose), was run using the *Fast Pose (DUC) - ResNet50 unshuffled* version with following parameters:

```
--cfg 256x192_res50_lr1e-3_1x-duc.yaml \
--checkpoint fast_421_res50-shuffle_256x192.pth \
--detbatch 2 --posebatch 40
```

DeepLabCut

DeepLabCut, publicly available at https://github.com/DeepLabCut/DeepLabCut, was run on version 2.2.1.1 with the *full_human* pre-trained model from DLC's Model Zoo trained on the MPII dataset, with the following parameters.

```
shuffle=1, trainingsetindex=0
```

Detectron2

Detectron2 is publicly available at https://github.com/facebookresearch/detectron2. Detectron2 is a library with models trained to solve several computer vision tasks, some of them being keypoint detection and pose estimation. Those have been trained on the COCO dataset. We used the model *R50-FPN* with model ID: *137849621* (see Detectron2/ModelZoo on github) on version 0.6. It has been run using the following parameters:

```
--config-file keypoint_rcnn_R_50_FPN_3x.yaml \
--opts MODEL.WEIGHTS model_final_a6e10b.pkl
```

MediaPipe

MediaPipe is publicly available at https://github.com/google-ai-edge/mediapipe. It is a library with models trained to solve several tasks, including human pose estimation with *BlazePose*. It has been trained with BlazeFace and BlazePalm on top of COCO, for extra face and hand keypoints, which we did not use. We used its Python library on version 0.10.14 with the *heavy* model, with default parameters and tracking enabled for both input types:

```
model_asset_path="pose_landmarker_heavy.task"
num_poses=1
running_mode=vision_running_mode.VIDEO
```

HRNet

We used the HRNet implementation of the MMPose environment on version 0.28.0, which is publicly available at https://github.com/open-mmlab/mmpose. This environment contains many pre-trained models, trained on different datasets including COCO, MPII, COCO-WholeBody and Halpe. Some, including HRNet, have been implemented using the two general approaches to pose estimation methods, *Bottom-Up* and *Top-Down*. The Bottom-Up version has been run using the following parameters:

```
HRNet_w32_coco_512x512-bcb8c247_20200816.pth
```

The Top-Down version has been run using the following parameters, with the default detector:

```
faster_rcnn_r50_caffe_fpn_mstrain_1x_coco-5324cff8.pth \
HRNet_w48_coco_384x288_dark-e881a4b6_20210203.pth
```

OpenPose

OpenPose, publicly available at https://github.com/CMU-Perceptual-Computing-Lab/openpose, has been run with the following parameters:

```
--display 0 --render_pose 0 --model_pose COCO \
--number_people_max 1 --net resolution "512x400" \
--scale_number 2 --scale_gap 0.25
```

ViTPose

We used the implementation of ViTPose available through the MMPose environment on version 1.1.0, which is publicly available at https://github.com/open-mmlab/mmpose . There are several pre-trained models, trained on different datasets, we used VitPose-H (Huge) trained on COCO, AIC and MPII, with the default detector. It has been run with the following parameters:

```
faster_rcnn_r50_fpn_1x_coco_20200130-047c8118.pth \
td-hm_ViTPose-huge_8xb64-210e_coco-256x192-e32adcd4_20230314.pth
```

Datasets

In addition to the real infants annotated data mentioned in the main manuscript, additional images were annotated.

We manually annotated 720 additional images from eight additional videos of the same two infants. The images were selected in a similar manner as described in the main manuscript, but were processed only with video input by the methods, hence why they are not included in the main manuscript where the results between videos and images input are compared. Thus, we reach a total of 1440 annotated images from 16 videos (191 862 images) that were processed as video inputs by all methods.

