

Keywords: human pose estimation, semi-automatic annotation, preterm infants, sapiens, videos

Greta DI MARINO ^{1*}, Emanuele CARDINALE ¹, Alessio CORREANI ²,
 Lucia MIGLIORELLI ³, Sara MOCCIA ¹

¹ Università degli Studi "G. d'Annunzio", Italy, greta.dimarino@phd.unich.it, emanuele.cardinale@phd.unich.it, sara.moccia@unich.it

² Università Politecnica delle Marche, Italy, a.correani@staff.univpm.it

³ Università di Teramo, Italy, lmigliorelli@unite.it

* Corresponding author: greta.dimarino@phd.unich.it

Enforcing label consistency and lowering labelling time in infants' pose data via semi-automatic annotation

Abstract

Infants' spontaneous movements provide clinically relevant information about neurodevelopment, but their assessment still relies on qualitative visual inspection, which is prone to variability. Video-based systems with automatic pose estimation algorithms have been proposed to address this, yet annotating data to train these algorithms is time-consuming and prone to intra- and inter-annotator variability. To improve the efficiency and consistency of annotation, semi-automatic annotation has been proposed in the broader human pose estimation literature. To assess the extent to which semi-automatic annotation may also be beneficial for infants' pose estimation, we benchmarked six state-of-the-art adult-trained models on a new dataset comprising 46 videos of preterm infants recorded in a neonatal unit. The predicted joints from the best-performing model were presented to human annotators for optional refinement, who were also asked to label the same joints manually for comparison. ViTPose and Sapiens achieved the highest joint localisation accuracy, with ViTPose requiring markedly lower computational cost. Semi-automatic annotation reduced inter-annotator variability by 7.82% and decreased the time to label a frame by 35.9% compared to full manual labelling. These findings support the use of semi-automatic annotation as an effective strategy to enhance the labelling process of infant pose estimation data from real clinical settings.

1. INTRODUCTION

It is well known in the clinical literature today that the quality, variability, and coordination of infants' spontaneous movements in the days immediately after birth provide early insights into nervous system development (Doi et al., 2023).

In fact, a subset of these spontaneous movements, known as general movements, is widely recognised to be a potential predictor of cerebral palsy and hypoxic-ischemic encephalopathy (Kali et al., 2025). More recently, atypical spontaneous movement patterns, including reduced variability, asymmetry or spasticity, have also been identified as potential early markers of neurodevelopmental disorders, such as autism spectrum disorder and Attention Deficit Hyperactivity Disorder (ADHD) (Einspieler et al., 2014).

Currently, the assessment of infants' movement patterns in neonatal, intensive and sub-intensive care units is performed through direct visual inspection at the crib side by expert neonatologists (Migliorelli et al., 2023b). This procedure is qualitative and may be susceptible to inter- and intra-observer variability (Moccia et al., 2019).

Video-based support systems, coupled with automated infant pose estimation, may offer a promising solution for enabling continuous, objective, and non-invasive monitoring of infants' spontaneous movements, standardising and democratising movement assessment across centres (Migliorelli et al., 2023b). Human pose estimation has gained increasing attention in healthcare applications for contactless movement monitoring.

Recent examples include (He et al., 2024), where RTMPose is used for the detection of genu valgum in adolescents, (Liu et al., 2025), where AlphaPose with YOLOv8s is used for automatic fall detection, (Sato et al., 2019), which uses OpenPose to study the gait of subjects with Parkinson's disease, and

(Feng et al., 2025), where ViTPose and Sapiens are used for continuous neurological monitoring of neurologically impaired subjects.

As for the infants' field, Tab. 1 summarises the most recent works in the literature. OpenPose is used in McCay et al. (2020) for pose estimation on 12 synthetic videos from the MINI-RGBD dataset. A synthetic dataset, called SMaL, is also used in (Kyrollos et al., 2023), where the performance of HRNet and ViTPose is assessed. Despite the encouraging results, the use of synthetic data may limit the representation of the true variability in real infant movements. Approaches tested on real data include the work in (Turner & Sharkey, 2024), which uses MediaPipe for pose estimation as a prior to study the movement of 12 infants of mixed genders aged 3 to 12 months, while in (Jahn et al., 2025) several 2D pose estimators, including ViTPose, are compared on a dataset including 31 participants recorded in a range of age from 4 to 16 weeks. A mix of synthetic and real data is used in Gama et al. (2025) to compare the performance of several pose estimation models. Specifically, the MINI-RGBD dataset is used together with two small home-recorded infant datasets, covering a range of ages from 2 to 18 months, and a subset of the SyRIP dataset (Huang et al., 2021), comprising 500 images of infants aged up to 1 year, collected from YouTube videos and Google Images. Unlike Turner and Sharkey (2024) and Gama et al. (2025), we decided to focus specifically on movements of hospitalised preterm infants recorded at term-equivalent age, rather than on a wider but more heterogeneous population in terms of age. The repertoire of infants' movements evolves quickly, making pose estimation in infants immediately after birth more challenging, as their movements differ greatly from those of toddlers and adults (Doi et al., 2023). Moreover, pose estimation is further complicated in the clinical setting due to the presence of sensors and devices on infants' bodies, which may result in joint occlusion (Moccia et al., 2019).

The closest work to ours is probably Shin et al. (2022), where AlphaPose is used to analyse spontaneous limb movements in 65 preterm infants, recorded at term-equivalent age, either at home or in hospital. However, with respect to Shin et al. (2022), we decided to consider only recordings of hospitalised infants to as closely as possible mimic the visual assessment performed by neonatologists in clinical practice.

Tab. 1. Summary of recent studies on infant pose estimation

Paper	Main architecture	Data
McCay et al. (2020)	OpenPose	12 synthetic videos (MINI-RGBD dataset)
Kyrollos et al. (2023)	HRNet, ViTPose and TransPose	Mannequin videos (SMaL dataset)
Turner & Sharkey (2024)	MediaPipe	12 infants aged 3 to 12 months
Jahn et al. (2025)	Comprehensive benchmark pose estimators (ViTPose best after fine-tuning)	31 infants aged 4 to 16 weeks
Gama et al. (2025)	Comprehensive benchmark pose estimators	MINI-RGBD and SyRIP dataset (infants aged between 2 and 18 months)
Shin et al. (2022)	AlphaPose	65 preterm infants recorded at term-equivalent age at home or in hospital

All the approaches summarized in Tab. 1 rely on supervised models, posing issues relevant to the cost of data annotation, which is time-consuming and may be prone to errors, possibly introducing biases and compromising the reliability of models trained and evaluated on such data (Migliorelli et al., 2023b). To enhance the efficiency and consistency of data annotation in the wider field of human pose estimation, semi-automatic annotation strategies have been proposed as effective tools. These strategies leverage pre-trained pose estimation architectures to provide an annotation proposal to be checked and eventually corrected by human annotators, possibly minimizing annotation time and enforcing inter- and intra-annotator consistency (Sagonas et al., 2013). However, no study has assessed the impact and benefit of semi-automatic annotation in the field of infant pose estimation. This may be partially explained considering the generally small size of infants' datasets studied in the literature, as reported in Tab. 1. In fact, when compared to human pose estimation, working with infants' data presents further issues, such as ethical, legal, and privacy constraints. In (a), infants' video frames are processed through six state-of-the-art pose estimation models: HRNet, RTMPose, AlphaPose, Sapiens, Stacked Hourglass and ViTPose. The keypoints predicted by each model are evaluated against the ground-truth manual annotations to assess their performance. In (b), a set of video frames is labelled by 4 independent annotators in both fully manual and semi-automatic fashions using the best-performing pose estimation model. The two annotation strategies are assessed in terms of average annotation time, Root Mean Square Error (RMSE) relative to the ground truth, and inter-annotator consistency.

