8. Supplementary Materials

In this section we offer figures and analyses supplementing our discussion from the main paper. Fig. 9 and Fig. 10 provide overviews of the human rater survey data, and the pose-based angle assessments, respectively. Fig. 11 shows further skeleton visualizations of human and machine assessments of symmetry, highlighting comparative differences. Finally, Table 6 contains one more regression variance analysis. These results further reinforce our findings regarding the advantages of 3D pose-based symmetry assessment over both human ratings (individual or in aggregate) and 2D pose-based systems.

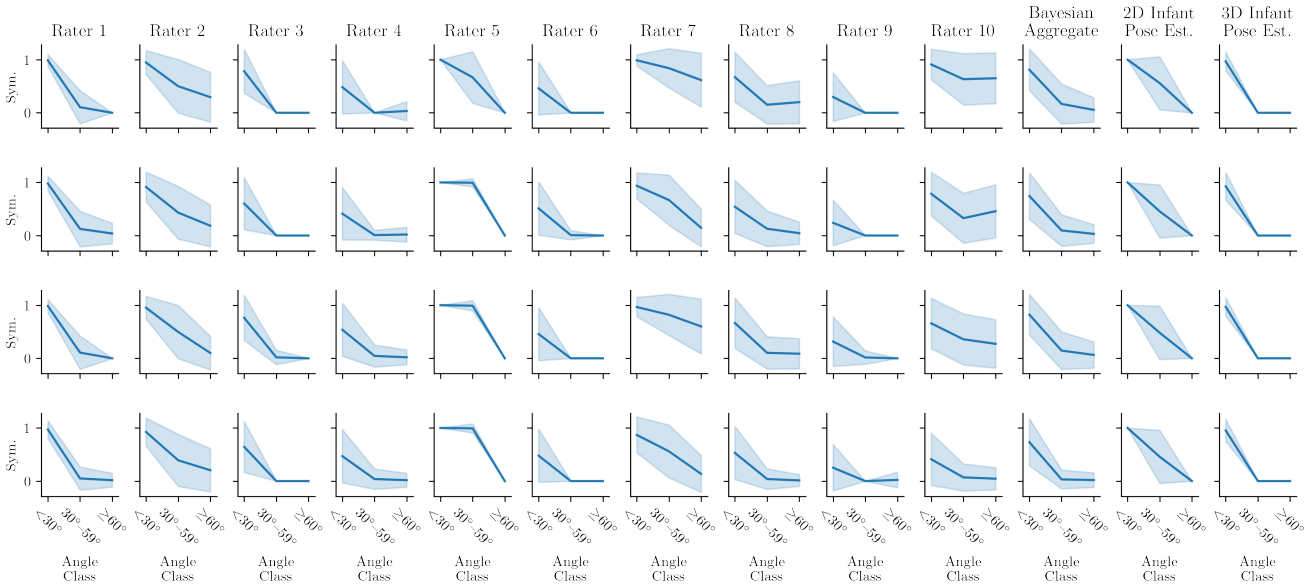

Figure 9: Mean assessments of symmetry level by assessed angle difference level, for all 10 raters plus the Bayesian aggregate rater and the predicted 2D- and 3D-pose-based models. Means are taken over all 700 SyRIP infant images per each of four pairs of limbs, with confidence intervals of one standard deviation at each angle level indicated. These statistics reveal wide variance in determination of symmetry versus angle difference across human raters, although most raters are fairly consistent across limbs. An upwardly sloped segment, as seen most prominently in Rater 8's upper arm assessments and or Rater 10's lower arm assessments, indicates an apparent inconsistency in aggregate assessments. Note that while small confidence intervals indicate consistent assessments, the converse does not necessarily hold.

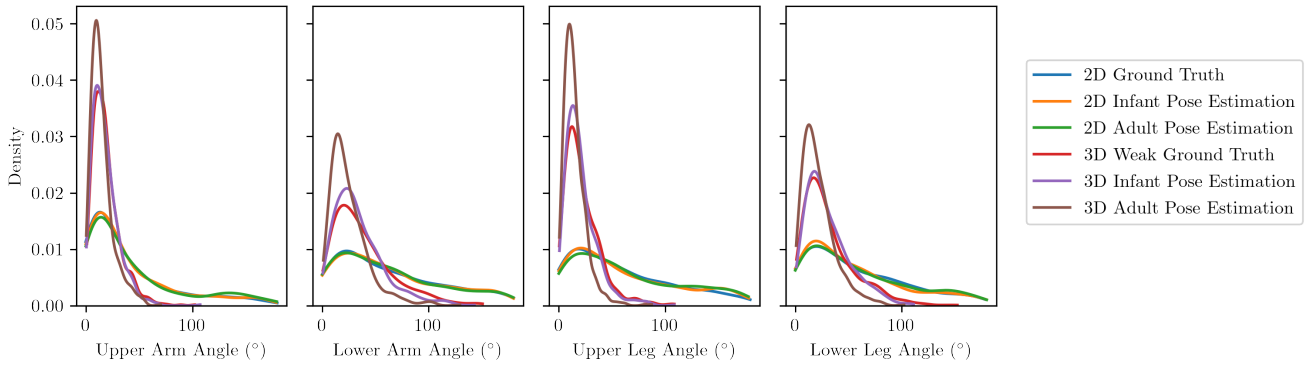

Figure 10: Distribution of raw angle differences, across four pairs of limbs and 700 real SyRIP infant images, as reported by a range of pose-based models. Models based on 3D poses yield far more consistent and seemingly realistic angles, compared with models based on 2D poses.