We also annotated 900 additional consecutive images from a single video, the one identified as "AA_17w" in several tables in the Supplementary Materials. This sequence was selected due to the presence of typical hand movements and self-touch occurrences that we are interested in studying in the future. In retrospective, we observed that this video is the one where all the methods performed the best (except for DeepLabCut and MediaPipe, see Tab. ST. 4), most likely because the camera is properly positioned and angled above the infant (see Figure 1., b) in the main manuscript), and the infant does not perform many complex leg movements. As such, it could be considered to have optimal conditions and show the upper bound performance of the methods on infants in supine position. Due to how this sequence was arbitrarily chosen and because it only concerns a single infant at a single age, we decided to not include it in the main manuscript, and leave it as an extra in the Supplementary Materials.

Results

Settings for detection selection with redundant detections

In most applications, when there are redundant detections, a selection process is necessary to choose one of them for further processing. This can only be done by using information that is provided by the pose estimation methods alongside the keypoints, such as the rank order in which the detections are output, or their scores and confidences. We computed Euclidean distances between estimations and their ground truth, the first processing step at the base of the OKS and Neck-MidHip computations, under each of these three detection selection settings:

- *Det 1*: using the first ranked detection provided by the method. Redundant detections proposed by the method are ignored.
- *Det 2*: the average Euclidean distance is computed for each detection of a given image. Then, the detection with the shortest distance is selected. This corresponds to the best detection that the method can offer, though it is not available without ground truth.
- *Det 3*: using the detection with the highest score, ignoring the rank order. For HRNet TD that did not provide scores, we used the score of the bounding-box provided by its detector as a substitute. For ViTPose, we used the median of its confidences.

These settings are used to 1) verify the general assumption that the detection with the highest score is always ranked first, and if not, 2) which of the two available selection metrics (rank or score) matches the best detection proposed by the methods and is a better choice.

Minimal differences were observed when comparing the results for all relevant metrics between the three detection selection settings, namely, the first-rank detection, the optimal detection, and the highest-scored detection. DeepLabCut, MediaPipe and OpenPose yielded identical results, as they only provided one detection. AlphaPose and ViTPose showed a few cases where the first-ranked detection did not have the highest score, though for ViTPose it might be due to how we estimated its score as the median of the confidences: a different estimation might have led to different or no changes. Both first-ranked and highest-score detections had differences with the optimal detection, meaning that there is no guaranteed way to select the best available detection for any of the methods when no ground-truth is available. Generally, using the highest-scored detection resulted in the closest performance to the optimal detection for all methods. Hence, the whole manuscript focuses on showing the results from the highest-scored detections only.

Results summary for the dataset of 1440 real infants annotated images processed by video input

The summary of the results including all 1440 annotated images from the 16 recordings is described in the Tab. ST 2.

We observe similar results as in the main manuscript, with a tendency for slightly higher AP, AR and OKS, and lower missing data, but also lower correlations between score and OKS values.

Metrics	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
OKS	0.88±0.07	0.11±0.12	0.87±0.09	0.32±0.20	0.90±0.06	0.92±0.05	0.81±0.17	**0.92±0.04**
AP	66.5	0.8	25.3	0.6	77.8	59.1	59.9	**86.7**
AR	76.9	0.3	79.1	4.0	85.9	88.5	66.8	**90.0**
Missing data	3.3%	**0%**	**0%**	5.6%	**0%**	0.9%	6.1%	0.1%
Redun. det.	4.6%	**0%**	259.5%	**0%**	5.1%	54.0%	**0%**	19.7%
Sco.-OKS corr.	0.26	0.08	0.21	N/A	**0.51**	0.47	0.35	0.39

Table 2. Summary of the highest-scored detection results for the full set of 1440 real infants annotated images processed by video input. For score-OKS Spearman Rank Coefficient Correlations, all p-values < 0.005.

Results summary for the 900 continuous annotated images from AA_17w

The results summary for the 900 continuous manually-annotated images from a single recording session, AA_17w, are summarized in Supplementary Tab. ST. 3.

The AP, AR, and OKS values are higher than the averages found in the main manuscript or in Supplementary Tab. ST 2, while the average Neck-MidHip errors are lower, except for DeepLabCut and MediaPipe.