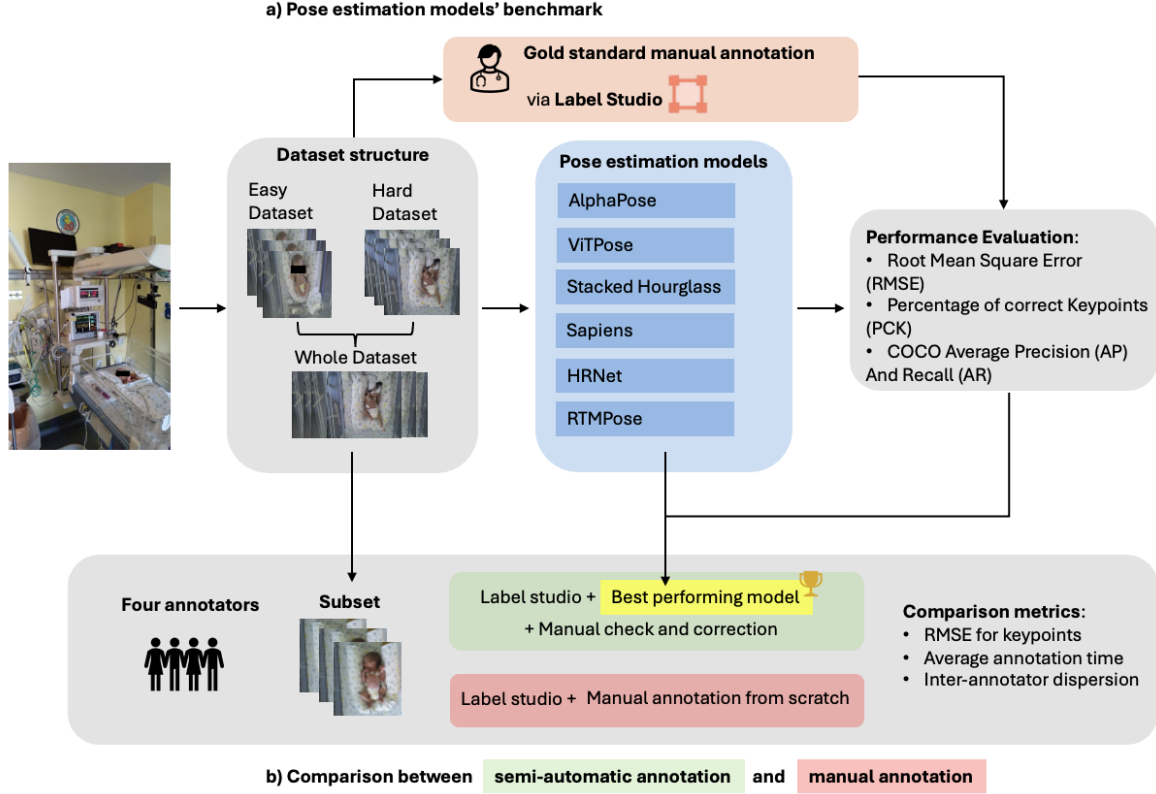


Fig. 1. Workflow of the proposed study.

Building on these considerations, this work examines the impact of semi-automated annotation of infants' joints using human pose estimation models pre-trained on adult data. With a new dataset of 46 videos acquired in the actual clinical practice from 46 preterm infants at term equivalent age, we here conduct a dedicated benchmarking of state-of-the-art pose estimation models: RTMPose (Jiang et al., 2023), ViTPose (Xu et al., 2022), Stacked Hourglass (Newell et al., 2016), HRNet (Sun et al., 2019), AlphaPose (Fang et al., 2022), and the foundation model Sapiens (Khirodkar et al., 2024). We choose these models as they are widely used in the infant literature (Tab. 1). After selecting the best-performing model, we compare semi-automatic and fully manual annotations performed by 4 experts in terms of inter-annotator consistency and average annotation time. This comparison highlights the extent to which pre-trained models can effectively reduce the annotation burden while maintaining annotation of quality standards in the infant domain.

2. METHODS

Figure 1 provides an overview of the workflow of the proposed study. Panel (a) shows the gold-standard annotation procedure obtained through Label Studio. Frames from infant video recordings are selected and carefully annotated by an expert clinician, following the COCO annotation format (Lin et al., 2014), the most popular benchmark for 2D body pose estimation. This format encompasses 17 landmarks, including facial keypoints (eyes, ears, nose) and upper- and lower-limb joints (shoulders, elbows, wrists, hips, knees, ankles). The panel further presents a comparative evaluation of HRNet, RTMPose, AlphaPose, Sapiens, Stacked Hourglass, and ViTPose. All these models follow a top-down pose estimation approach, in which subjects are first localised using bounding boxes. Here, for this task, we use YOLOv11l in an off-the-shelf configuration that achieves state-of-the-art performance on the COCOval2017 dataset, with reported mean Average Precision $mAP@[0.5:0.95]$ of 54.7% (Ultralytics, 2024).

Figure 1(b) shows the experimental phase involving four annotators who were asked to label infant frames both semi-automatically, using keypoints predicted by the best-performing pose estimation model, and manually. The performance of the two strategies is evaluated in terms of average annotation time, Root Mean Square Error (RMSE) per keypoint and inter-annotator dispersion.

2.1. Architectures under investigation

For our benchmark analysis, we assess a set of pre-trained models on our custom dataset. This set includes the models previously tested on infant data in the literature (HRNet, ViTPose, and AlphaPose), as shown in Tab. 1, as well as additional widely used pose estimation models for adults (Hourglass, Sapiens, and RTMPose). Tab. 2 provides a brief overview of the models considered. Among CNN-based models, RTMPose (Jiang et al., 2023) is a real-time top-down estimator with a detection-optimised CSPNext backbone and a classification-based keypoint formulation, offering a strong accuracy-efficiency trade-off. HRNet (Sun et al., 2019) preserves high-resolution features via parallel multi-resolution branches and is widely used in pose estimation research, including in infant pose estimation (Cacciatore et al., 2025; Grafton et al., 2025). Hourglass (Newell et al., 2016) captures multi-scale context through symmetric down- and up-sampling with residual connections and intermediate supervision, and remains a standard benchmark. AlphaPose (Fang et al., 2022) is a real-time, top-down multi-person framework commonly cited in the field of pose estimation. Among transformer-based models, ViTPose (Xu et al., 2022) uses a plain ViT encoder with a lightweight decoder and attains state-of-the-art results on COCO Keypoints, making it a representative baseline. Sapiens (Khrodkar et al., 2024) is a family of high-resolution ViT foundation models, pre-trained with MAE on 300M+ in-the-wild human images and fine-tuned for 2D pose, part segmentation, depth, and surface normals, with reported gains over ViTPose on Human-5K.