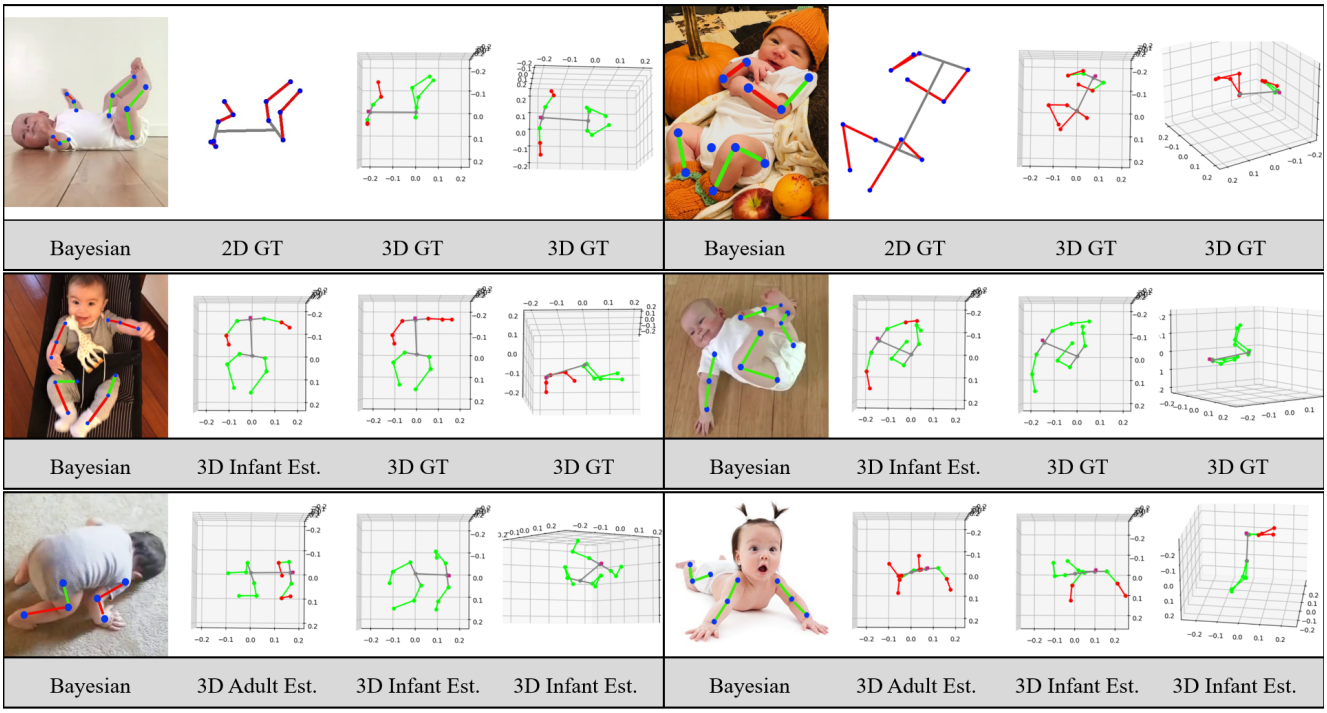

Figure 11: Comparison of the performance of different pose-based symmetry measurement results. In each row we compare symmetry assessments for each pair of limbs from the following models: *1st row:* SyRIP 2D ground truth pose (2D GT) vs. weak 3D ground truth pose (3D GT); *2nd row:* 3D predicted pose by using infant HW-HuP model (3D Infant Est.) vs. weak 3D ground truth pose (3D GT); *3rd row:* 3D predicted pose by using adult SPIN model (3D Adult Est.) vs. 3D predicted pose by using infant HW-HuP model (3D Infant Est.). Bayesian aggregate results are overlaid on the original images, as a kind of weak ground truth. Labeling conventions as in Fig. 1. 3D infant pose estimation yields better results than those obtained from the 3D adult pose estimation, but the best results come from the weak 3D ground truth. The Bayesian result of left-side example in the 3rd row is incorrect, possibly, due to the effects of occlusion on human judgement. Our pose-based methods have been trained to be robust to occlusion, and can produce objective evaluations where human assessments falter.

Table 6: Proportion of variance R^2 values from three linear regression models, all with the 3D weak ground truth raw angle as the dependent variable (and each of four pairs of limbs across 700 infants as the sample space). All three models include the labels for the limb part, the posture, and occlusion as independent variables, together with an angle class label drawn respectively the 2D infant pose estimation, 3D infant pose estimation, or the Bayesian aggregate assessment. The resulting R^2 and adjusted R^2 scores can be interpreted as gauging the predictive power of each respective model, with our 3D infant pose estimation method holding a clear advantage.

Model	R^2	Adjusted R^2
Bayesian Aggregate	0.365	0.360
2D Infant Pose Estimation	0.229	0.223
3D Infant Pose Estimation	**0.496**	**0.492**