Concerning the correlations, we observe that they are particularly low, especially for Detectron2, where they are negative.

Input	Metrics	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Images	OKS	0.90±0.04	N/A	0.93±0.04	0.40±0.10	0.94±0.03	**0.95±0.03**	0.93±0.05	N/A
	AP	77.9	N/A	81.4	0.1	**91.2**	66.4	87.8	N/A
	AR	81.7	N/A	90.0	2.4	94.3	**95.2**	90.6	N/A
	Neck-MidHip error	5.0±1.7%	N/A	4.7±2.4%	24.7±12.0%	**4.0±2.0%**	**3.9±2.2%**	4.5±1.9%	N/A
	Sco.-OKS corr.	0.11	N/A	-0.35	N/A	0.04 ($p = 0.26$)	**0.30**	0.06 ($p = 0.06$)	N/A
Videos	OKS	0.90±0.04	0.08±0.06	0.92±0.04	0.40±0.10	0.94±0.03	0.95±0.03	0.93±0.05	**0.95±0.02**
	AP	76.9	0.0	34.2	0.3	90.8	63.9	87.3	**91.8**
	AR	81.1	0.0	89.9	2.5	93.8	**95.1**	90.2	94.5
	Neck-MidHip error	5.1±1.8%	79.3±15.3%	4.9±2.5%	24.8±0.12%	4.0±2.0%	3.9±2.2%	4.6±1.9%	**3.9±2.0%**
	Sco.-OKS corr.	0.10 ($p < 0.006$)	0.11	-0.32	N/A	0.01 ($p = 0.70$)	**0.28**	0.10	0.18

Table 3. Summary of the highest-scored detections results from each metric for the set of 900 consecutive annotated images from video AA_17w. For score-OKS Spearman Rank Coefficient Correlations, all p-values < 0.005 when not stated.

Object Keypoint Similarity (OKS)

The average OKS of individual videos can be seen in Tables ST 4 and ST 5. It can be used to identify the most challenging or easiest videos for each method.

As the synthetic infants from the MINI-RGBD dataset come with an estimation of the complexity of their sequence, we can observe that the performance between the easy group (IDs 1-4) and the medium group (5-9) seems close, except for synthetic infant 1 that some methods seem to struggle with, and synthetic infant 9 for which all methods show a drop of performance. Synthetic infants 7 and 8 are handled differently by the methods, some keeping their performance levels, while others display drops of performance. However, for all methods, we can observe a performance drop between the easy and medium video group and the difficult video group (IDs 10-12).