Tab. 2. Models' architecture. For the backbone employed in Sapiens (Khrodkar et al., 2024), they clarify that the Sapiens-0.3B model is architecturally similar to the traditional ViT-Large, but it consists of twenty-fold more Floating Point Operation (FLOPs) due to its higher resolution

Model	Backbone	GFLOPs	Input Image Size [pixels]
AlphaPose	FastPose(DUC)	15.99	256x192
RTMPose-m	CSPNext-m	1.93	256x192
HRNet	HRNet-w32	7.71	256x192
Hourglass	Hourglass-52	64.53	384x384
ViTPose-h	ViT-h	122.91	256x192
Sapiens-0.3B	ViT	1242	1024x768

For the backbone employed in Sapiens, (Khrodkar et al., 2024) clarify that the Sapiens-0.3B model is architecturally similar to the traditional ViT-Large, but it consists of twenty-fold more Floating Point Operation (FLOPs) due to its higher resolution

RTMPose, HRNet, Hourglass and ViTPose are available via MMPose (OpenMMLab, n.d.), an open-source toolbox for human pose estimation that provides implementations of various state-of-the-art models, and are trained on the COCO dataset. Sapiens is trained on a proprietary dataset, Humans-300M, and AlphaPose is trained on a combination of COCO and PoseTrack (Andriluka et al., 2018). We selected "huge" variant for ViTPose as in (Jahn et al., 2025), "medium" for RTMPose, "w32" for HRNet as in (Cacciatore et al., 2025). Due to computational limitations, we adopted the "Sapiens-0.3B"

2.2. Dataset

In this study, we collect and curate a dataset of 46 RGB video recordings of 46 infants born preterm, recorded at term-equivalent age or just before discharge in the neonatology department of the G. Salesi Hospital in Ancona, Italy. The dataset shows high variability in gestational age (mean = 31.87 ± 3.77), weight (mean = 2021 ± 790), and length (mean = 44.13 ± 4.12). The infants were identified by clinicians among those in stable condition and capable of spontaneous movement.

The acquisition setup is designed to minimise interference with clinical workflows, as shown in Figure 2. A top-mounted Astra Mini S-Orbbec RGB-D camera is positioned approximately 50 cm above the crib. Each video is recorded at 30 frames per second (fps) with a spatial resolution of 640×480 pixels. For each infant, a continuous five-minute recording is obtained. From this, a 20-second segment is extracted, comprising 10 seconds of spontaneous movement and 10 seconds of rest. To speed the annotation process, videos are downsampled to 5 fps, resulting in 120 frames per infant, and a total of 5,520 frames across the dataset. To generate high-quality ground-truth labels for pose estimation, an expert annotator is asked to label the images using Label Studio (HumanSignal, n.d.).



Fig. 2. RGB-D video acquisition setup. The RGB-D camera is positioned at about 50 cm above the infant’s crib and does not hinder health operator movements

To account for varying levels of movement visibility and occlusion, we divide the dataset into two subsets: easy and hard (Figure 3). The easy subset comprises infants lying in the crib with unobstructed views and free limb movement. It is characterised by reduced occlusions, which allow for reliable movement analysis. Conversely, the hard subset includes recordings in which infants are partially occluded by external factors such as medical devices (e.g., oxygen masks), blankets, or orthopaedic casts. These factors pose significant visual challenges for pose estimation by obscuring limb visibility. Sequences in this subset may not be suitable for downstream movement analysis but serve as a realistic benchmark for assessing model robustness in clinical scenarios.



Fig. 3. Samples of infant images from our dataset. The first row shows infant images from the hard subset, characterised by medical devices, casts, and several occlusions. The second row shows infants from the easy subset, who generally have clearer visibility of limbs and joints.

2.3. Metrics

To quantitatively assess the performance of human pose estimation models, we adopt evaluation metrics that compare the predicted keypoints against the gold-standard manual annotations: Percentage of Correct Keypoints (PCK), RMSE, COCO Average Precision (AP), and COCO Average Recall (AR).

PCK is a standard benchmark in human pose estimation, measuring the proportion of predicted keypoints that fall within a specified distance of the corresponding ground-truth keypoints.

A predicted keypoint $\hat{k}_i$ is considered correct if its Euclidean distance from the ground-truth keypoint k_i is less than a threshold proportional to a reference length l :

$$\|\hat{k}_i - k_i\| < \alpha \cdot l \quad (1)$$

Where alpha is the tolerance threshold that sets the maximum allowable error as a fraction of l , a subject-specific normalisation parameter, equal to the infant's torso length, introduced for fair comparison across infants of different sizes. The torso length is measured as the Euclidean distance between the base of the neck (midpoint of the left and right shoulders) and the hip centre (midpoint of the left and right hips). Four values of alpha are selected: 0.05, 0.1, 0.15, and 0.2.

The COCO AP is calculated as the area under the precision-recall curve, averaged over multiple Object Keypoint Similarity (OKS) thresholds (Maji et al., 2022), ranging from 0.50 to 0.95 in increments of 0.05. We also included AP values at commonly used thresholds of 0.75 and 0.5. Similar to the COCO AP, the COCO AR metric computes the mean AR across multiple thresholds of OKS, ranging from 0.5 to 0.95 with a step size of 0.05, as well as COCO AR values at thresholds of 0.75 and 0.5.

To compare manual and semi-automatic annotations, RMSE and the dispersion of annotations across annotators for each keypoint are used. RMSE quantifies the average deviation between predicted $\hat{k}_i = (\hat{x}_i, \hat{y}_i)$ and ground-truth $k_i = (x_i, y_i)$ keypoint locations in pixels. The RMSE over N keypoints is computed as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\hat{k}_i - k_i\|^2} \quad (2)$$

Dispersion is defined as the average Euclidean distance of the annotated keypoints from their centroid. For a given keypoint k_p , let $p_i = (x_i, y_i)$ be the coordinates provided by annotator i and N the number of annotators. The centroid c is then computed as the mean over these coordinates.

3. RESULTS AND DISCUSSION

Tab. 3 reports PCK values computed at torso-normalised thresholds of 0.05, 0.10, 0.15 and 0.20, and Tab. 4 presents COCO AP and COCO AR following the standard multi-threshold protocol (0.5:0.05:0.95) and their single-threshold variants (AP@0.50, AP@0.75; AR@0.50, AR@0.75) across the easy subset, the hard subset and the whole dataset. As expected, performance is consistently higher in the easy subset, which is characterised by homogeneous backgrounds, minimal occlusions, and controlled recording conditions that facilitate accurate keypoint localisation. Conversely, the hard subset incorporates recordings characterised by heterogeneous backgrounds, frequent occlusions, and greater variability in infant posture and spontaneous movement, resulting in lower performance. Both subsets, easy and hard, originate from real clinical acquisitions and therefore capture complementary aspects of the neonatal environment; the combined whole dataset offers the most representative benchmark of model performance.

Comparing our findings with previous infant pose-estimation studies summarised in Tab. 1, Jahn et al. (2025) reported ViTPose-huge as the best-performing model, with $PCK@0.10 = 0.846$ on a dataset comprising 31 infants aged 4 to 16 weeks. In our dataset, ViTPose also achieved the best performance among State-of-the-art models, with $PCK@0.10 = 0.784$ on the whole dataset. Similarly, Gama et al. (2025) evaluated across infant datasets with varying levels of complexity, including their real Supine dataset, the real Lap dataset, the SyRIP dataset, and the synthetic MINI-RGBD dataset. Among these, the real Supine dataset, composed of videos of infants aged 2 to 6 months, is the closest to our dataset in terms of infant posture. In this setting, they reported AP/AR values of 0.885/0.909 for ViTPose, while in our clinically acquired dataset, ViTPose achieved $AP = 0.804$ and $AR = 0.826$. For AlphaPose, Gama et al. (2025) reported AP/AR values of 0.675/0.743, whereas in our dataset, AlphaPose achieved $AP = 0.768$ and $AR = 0.794$. Therefore, our results are consistent with the current state of the art. The slightly lower performance of ViTPose can be explained by the more challenging nature of our dataset, which includes 46 hospitalised preterm infants recorded in a real neonatal unit, with clinical equipment, blankets, casts, oxygen masks, heterogeneous backgrounds, and partial occlusions.

Tab. 3. Proportion of Corrected Keypoints (PCK) measured at different torso length thresholds for the whole (easy + hard), hard and easy subsets (mean Standard Deviation (SD)), respectively

Dataset	Model	PCK@0.05	PCK@0.10	PCK@0.15	PCK@0.20
Easy + Hard	HRNet	0.330 (0.125)	0.639 (0.204)	0.763 (0.218)	0.824 (0.213)
Easy + Hard	Hourglass	0.386 (0.123)	0.675 (0.187)	0.775 (0.200)	0.827 (0.199)
Easy + Hard	RTMPose	0.410 (0.139)	0.739 (0.192)	0.863 (0.185)	0.913 (0.165)
Easy + Hard	AlphaPose	0.415 (0.120)	0.757 (0.155)	0.878 (0.154)	0.926 (0.142)
Easy + Hard	Sapiens	0.435 (0.104)	0.771 (0.134)	0.890 (0.133)	0.939 (0.117)
Easy + Hard	ViTPose	0.432 (0.117)	0.784 (0.156)	0.907 (0.145)	0.950 (0.125)
Hard	HRNet	0.279 (0.147)	0.552 (0.237)	0.687 (0.261)	0.755 (0.261)
Hard	Hourglass	0.333 (0.147)	0.599 (0.220)	0.704 (0.234)	0.767 (0.234)
Hard	AlphaPose	0.349 (0.105)	0.682 (0.134)	0.835 (0.140)	0.894 (0.137)
Hard	RTMPose	0.350 (0.161)	0.662 (0.203)	0.811 (0.181)	0.892 (0.160)
Hard	Sapiens	0.393 (0.085)	0.717 (0.065)	0.861 (0.041)	0.936 (0.032)
Hard	ViTPose	0.376 (0.125)	0.718 (0.139)	0.868 (0.095)	0.939 (0.063)
Easy	HRNet	0.354 (0.104)	0.680 (0.171)	0.800 (0.182)	0.857 (0.176)
Easy	Hourglass	0.411 (0.100)	0.711 (0.157)	0.809 (0.171)	0.856 (0.172)
Easy	RTMPose	0.440 (0.117)	0.775 (0.175)	0.888 (0.182)	0.924 (0.167)
Easy	AlphaPose	0.447 (0.113)	0.793 (0.152)	0.898 (0.156)	0.942 (0.142)
Easy	Sapiens	0.455 (0.106)	0.797 (0.149)	0.904 (0.158)	0.941 (0.140)
Easy	ViTPose	0.459 (0.103)	0.816 (0.154)	0.925 (0.160)	0.956 (0.145)

Tab. 4. COCO AP (Average Precision) and COCO AR (Average Recall) across multiple OKS (Object Keypoint Similarity) thresholds for the whole (easy + hard), hard and easy subsets (mean Standard Deviation (SD)), respectively

Model	AP@[0.5:0.05:0.95]	AP@0.75	AR@[0.5:0.05:0.95]	AR@0.75
Easy + Hard - Hourglass	0.671 (0.244)	0.793 (0.305)	0.692 (0.245)	0.805 (0.299)
Easy + Hard - HRNet	0.660 (0.241)	0.781 (0.308)	0.679 (0.244)	0.794 (0.304)
Easy + Hard - RTMPose	0.757 (0.216)	0.871 (0.275)	0.777 (0.217)	0.883 (0.265)
Easy + Hard - AlphaPose	0.768 (0.184)	0.905 (0.217)	0.794 (0.171)	0.922 (0.190)
Easy + Hard - Sapiens	0.784 (0.152)	0.943 (0.176)	0.810 (0.146)	0.954 (0.159)
Easy + Hard - ViTPose	0.804 (0.160)	0.943 (0.180)	0.826 (0.158)	0.951 (0.169)
Hard - HRNet	0.572 (0.270)	0.673 (0.356)	0.596 (0.273)	0.706 (0.350)
Hard - Hourglass	0.604 (0.254)	0.739 (0.334)	0.629 (0.254)	0.762 (0.326)
Hard - AlphaPose	0.667 (0.196)	0.802 (0.253)	0.710 (0.164)	0.851 (0.196)
Hard - RTMPose	0.693 (0.208)	0.807 (0.303)	0.720 (0.207)	0.839 (0.282)
Hard - Sapiens	0.719 (0.116)	0.909 (0.171)	0.759 (0.092)	0.936 (0.114)
Hard - ViTPose	0.737 (0.136)	0.897 (0.179)	0.768 (0.131)	0.918 (0.146)
Easy - HRNet	0.702 (0.214)	0.832 (0.268)	0.720 (0.217)	0.837 (0.269)
Easy - Hourglass	0.703 (0.232)	0.819 (0.286)	0.723 (0.235)	0.826 (0.283)
Easy - RTMPose	0.788 (0.213)	0.902 (0.254)	0.805 (0.216)	0.904 (0.254)
Easy - Sapiens	0.816 (0.157)	0.960 (0.176)	0.835 (0.159)	0.962 (0.176)
Easy - AlphaPose	0.817 (0.156)	0.955 (0.177)	0.835 (0.159)	0.957 (0.177)
Easy - ViTPose	0.837 (0.160)	0.966 (0.176)	0.854 (0.162)	0.966 (0.176)