Real	Video ID	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Images	AA_8w	0.86	N/A	0.78	0.50	0.81	**0.89**	0.79	N/A
	AA_17w	0.90	N/A	0.92	0.39	**0.94**	**0.94**	0.92	N/A
	TH_8w_st1	0.89	N/A	0.88	0.43	0.90	**0.91**	0.88	N/A
	TH_8w_st2	0.90	N/A	0.89	0.39	0.91	**0.93**	0.89	N/A
	TH_8w_st3	0.88	N/A	0.89	0.32	0.90	**0.91**	0.89	N/A
	TH_15w	0.77	N/A	0.86	0.17	0.90	**0.92**	0.80	N/A
	TH_19w	0.87	N/A	0.86	0.19	0.90	**0.93**	0.88	N/A
	TH_25w	0.89	N/A	0.87	0.50	**0.93**	**0.93**	0.88	N/A
	Mean	0.87	N/A	0.87	0.39	0.90	**0.92**	0.87	N/A
Videos	AA_8w	0.86	0.31	0.77	0.52	0.81	0.89	0.79	**0.94**
	AA_11w	0.86	0.19	0.85	0.19	0.90	**0.91**	0.85	**0.91**
	AA_13w	0.90	0.48	0.91	0.19	0.91	**0.94**	0.81	**0.94**
	AA_17w	0.90	0.09	0.92	0.37	0.94	0.94	0.92	**0.95**
	AA_19w	0.89	0.02	0.90	0.15	**0.93**	**0.93**	0.91	**0.93**
	TH_8w_st1	0.88	0.15	0.87	0.42	0.90	**0.91**	0.87	**0.91**
	TH_8w_st2	0.90	0.07	0.88	0.39	0.91	**0.93**	0.90	**0.93**
	TH_8w_st3	0.88	0.10	0.89	0.36	0.90	0.90	0.89	**0.92**
	TH_10w_st1	0.90	0.15	0.89	0.47	0.91	**0.93**	0.89	**0.93**
	TH_10w_st2	0.84	0.08	0.87	0.33	**0.90**	0.89	0.80	**0.90**
	TH_12w	0.89	0.20	0.88	0.40	0.92	0.90	0.84	**0.93**
	TH_15w	0.75	0.04	0.86	0.19	0.90	**0.92**	0.58	**0.92**
	TH_17w	0.87	0.04	0.86	0.07	**0.90**	0.88	0.83	0.89
	TH_19w	0.88	0.05	0.86	0.11	0.90	0.92	0.52	**0.93**
	TH_23w	0.87	0.06	0.82	0.24	0.91	0.91	0.68	**0.93**
	TH_25w	0.88	0.17	0.87	0.49	0.93	0.93	0.76	**0.94**
	Mean	0.88	0.11	0.87	0.32	0.90	**0.92**	0.81	**0.92**

Table 4. Average OKS values over all manually-annotated images for each method and input type on real infants.

Synth.	Video ID	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Images	Syn. 1	0.83	N/A	0.84	0.34	0.61	**0.90**	0.84	N/A
	Syn. 2	0.87	N/A	0.83	N/A	0.90	**0.91**	0.86	N/A
	Syn. 3	0.89	N/A	0.90	0.52	0.90	**0.91**	0.89	N/A
	Syn. 4	0.90	N/A	0.90	0.62	**0.90**	**0.90**	0.84	N/A
	Syn. 5	0.89	N/A	0.87	0.59	**0.90**	**0.90**	0.85	N/A
	Syn. 6	0.86	N/A	0.88	0.57	**0.91**	0.90	0.88	N/A
	Syn. 7	0.89	N/A	0.89	0.59	**0.91**	**0.91**	0.88	N/A
	Syn. 8	0.85	N/A	0.85	0.45	0.87	**0.89**	0.84	N/A
	Syn. 9	0.73	N/A	0.82	0.52	0.86	**0.90**	0.72	N/A
	Syn. 10	0.79	N/A	0.79	0.39	0.76	**0.81**	0.75	N/A
	Syn. 11	0.84	N/A	0.87	0.55	0.87	**0.88**	0.86	N/A
	Syn. 12	0.67	N/A	0.66	0.35	0.75	**0.78**	0.75	N/A
	Mean	0.84	N/A	0.84	0.50	0.83	**0.88**	0.83	N/A
Videos	Syn. 1	0.76	0.37	0.85	0.22	0.59	0.87	0.85	**0.89**
	Syn. 2	0.86	0.46	0.82	N/A	0.89	**0.90**	0.84	0.89
	Syn. 3	0.88	0.49	0.88	0.51	0.88	**0.89**	0.87	**0.89**
	Syn. 4	0.86	0.40	0.86	0.61	0.82	0.84	0.80	**0.88**
	Syn. 5	0.88	0.59	0.85	0.57	0.87	**0.90**	0.85	**0.90**
	Syn. 6	0.86	0.51	0.87	0.57	**0.90**	**0.90**	0.86	0.89
	Syn. 7	0.87	0.55	0.87	0.58	**0.89**	**0.89**	0.86	0.88
	Syn. 8	0.83	0.36	0.83	0.44	0.85	**0.87**	0.82	0.85
	Syn. 9	0.48	0.38	0.74	0.42	0.78	0.87	0.70	**0.89**
	Syn. 10	0.78	0.35	0.75	0.38	0.74	0.78	0.72	**0.83**
	Syn. 11	0.83	0.50	0.85	0.54	**0.86**	**0.86**	0.84	0.84
	Syn. 12	0.56	0.19	0.59	0.35	0.69	0.75	0.69	**0.79**
	Mean	0.81	0.43	0.81	0.47	0.81	0.86	0.81	**0.87**

Table 5. Average OKS values over all manually-annotated images for each method and input type on synthetic infants.

Neck-MidHip error

Figures showing the average Neck-MidHip errors, including DeepLabCut, for each individual keypoint with available ground truth are shown in Fig. SF 1.

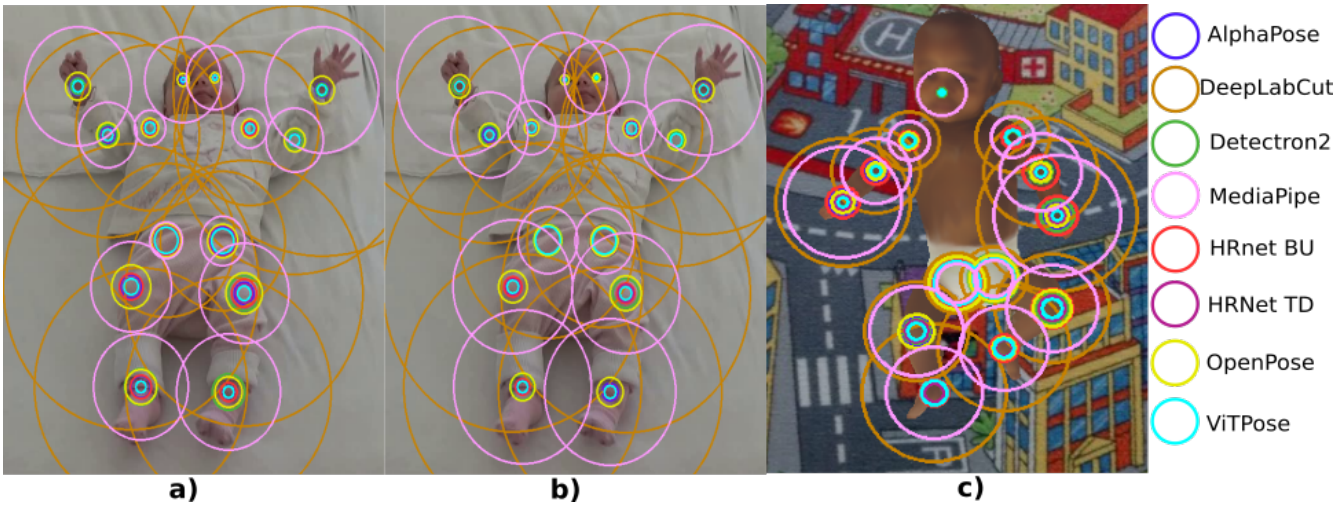

Figure 1. Average Neck-MidHip errors for each keypoint with available ground truth. The centre of the circles is the ground-truth position for that keypoint. The radius of each circle shows the average amplitude of the errors, scaled to the Neck-MidHip segment. The colors separately represent each pose estimation method. (a) Real infants, video input (720 annotations); (b) Real infants, video input (1440 annotations); (c) Synthetic infants, video input

The details of the overall average Neck-MidHip errors across all keypoints with their standard deviations for each method are shown in Tab. ST. 6. The best method is ViTPose, achieving average errors as low as 6.0±2.7% of the Neck-MidHip segment. HRNet Top-Down is slightly behind. However, the variability between each keypoint is generally high: the standard deviations across all keypoints are often above one third, and sometimes even half, of the average error. This can also be see more visually in the main manuscripts' Fig. 2 and in the Supplementary Materials Fig. SF. 1.