Statistical analysis is conducted using post hoc Wilcoxon pairwise tests with the Holm correction. P-values are provided in the Supplementary material (see Tab. 1). ViTPose significantly outperforms Hourglass and HRNet across all sets (whole, easy, and hard) and all metrics. By contrast, comparisons with AlphaPose and RTMPose are threshold-dependent; differences are significant under more tolerant thresholds, PCK@0.10-0.20, AP@0.75, and AR@0.75. Multi-threshold aggregates are significant relative to AlphaPose but not to RTMPose. Sapiens shows mostly non-significant differences across metrics, with significance limited to PCK@0.15 ($p=0.0048$) and PCK@0.20 ($p=0.0046$). The statistical significance of the differences between the easy and hard subsets is evaluated using two-sided Mann-Whitney U tests with Holm-Bonferroni correction (see Tab. 2 in Supplementary). ViTPose, RTMPose, AlphaPose, and Sapiens all exhibit significant differences between easy and hard subsets (corrected p -value < 0.05). In contrast, HRNet and Hourglass show no

significant differences (corrected p -value > 0.05), which also aligns with their relatively poor overall performance.

Figure 4 reports the mean RMSE per keypoint for the easy subset, hard subset, and whole dataset, as shown in Figure 5. To test for significant global differences in RMSE across models for each anatomical keypoint, we perform a nonparametric Friedman test. Statistically significant differences are found for most keypoints after Holm correction, with exceptions for the Left Eye and Left Ear keypoints (p -value = 0.2598 and p -value = 0.2571, respectively), as illustrated in Tab. 4 in the Supplementary.

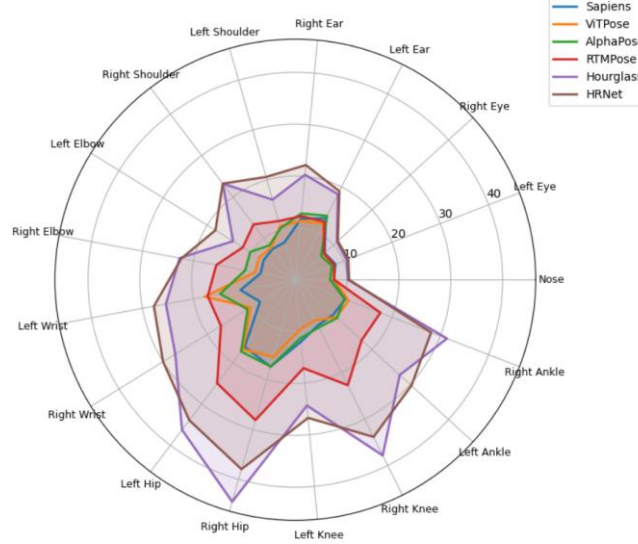


Fig. 4. Radar plot of Root Mean Square Error (RMSE) values on the whole dataset for each of the 17 keypoints and each evaluated model

In accordance with the findings outlined above, ViTPose generally achieves higher PCK, COCO AP, and COCO AR values. Sapiens yields lower per-keypoint RMSE, suggesting fewer large errors and more stable localisation. However, these RMSE advantages are mostly not statistically significant after correction and are therefore insufficient to claim superiority over ViTPose.

Sapiens’ advantage can be attributed to the ViT architecture pre-trained on the large-scale, multi-task Humans-300M dataset, which facilitates fine-grained spatial representation. However, this comes at a higher computational cost, making Sapiens less suitable for scalable or real-time NICU deployment. All models exhibit larger errors in distal joints -hips, knees, ankles, and wrists- stemming from the combined effects of motion blur and frequent occlusions. Hips are additionally challenging because the diaper masks anatomical landmarks, increasing ambiguity. Ears are also difficult to localize because infants often present with the head rotated; therefore, the target ear can be partially occluded or visible under an extreme pose, as qualitatively shown in Figure 6. Consistent with this anatomy-dependent difficulty, the keypoint-wise Friedman analysis shows no significant differences ($p \geq 0.05$) for the left eye and left ear, indicating that models behave similarly where the signal is weak or intermittently invisible, whereas significant differences emerge for clearer, less-occluded regions (e.g., shoulders, elbows, knees).

Our analysis does not reveal any significant discrepancies between ViTPose and other advanced CNN-based architectures, such as RTMPose and AlphaPose. RTMPose improves keypoint localisation by reformulating it as a classification task with ordinal targets and by employing a lightweight prediction head to mitigate the effects of heatmap quantisation. AlphaPose, instead, relies on a robust top-down pipeline that incorporates resilient crop normalisation and pose-level de-duplication, enhancing robustness to detector noise and facilitating cross-domain transfer. In contrast, statistically significant differences emerge between ViTPose and more conventional convolutional models such as HRNet and Stacked Hourglass, which struggle particularly under occlusions and complex postures, likely due to their limited capacity to model global spatial dependencies across joints. Collectively, these findings emphasise the robustness of transformer-based approaches: through the usage of global self-attention mechanisms, they preserve structural consistency even in challenging scenarios, making them the most reliable option.