Dataset	Input	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Real	Images [720]	8.3 ± 3.1	N/A	10.2 ± 5.2	38.0±13.8	7.8 ± 3.8	**6.7±3.6**	8.8 ± 4.1	N/A
	Videos [720]	8.3 ± 2.9	99.6 ± 25.9	10.5 ± 5.4	35.1±13.7	7.9 ± 3.9	6.7 ± 3.6	12.3 ± 4.3	**6.0±2.7**
	Videos [1440]	7.6 ± 2.8	103.6 ± 20.5	9.4 ± 4.3	44.6±15.5	7.1 ± 3.4	6.6 ± 3.5	10.7 ± 3.4	**6.0±2.9**
Synth.	Images	9.8±4.6	N/A	9.6±5.1	28.2±10.9	11.7±3.8	**7.6±5.0**	10.7±5.4	N/A
	Videos	11.6±4.3	43.5±13.5	11.3±5.0	31.2±12.1	13.5±4.5	9.1±4.7	11.4±5.4	**8.2±4.7**

Table 6. Average errors across all keypoints as a percentage of the Neck-MidHip segment, with standard deviation, for each method.

Redundant detections

The complete table with the percentage of redundant detections is shown in Tab. ST. 7 and ST. 8 for real and synthetic infants respectively.

For real infants, it is difficult to estimate the difficulty of each sequence exactly. We observe that video input has a tendency to produce more redundant detections.

For synthetic infants, we observe that Detectron2 and HRNet RD seem to produce more redundant detections when the sequence is difficult (IDs 10-12), while HRNet BU seems to produce more redundant detections when the sequence is easy (IDs 1-4). We observe a tendency to produce more redundant detections with video input than for image input.

Real	Video ID	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Images	AA_8w	**0**	N/A	11.5	**0**	1.4	1.1	**0**	N/A
	AA_17w	0.6	N/A	8.0	**0**	1.1	40.5	**0**	N/A
	TH_8w_st1	**0**	N/A	9.9	**0**	0.5	3.3	**0**	N/A
	TH_8w_st2	**0**	N/A	34.6	**0**	5.3	22.1	**0**	N/A
	TH_8w_st3	1.4	N/A	34.3	**0**	2.6	12.1	**0**	N/A
	TH_15w	14.4	N/A	86.3	**0**	11.8	103.5	**0**	N/A
	TH_19w	0.4	N/A	28.9	**0**	7.8	53.2	**0**	N/A
	TH_25w	0.4	N/A	7.1	**0**	5.1	54.6	**0**	N/A
Videos	AA_8w	**0**	**0**	92.0	**0**	9	1.0	**0**	1.4
	AA_11w	0.3	**0**	199.8	**0**	0.3	44.9	**0**	0.1
	AA_13w	4.7	**0**	290.9	**0**	4.9	66.7	**0**	4.0
	AA_17w	1.2	**0**	170.1	**0**	1.4	47.7	**0**	6.4
	AA_19w	6.7	**0**	308.6	**0**	10.0	26.4	**0**	4.0
	TH_8w_st1	0.1	**0**	197.0	**0**	0.9	1.9	**0**	2.9
	TH_8w_st2	**0**	**0**	188.7	**0**	6.5	20.6	**0**	38.6
	TH_8w_st3	0.9	**0**	249.2	**0**	2.9	12.7	**0**	26.6
	TH_10w_st1	0.1	**0**	197.0	**0**	0.1	7.7	**0**	22.5
	TH_10w_st2	1.2	**0**	388.0	**0**	3.3	130.0	**0**	82.3
	TH_12w	0.1	**0**	164.4	**0**	0.3	5.4	**0**	5.4
	TH_15w	15.2	**0**	522.8	**0**	13.9	98.8	**0**	28.8
	TH_17w	2.4	**0**	198.9	**0**	9.4	101.2	**0**	40.3
	TH_19w	0.7	**0**	257.6	**0**	8.2	51.5	**0**	2.7
	TH_23w	51.9	**0**	198.0	**0**	2.3	91.0	**0**	43.9
	TH_25w	0.5	**0**	211.3	**0**	4.9	43.1	**0**	36.6

Table 7. Percentage of redundant detections for each method and input type on real infants videos.