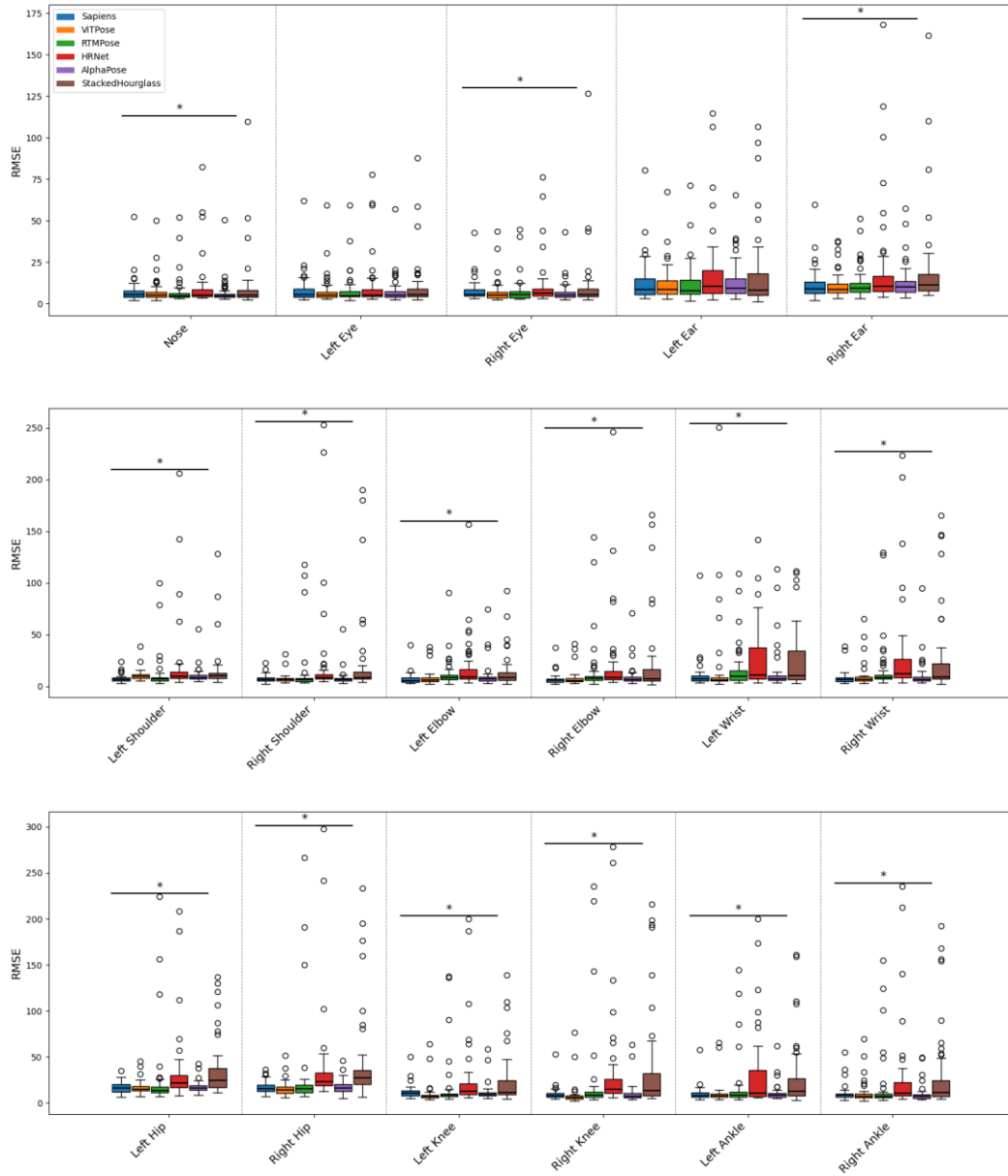


Fig. 5. Root Mean Square Error (RMSE) box plot over the whole dataset for (up) facial, (mid) upper limbs and (down) lower limbs keypoints.



Fig. 6. Qualitative results of the tested models (RTMPose, HRNet, Stacked Hourglass, ViTPose, Sapiens, AlphaPose) and ground-truth annotation for a (left) hard and (right) easy sample.

Across all recordings, the camera-to-crib distance is kept approximately constant (about 50 cm). This distance is chosen to ensure the infant remains within the field of view while respecting the spatial constraints imposed by the overhead infant-warming system above the crib. Within this controlled acquisition range, we do not observe any difference in pose estimation performance associated with slight changes in the acquisition distance between the camera and the recorded infant. We may expect that, given the prior localisation of an infant in the image before proceeding with pose estimation, even at different distances, we should not see major problems.

Figure 7 shows per-keypoint RMSE across the four annotators, for manual and semi-automatic annotation, respectively. Manual annotations (Figure 7 – upper part) yield lower RMSE values than semi-automatic annotations (Figure 7 – lower part) when compared against the gold standard. This finding indicates that fully manual placement yielded higher overall precision. The largest errors observed in both strategies pertain to the hips and ears, reflecting the inherent challenges associated with localising these keypoints: the hips are characterised by the absence of clearly visible anatomical boundaries, while the ears are frequently obscured during lateral head rotation.

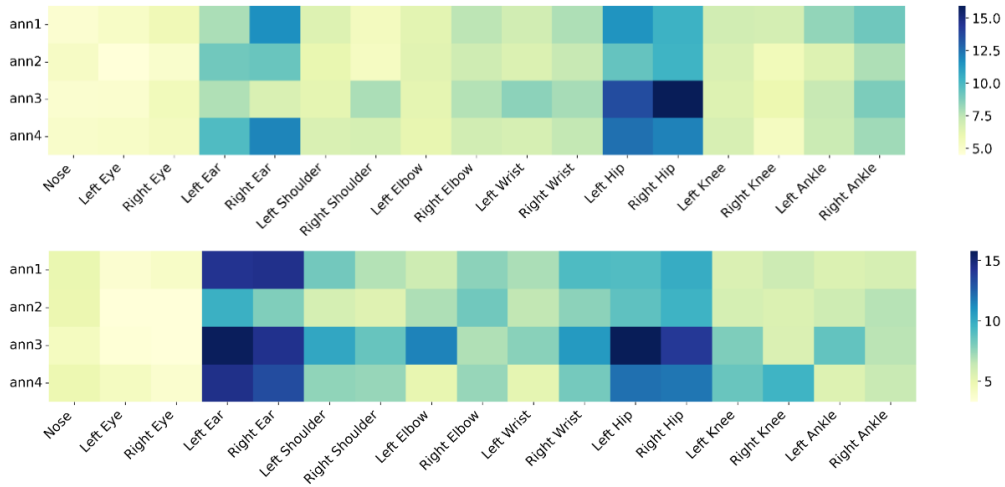


Fig. 7. Root Mean Square Error (RMSE) for (up) manual annotations and (down) semi-automatic annotations with respect to the ground-truth

Figure 8 shows a box plot of dispersion for manual and semi-automatic annotations. Per-frame dispersion scores are calculated by averaging the dispersion of all keypoints. Inter-annotator dispersion is significantly higher in manual annotations than in semi-automatic ones (Wilcoxon test, $p = 0.026$), with semi-automatic annotations reducing inter-annotator variability by 7.82% RA-2, whereas they yield higher RMSE. A per-keypoint dispersion plot is provided in Supplementary Figure 1. A likely explanation is an anchoring bias: when presented with model-generated keypoints, annotators may adjust them only minimally, accepting slightly inaccurate predictions as “good enough”.

Manual annotation from scratch encourages more independent placement, which, despite greater variability among annotators, tends to align more closely with the ground truth on average. This trade-off should be considered when creating datasets in neonatal pose estimation: for large-scale annotations where speed and consistency are critical, semi-automatic initialisation may be advantageous, while for tasks requiring maximal accuracy, manual annotation may still be preferable.

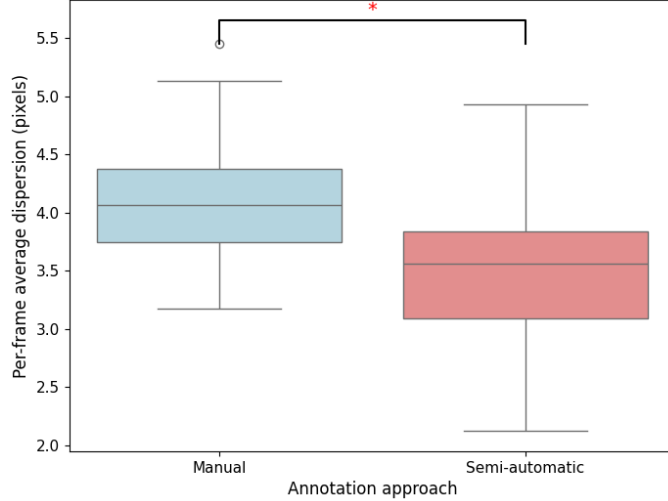


Fig. 8. Inter-annotator per-frame dispersion for the manual (left) and for the semi-automatic (right) approaches

As shown in Tab. 5, the implementation of semi-automatic annotation results in a relevant reduction in the average processing time per frame, from 72.1 seconds to 46.2 seconds, corresponding to a 35.9% decrease (i.e., 25.9 seconds saved per frame).

Tab. 5. Average annotation time per frame (in seconds) for each annotator, comparing fully manual annotation versus semi-automatic annotation (60 frames per annotator). Values are reported as mean ($\pm$ standard deviation)

Annotator	Manual (s/frame)	Semi-automatic (s/frame)
Ann 1	83.1	61.4
Ann 2	68.6	28.4
Ann 3	57.7	49.6
Ann 4	79.1	45.5
Average	72.1 ($\pm$ 9.9)	46.2 ($\pm$ 11.8)

Although this improvement may appear limited at the single-frame level, it becomes relevant when scaled to realistic annotation workloads. For the entire dataset, the estimated annotation time decreases from 110.55 hours with fully manual annotation to 70.84 hours with semi-automatic annotation. Thus, the semi-automatic approach allows clinicians’ involvement to shift from time-consuming annotation from scratch to review and validation of predicted keypoints, making the workflow more compatible with limited clinical time availability, in line with current social considerations about clinician burden and burn-out (Bishop et al., 2023).

4. CONCLUSION

This work represents a benchmark for obtaining frame-based ground-truth annotations of infant videos, a crucial preliminary step for subsequent movement analysis aimed at the early assessment of neurodevelopmental disorders. We benchmarked six state-of-the-art human pose estimation models on infant data and evaluated the benefits of using the best-performing model for semi-automatic joint annotation. Our results show that ViTPose and Sapiens provided the best performance in keypoint localisation, with Sapiens having a higher computational cost. We further show that semi-automatic annotation reduced inter-annotator variability and annotation time compared to manual labelling.

The main limitation of this study may be seen in the dataset structure, which is collected at a single clinical centre and includes a relatively small number of infants. Other neonatal units and patient populations might show different findings, since pose estimation performance may depend on infant size, background conditions, and acquisition setups (camera viewpoint, lighting conditions, occlusions (Migliorelli et al., 2023a)). Moreover, the single-centre design may introduce biases related to the local clinical environment, acquisition protocol, and patient characteristics. To address these limitations, future work will focus on external validation on multi-centre datasets and on bias analyses across acquisition conditions and infant subgroups. We will also explore domain adaptation and privacy-preserving federated learning strategies to improve model robustness

across clinical centres while limiting the need for direct data sharing. Nevertheless, as shown in Tab. 1, our dataset is the largest currently available dataset that focuses exclusively on videos of preterm infants recorded in neonatal settings, thereby mimicking how movement monitoring is performed in clinical practice. Acquiring such data involves several legal and privacy constraints, as well as practical challenges associated with recording a highly fragile population in critical clinical conditions. These factors pose management challenges for healthcare professionals and caregivers. Furthermore, spontaneous movements must be recorded when newborns are awake and unaffected by external factors (e.g., pacifier, toys); these recording windows are limited and can often be interrupted by interventions from healthcare professionals or the presence of parents. To further enhance the potential of our study, we are working to obtain ethical approval to collect a second, larger dataset at another clinical institution in Italy, where longitudinal neurodevelopmental follow-up is routinely performed. Future work will investigate pose-derived movement features for assessing neurodevelopmental risk. Movement patterns may, in fact, be derived by tracking keypoint trajectories across consecutive frames, enabling a lightweight pipeline with low computational cost that supports privacy-preserving edge deployment, in accordance with ethics guidelines for trustworthy AI and the General Data Protection Regulation (GDPR). Beyond ASD, the proposed annotation workflow is also applicable to other populations, such as those with cerebral palsy, as shown in Letzkus et al. (2024). We will also consider exploring image retrieval strategies (e.g., Shamsipour et al., 2024) to automatically select relevant video clips based on content and semantics.

Funding

This research was funded by the ITALIAN FUND FOR APPLIED SCIENCES (FISA), grant number: FISA2022-00696.

Acknowledgments

The authors acknowledge ISCRA for granting access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking and hosted by CINECA (Italy).

Conflicts of interest

No potential competing interest was reported by the authors.

Ethics Approval and Consent to Participate

All recordings were conducted in accordance with hospital protocol, and written informed consent was obtained from each infant’s legal guardian. The study received the approval of the Ethics Committee of the “Ospedali Riuniti di Ancona”, Italy (ID: Prot. 2019-399).

Data Availability

The dataset used in this study will be made available upon request.

REFERENCES

- Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018). PoseTrack: A benchmark for human pose estimation and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5167–5176). <https://doi.org/10.1109/CVPR.2018.00542>
- Bishop, C. E., Machut, K. Z., Dammann, C. E., Cuevas Guaman, M., Miller, E. R., & Lakshminrusimha, S. (2023). Academic neonatologist—a species at the brink of extinction? *Journal of Perinatology*, 43(12), 1526–1529. <https://doi.org/10.1038/s41372-023-01803-4>
- Cacciatore, A., Berardini, D., Scaraggi, V., Mancini, A., Moccia, S., & Migliorelli, L. (2025). Online knowledge distillation and deep supervision in HRNet: Green AI for preterm infants’ pose estimation. *ACM Transactions on Computing for Healthcare*, 6(4), 1–19. <https://doi.org/10.1145/3757067>