Synth.	Video ID	AlphaPose	DeepLabCut	Detectron 2	MediaPipe	HRNet BU	HRNet TD	OpenPose	ViTPose
Images	Syn. 1	0.3	N/A	5.5	**0**	56.8	0.6	**0**	N/A
	Syn. 2	0.5	N/A	1.3	**0**	**0**	0.6	**0**	N/A
	Syn. 3	**0**	N/A	**0**	**0**	**0**	0.2	**0**	N/A
	Syn. 4	**0**	N/A	0.1	**0**	20.7	3.6	**0**	N/A
	Syn. 5	**0**	N/A	**0**	**0**	3.6	1.6	**0**	N/A
	Syn. 6	**0**	N/A	**0**	**0**	**0**	7.1	**0**	N/A
	Syn. 7	**0**	N/A	2.0	**0**	0.1	0.3	**0**	N/A
	Syn. 8	**0**	N/A	0.5	**0**	1.9	0.5	**0**	N/A
	Syn. 9	**0**	N/A	29.0	**0**	6.2	95.4	**0**	N/A
	Syn. 10	0.1	N/A	240.2	**0**	1.9	101.6	**0**	N/A
	Syn. 11	**0**	N/A	102.2	**0**	1.0	0.4	**0**	N/A
	Syn. 12	**0**	N/A	31.7	**0**	7.6	106.8	**0**	N/A
Videos	Syn. 1	0.5	**0**	20.5	**0**	61.8	1.5	**0**	2.6
	Syn. 2	**0**	**0**	38.1	**0**	**0**	2.7	**0**	**0**
	Syn. 3	**0**	**0**	**0**	**0**	0.1	0.1	**0**	**0**
	Syn. 4	1	**0**	92.2	**0**	31.3	15.1	**0**	2.4
	Syn. 5	**0**	**0**	13.6	**0**	9.1	3.7	**0**	3.6
	Syn. 6	**0**	**0**	42.2	**0**	0.9	24.0	**0**	9.2
	Syn. 7	**0**	**0**	98.7	**0**	1.3	0.9	**0**	0.1
	Syn. 8	**0**	**0**	70.7	**0**	2.9	2.2	**0**	**0**
	Syn. 9	0.1	**0**	97.2	**0**	23.7	93.1	**0**	50.4
	Syn. 10	0.4	**0**	378.1	**0**	3.3	30.6	**0**	2.5
	Syn. 11	**0**	**0**	137.1	**0**	1.0	0.8	**0**	74.5
	Syn. 12	0.1	**0**	121.6	**0**	5.8	110.5	**0**	30.5

Table 8. Percentage of redundant detections for each method and input type on synthetic infants.

Supplementary results summary

In a best-case scenario (see Tab. S 3, pose estimation methods car reach high levels of performance on infants in supine position, with low variability, with ViTPose reaching an Average Precision up to 91.8, and average errors around 3.9±2.0% of the Neck-MidHip segment (which corresponds roughly to the infant's torso length).

With the detailed OKS values for each video (Tables ST. 4 and 5), we could get further insight with regards to the complexity estimation made by Hesse et al. on their MINI-RGBD dataset based on the actual performance of the methods' estimations. Such table can also help to identify for each specific method which kind of sequences they seem to struggle with to identify the possible causes of keypoints misplacement, so that more of such examples can be included in future methods' training or fine-tuning. For example, MediaPipe did not manage to identify a single image among the sequence for synthetic infant 2, despite it being deemed "easy", possibly due to its unique background.

From Table 6 of the main manuscripts and Tables of Supplementary Materials ST 2 and ST 3, it seems that the better the estimations and the easier the videos, the lower the correlation between scores and OKS values. This could be explained by a lower range of high values from OKS, while the scores might be less reliable and might not reflect a similar reduction in their variability and range of values.