- Doi, H., Furui, A., Ueda, R., Shimatani, K., Yamamoto, M., Sakurai, K., Mori, C., & Tsuji, T. (2023). Spatiotemporal patterns of spontaneous movement in neonates are significantly linked to risk of autism spectrum disorders at 18 months old. *Scientific Reports*, 13(1), Article 13869. <https://doi.org/10.1038/s41598-023-40368-2>
- Einspieler, C., Sigafos, J., Bartl-Pokorny, K. D., Landa, R., Marschik, P. B., & Bölte, S. (2014). Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or Rett syndrome. *Research in Autism Spectrum Disorders*, 8(3), 286–291. <https://doi.org/10.1016/j.rasd.2013.12.013>
- Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., & Lu, C. (2022). Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), 7157–7173. <https://doi.org/10.1109/TPAMI.2022.3222784>
- Feng, R., Richter, F., Mari, E., Gleason, A., Le, C., Kellner, C., Shrivastava, R., Fields, M., Rapoport, B., Bederson, J., Schadt, E. E., Glicksberg, B. S., Richter, F., & Dangayach, N. (2025). Pose AI prediction of neurological status in the Neuroscience Intensive Care Unit. *MedRxiv*.
- Gama, F., Misař, M., Navara, L., Popescu, S. T., & Hoffmann, M. (2025). Automatic infant 2D pose estimation from videos: Comparing seven deep neural network methods. *Behavior Research Methods*, 57(10), Article 280. <https://doi.org/10.3758/s13428-025-02816-x>
- Grafton, A., Warnecke, J. M., Li, M., He, E., Thomson, L., Beardsall, K., & Lasenby, J. (2025). Neonatal pose estimation in the unaltered clinical environment with fusion of RGB, depth and IR images. *npj Digital Medicine*, 8(1), Article 539. <https://doi.org/10.1038/s41746-025-01929-z>
- He, S., Wei, M., Meng, D., Lv, Z., Guo, H., Yang, G., & Wang, Z. (2024). Adversarially trained RTMpose: A high-performance, non-contact method for detecting Genu valgum in adolescents. *Computers in Biology and Medicine*, 183, Article 109214. <https://doi.org/10.1016/j.compbiomed.2024.109214>
- Huang, X., Fu, N., Liu, S., & Ostadabbas, S. (2021). Invariant representation learning for infant pose estimation with small data. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1–8). <https://doi.org/10.1109/FG52635.2021.9667056>
- HumanSignal. (n.d.). *Label Studio: Open source data labeling and AI evaluation*. <https://labelstud.io/>
- Jahn, L., Flüge, S., Zhang, D., Poustka, L., Bölte, S., Wörgötter, F., Marschik, P. B., & Kulvicius, T. (2025). Comparison of markerless 2D image-based methods for infant pose estimation. *Scientific Reports*, 15(1), Article 12148. <https://doi.org/10.1038/s41598-025-96206-0>
- Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., & Chen, K. (2023). RTMPose: Real-time multi-person pose estimation based on MMPose. *ArXiv, abs/2303.07399*. <https://doi.org/10.48550/arXiv.2303.07399>
- Kali, G. T., Du Preez, J. C., Van Zyl, J. I., Burger, M., Katsabola, H., & Pepper, M. S. (2025). Predicting cerebral palsy and 18 month neurodevelopmental outcome in infants with presumed Hypoxic Ischaemic Encephalopathy: Role of General Movements Assessment and early neurological examination. *Frontiers in Pediatrics*, 13, Article 1638584. <https://doi.org/10.3389/fped.2025.1638584>
- Khironkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., & Saito, S. (2024). Sapiens: Foundation for human vision models. *ArXiv, abs/2408.12569*. <https://doi.org/10.48550/arXiv.2408.12569>
- Kyrollos, D. G., Fuller, A., Greenwood, K., Harrold, J., & Green, J. R. (2023). Under the cover infant pose estimation using multimodal data. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–12. <https://doi.org/10.1109/TIM.2023.3244220>
- Letzkus, L., Pulido, J. V., Adeyemo, A., Baek, S., & Zanelli, S. (2024). Machine learning approaches to evaluate infants' general movements in the writhing stage—a pilot study. *Scientific Reports*, 14(1), Article 4522. <https://doi.org/10.1038/s41598-024-54297-1>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48
- Liu, L., Sun, Y., Li, Y., & Liu, Y. (2025). A hybrid human fall detection method based on modified YOLOv8s and AlphaPose. *Scientific Reports*, 15(1), Article 2636. <https://doi.org/10.1038/s41598-025-86429-6>
- Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2636–2645). <https://doi.org/10.1109/CVPRW56347.2022.00297>
- McCay, K. D., Ho, E. S., Shum, H. P., Fehringer, G., Marcroft, C., & Embleton, N. D. (2020). Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access*, 8, 51582–51592. <https://doi.org/10.1109/ACCESS.2020.2980269>
- Migliorelli, L., Cacciatore, A., Ottaviani, V., Berardini, D., Dellacà, R. L., Frontoni, E., & Moccia, S. (2023a). TwinEDA: A sustainable deep-learning approach for limb-position estimation in preterm infants' depth images. *Medical & Biological Engineering & Computing*, 61(2), 387–397. <https://doi.org/10.1007/s11517-022-02724-z>
- Migliorelli, L., Tiribelli, S., Cacciatore, A., Giovanola, B., Frontoni, E., & Moccia, S. (2023b). Accountable deep-learning-based vision systems for preterm infant monitoring. *Computer*, 56(8), 84–93. <https://doi.org/10.1109/MC.2023.3235987>
- Moccia, S., Migliorelli, L., Carnielli, V., & Frontoni, E. (2019). Preterm infants' pose estimation with spatio-temporal features. *IEEE Transactions on Biomedical Engineering*, 67(8), 2370–2380. <https://doi.org/10.1109/tbme.2019.2961448>
- Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision* (pp. 483–499). Springer. https://doi.org/10.1007/978-3-319-46484-8_29
- OpenMMLab. (n.d.). *Welcome to MMPose's documentation* (wersja 1.3.2). <https://mmpose.readthedocs.io/en/latest/mmpose.readthedocs>
- Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). A semi-automatic methodology for facial landmark annotation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 896–903). <https://doi.org/10.1109/CVPRW.2013.132>
- Sato, K., Nagashima, Y., Mano, T., Iwata, A., & Toda, T. (2019). Quantifying normal and parkinsonian gait features from home movies: Practical application of a deep learning-based 2D pose estimator. *PLoS ONE*, 14(11), Article e0223549. <https://doi.org/10.1371/journal.pone.0223549>

- Shamsipour, G., Fekri-Ershad, S., Sharifi, M., & Alaei, A. (2024). Improve the efficiency of handcrafted features in image retrieval by adding selected feature generating layers of deep convolutional neural networks. *Signal, Image and Video Processing*, 18(3), 2607–2620. <https://doi.org/10.1007/s11760-023-02934-z>
- Shin, H. I., Shin, H.-I., Bang, M. S., Kim, D.-K., Shin, S. H., Kim, E.-K., Kim, Y.-J., Lee, E. S., Park, S. G., Ji, H. M., & Lee, W. H. (2022). Deep learning-based quantitative analyses of spontaneous movements and their association with early neurological development in preterm infants. *Scientific Reports*, 12(1), Article 3138. <https://doi.org/10.1038/s41598-022-07139-x>
- Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5693–5703). <https://doi.org/10.1109/CVPR.2019.00584>
- Turner, A., & Sharkey, D. (2024). Enhanced infant movement analysis using transformer-based fusion of diverse video features for neurodevelopmental monitoring. *Sensors*, 24(20), Article 6619. <https://doi.org/10.3390/s24206619>
- Ultralytics. (2024). *Ultralytics YOLO11*. <https://docs.ultralytics.com/models/yolo11/>
- Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35, 38571–38584. <https://doi.org/10.52202/068431-2795>